"SCAR - SPEECH CONVERSION & ASSIST RESPONSES"

Rahul B. Adakar
Computer Engineering
VPSCET Lonavla
Student,SPPU India
rahuladkar2@gmail.com

Siddhesh B. Balghare
Computer Engineering
VPSCET Lonavla
Student,SPPU India
balgharesiddhesh@gmail.com

Abhishek A. Jadhav
Computer Engineering
VPSCET Lonavla
Student,SPPU India
abhishek17874@gmail.com

Chaitanya S. Waghmare
Computer Engineering
VPSCET Lonavla
Student,SPPU India
chaitanya080999@gmail.com

Soni R. Ragho
Computer Engineering
VPSCET Lonavla
HOD,SPPU India
sonirragho@gmail.com

ABSTRACT:

Voice assistants are inter mediator between agents that can interpret human speech and respond via synthesized voices. It's like a magical friend in your device. For people who can't see, it's like having superpowers. This amazing assistant can do lots of things! It can change different types of files, making them easy to understand. If you want a file in a special way, it can do that too, just like magic! But there's more! It can read words on the screen out loud, so you can listen instead of read. And if you talk to it, it can write down what you say. It makes technology easy and fun for everyone. Imagine it as a bridge, connecting people to the digital world. It doesn't matter if you can see or not – the voice assistant makes sure everyone can use technology without any trouble. It's all about making friends with technology and making the world a friendlier place for everyone!

KEYWORDS:

SCAR , Text to Speech , Speech to Text , Automatic Speech recognition, Mel-frequency cepstral coefficients , Acoustic Training , Wiener Filtering .

I. INTRODUCTION

Voice assistants offer a wide range of benefits to users, making everyday tasks easier and more convenient. In the heart of the digital age, where technology continues to transform the way we live, voice assistants stand out as a beacon of innovation and exclusivity. Imagine a world where anyone, regardless of their abilities, can effortlessly interact with devices and access a wealth of information, all through the power of speech. This trans formative force has not only revolutionized convenience for people worldwide but has

also become a vital source of motivation, particularly for those with disabilities, such as the visually impaired. In this journey, we will explore the incredible evolution of voice assistants, understanding the motivation behind their creation, and delving into the sophisticated techniques that empower them, including Automatic Speech Recognition (ASR), acoustic modeling, and natural language processing. The motivation behind the development of voice assistants was rooted in the simple yet profound idea of democratizing technology. It aimed to bridge the gap between the digital world and individuals, ensuring that no one was left behind. One of the primary driving forces

was the aspiration to empower people, especially those with disabilities. For the visually impaired, voice assistants became not just a convenience but a lifeline, granting them the freedom to independently navigate the digital landscape. Imagine the motivation it instilled, as individuals who once felt limited in their interactions with technology now found a powerful and accessible tool at their fingertips. Exclusivity Beyond Boundaries Voice assistants have transcended barriers, making technology universally accessible. For the visually impaired, these smart companions provide more than just convenience; they offer independence. With a simple voice command, a blind person can effortlessly send messages, make calls, read books, or even control smart home devices. This newfound accessibility empowers them to pursue education, connect with others, and engage with the world in ways that were once unimaginable. The motivation these advancements instill is immeasurable, fostering a sense of belonging and equality.

At the core of every voice assistant lies a complex web of technologies, with Automatic Speech Recognition (ASR) and acoustic modeling playing pivotal roles. ASR, a fascinating technique, enables the system to convert spoken language into written text accurately. Through intricate algorithms, ASR deciphers the nuances of speech, distinguishing words and phrases with remarkable precision. Acoustic modeling, on the other hand, focuses on understanding the unique characteristics of sounds in different environments. By analyzing acoustic patterns, voice assistants can adapt to various situations, ensuring seamless communication regardless of background noise or accents. Another marvel in the realm of voice assistants is Natural Language Processing (NLP), a technique that allows machines to comprehend human language naturally. By integrating NLP, these assistants can understand context, recognize intent, and respond in a conversational manner. This human-like interaction elevates the user experience, making interactions with technology feel intuitive and comfortable.

