ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Research Article

Open Access

Crafting Multi-Modal Interactions on Voice Assistants

Ashlesha Vishnu Kadam

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

ABSTRACT

Voice assistants are finding adoption because of their ease and intuitiveness of use. While voice has been the dominant mode of interaction of humans with voice assistants, some embodiments of voice assistants also provide alternative modalities for interaction, popularly, via visual or touch interface. In this paper, the end-to-end working of a multi-modal voice assistant is provided, followed by a deep dive into the challenges associated specifically with multimodal voice assistants. This is followed by mitigation strategies for the challenges arising out of multimodal interaction scenarios. Next, architectural and design guidelines are provided that can provide a seamless user experience. Finally, future research areas have been identified.

***Corresponding author**

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

Received: October 10, 2022; **Accepted:** November 16, 2022; **Published:** December 20, 2022

Keywords: AI, ASR, Human Computer Interaction, NLU, Multi-Modal, Voice Assistant

Introduction

Voice Assistants are an application of Artificial Intelligence (AI) and Natural Language Processing (NLP) that perform the task of recognizing and understanding human speech and responding to it in a way that is understandable by humans [1]. Voice