The motivation behind perfecting NLP lies in creating a bridge between humans and machines, enabling effortless communication and enhancing user satisfaction. The evolution of voice assistants has expanded beyond speech recognition, embracing multi-modal interfaces. These interfaces combine speech with other forms of input, such as gestures and touch, offering a more immersive and versatile user experience. For instance, individuals with mobility impairments can use gestures in conjunction with voice commands, widening the scope of accessibility. The motivation to develop multi-modal interfaces is grounded in the belief that technology should adapt to the user's needs, providing tailored solutions for every individual. As we gaze into the future, the possibilities with voice assistants seem boundless. Innovations continue to emerge, from advanced language understanding to emotion recognition, ushering in an era where technology is not just a tool but a companion that understands, empathizes, and motivates.

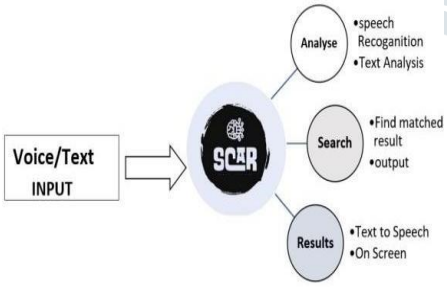

Fig.1. working flow of SCAR (Speech Conversion & Assist Responses)

The diagram is a flowchart that shows the process of how a voice assistant works. The process starts with the device gathering audio with the microphone. Recorded speech waveforms get straight to acoustic analysis, which is performed on three different levels. Then, the speech is digitized into a format that the machine can read, and analyzed for meaning. The voice assistant then decides what the user needs based on previous input and algorithms. The above diagram shows entities and their relationship for a virtual assistant system. It also includes a data flow diagram that shows the flow of data between different entities in the system.

II. LITERATURE SURVEY

A. K. Sahu , S. Dubey , A. K. Jha , R. Bhargava , P. Priya , R. Kumari explore factors influencing user trust in virtual assistant services, combining ISSM, SET, and HCI theories. They find interaction quality is key, validating their hypothesis at a 0.05 significance level, revealing a 54.9% explanatory power. Trust, innovative behavior, and use intent show positive correlation. The study emphasizes the importance of interaction quality for added value and confidence, while highlighting the significance of trust satisfaction and high-quality information in preventing trust erosion. The paper serves as a valuable literature

reference on AI-based technology adoption, especially virtual assistant services [1].

Mugdha Bapat, Pushpak Bhattacharyya et al, described morphological analyzer for almost of the Indian languages. At the starting phase the planning was about some extent homomorphism “boos trappable” encryption technique. The research proved out to be a great success for Marathi language that resulted in engagement of the Finite State Systems for the demonstration of language in a sophisticated way. Since Marathi has a really difficult morphotactics hence the growth of FSA is one of significant assistances. [2].

In the existing system of virtual assistant there are everal virtual assistants in market by using Artificial Intelligence technology. Many companies have used the dialogue systems technology to establish various kinds of Virtual Personal Assistants (VPAs) based on their applications and areas, such as Microsoft’s Cortona for Windows and Espeak for Linux, Siri for Apple, Google Assistants For Android [3].

B. S. Atal and L. R. Rabiner et al, explained regarding speech analysis, and result is regularly completed in combination with pitch analysis. The research described a pattern recognition technique for

determining whether a given slice of a speech signal should be categorized as voiced speech, unvoiced speech, or silence, depending on dimensions finished

on signal. The main restriction of the technique is the

requirement for exercise the algorithm on exact set of

dimensions picked, and for the specific recording circumstances [4].

In this paper,the benefits of empowering people with disabilities via employment goes well beyond offering opportunities for social participation and to live dignified and productive lives without seeking any help or guidance.In the workplace, people with disabilities are reported to be highly motivated and loyal, translating into extremely low turnover rates. With this observation we emerged with a solution to propose a model which can boost the confidence of these people [5].

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum. This frequency warping can allow for better representation of sound, for example, in audio compression that might potentially reduce the transmission bandwidth

and the storage requirements of audio signals [6].

The goal of the Wiener filter is to compute a statistical estimate of an unknown signal using a related signal as an input and filtering that known signal to produce the estimate as an output. For example, the known signal might consist of an unknown signal of interest that has been corrupted by additive noise. The Wiener filter can be used to filter out the noise from the corrupted signal to provide an estimate of the underlying signal of interest. The Wiener filter is based on a statistical approach, and a more statistical account of the theory is given in the minimum mean square error (MMSE) estimator article. Typical deterministic filters are designed for a desired frequency response. However, the design of the Wiener filter takes a different approach. One is assumed to have knowledge of the spectral properties of the original signal and the noise, and one seeks the linear time-invariant filter whose output would come as close to the original signal as possible. Wiener filters are characterized by the following: [7]

Assumption: signal and (additive) noise are stationary linear stochastic processes with known spectral characteristics or known autocorrelation and cross-correlation

Requirement: the filter must be physically realizable/causal .

Performance criterion: minimum mean-square error (MMSE)

III. EXISTING TECHNIQUE

- **Wiener Filtering :** Wiener filtering works by estimating and reducing noise signals in audio data through a multi-stage process. The accuracy of Wiener filtering depends on the quality of the input data and the effectiveness of preprocessing techniques. Typically, it provides substantial noise reduction, improving the clarity and intelligibility of speech signals. However, the exact accuracy can vary based on the complexity of the noise and the specific implementation parameters used. Performance is often evaluated in terms of Signal-to-Noise Ratio (SNR), with higher SNR values indicating better noise reduction.

- **BRIL :** It is a versatile block-processing frequency domain algorithm designed for noise reduction in various environments, including challenging applications like military communication systems. It is implemented in standard ANSI C and supports both floating-point and fixed-point configurations. The algorithm, with adjustable block length and noise reduction levels, efficiently processes audio signals without distorting speech quality. BRIL can be seamlessly integrated into different platforms, ranging from DSPs to CPUs, and works across a wide range of sampling frequencies without requiring calibration.

BRIL achieves up to 25 dB of background noise reduction while maintaining speech clarity. It allows users to fine-tune noise reduction levels, offering flexibility in balancing noise reduction and speech distortion. With a noise reduction level set to 20 dB, BRIL demonstrates significant noise reduction without compromising speech quality. While further noise reduction beyond 25 dB is possible, it might lead to gradual speech degradation. The algorithm has been extensively tested in both lab and field environments, proving its effectiveness across various fixed-point and floating-point processors, DSPs, and platforms, including desktop, mobile, and embedded systems.

- **Adaptive Noise Cancelling (ANC) :** It works by utilizing a secondary sensor near the source of known interference, capturing a clean reference signal devoid of the target signal and other disturbances. An adaptive filter processes this reference signal, creating an optimal estimate of the interference affecting the desired signal. By subtracting this estimated interference from the received signal, ANC effectively reduces or cancels out the unwanted interference, enhancing the clarity of the target signal. The adaptive filter continuously adjusts itself based on ongoing sampling of the reference input and the noise canceller output, adapting to changing environmental conditions. The accuracy of Adaptive Noise Cancelling depends on several factors, including the quality of the reference signal, the adaptability of the filter, and the level of correlation between the target signal and the interference. In situations where the interference is well-defined, uncorrelated with the target signal, and the reference sensor provides an accurate representation of the interference, ANC can achieve significant noise reduction, enhancing the accuracy and clarity of the desired signal. However, the effectiveness of ANC may decrease if the interference is strongly correlated with the target signal or if the reference signal is not representative of the interference. Proper calibration and selection of reference sensors are crucial for achieving high accuracy in noise reduction.

IV. PROPOSED TECHNIQUE

- **Acoustic Traning :** Acoustic training for noise cancellation involves training neural networks, such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM), using two audio signals: one with speech and background noise, and another capturing only background noise. The networks compare their predictions of clean speech against the actual cleaner speech signals. Through iterative adjustments using reset and update gates, the network learns to isolate and remove background noise from the speech signal effectively. The accuracy of acoustic training for noise cancellation

depends on the quality of the training dataset, the complexity of the neural network architecture, and the adaptability of the model. When properly trained with diverse and representative data, neural networks like GRU and LSTM can achieve high accuracy in isolating and canceling background noise from speech signals, leading to improved speech clarity. Regular fine-tuning and training with varied noise patterns enhance the accuracy and efficiency of noise cancellation models.

- **Gated Recurrent Unit (GRU) :** It is designed to process sequential data like text, speech, and time-series data. It utilizes gating mechanisms to selectively update the hidden state at each time step.

- There are three main gates in a GRU:

- **Update Gate (z):** Determines how much past knowledge should be passed to the future, similar to the Output Gate in LSTM.

- **Reset Gate (r):** Determines how much past knowledge to forget, similar to a combination of the Input Gate and the Forget Gate in LSTM.

  Current Memory Gate (\overline{h}_{t}): Introduces non-linearity into the input and ensures a zero-mean input. It is part of the Reset Gate, reducing the impact of previous information on current data passed into the future.

- **Equations for GRU Gates:**

  Reset Gate : $r_t = \text{sigmoid}(W_r . [h_{t-1},x_t])$

  Update Gate : $z_t = \text{sigmoid}(W_z . [h_{t-1},x_t])$

  Candidate Hidden State : $h_t' = \tanh(W_h . [r_t . h_t-1,x_t])$

  Hidden State : $h_t = (1 - z_t) . h_{t-1} + z_t . h_t'$

V. FUTURE SCOPE

Looking ahead, our project is set to achieve remarkable advancements. We're focusing on making our voice assistant even better. First, we want it to understand voices from far away accurately. Imagine being able to talk to your device from a distance, and it understands you perfectly - that's our goal. We're also working on making the voice assistant smart enough to recognize different voices in a noisy place. So, even if many people are talking, it will know who is speaking. Moreover, we're not stopping there. We plan to turn our voice assistant into a physical device using "Raspberry Pi." This means you could have a gadget at home that listens and responds to your voice commands. Our ultimate goal is to bring this smart voice assistant both as an app and as a physical device to the market. We're striving to make technology easy and accessible for everyone, ensuring that talking to your devices becomes simpler and more useful in your everyday life.

VI. CONCLUSION

In wrapping up our study, we've learned a lot about voice assistants. These helpful digital tools are making our lives easier every day. They can assist people with disabilities, simplify tasks, and bring convenience to our fingertips. Looking forward, we see a future where voice assistants will become even more amazing. They'll keep getting smarter and more integrated into our devices and apps. But as they advance, we must also be careful about how they handle our privacy and security. In a nutshell, voice assistants are here to stay and will keep making our lives simpler and more connected. Our research gives us a glimpse into this exciting future, where technology truly works hand in hand with people, making our lives better in every way

VII. REFERENCES

[1]. Author, A. K. Sahu , S. Dubey , A. K. Jha , R. Bhargava , P. Priya , R. Kumari (2023). nderstanding the Adoption of AI-based Innovations. Electronic copy available at: https://ssrn.com/abstract=4384623

[2]. M. Bapat, H. Gune, and P. Bhattacharyya, "A paradigm-based finite state morphological analyzer for marathi," in Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pp. 26–34, 2010.

[3]. Cortana Intelligence, Google Assistant, Apple Siri

[4]. B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced unvoiced-silence classification with applications to speech recognition," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 24, no. 3,pp. 201–212, 1976.

[5]. Empowering people with disabilities through AI https://futureofwork.wbcsd.org (accessed Sep. 26, 2021)

[6]. Min Xu; et al. (2004). "HMM-based audio keyword generation" (PDF). In Kiyoharu Aizawa; Yuichi Nakamura; Shin'ichi Satoh (eds.). Advances in Multimedia Information Processing – PCM 2004: 5th Pacific Rim Conference on Multimedia. Springer. ISBN. Archived from the original (PDF) on 2007-05-10.

[7]. Brown, Robert Grover; Hwang, Patrick Y.C. (1996). Introduction to Random Signals and Applied Kalman Filtering (3 ed.). New York: John Wiley & Sons. ISBN .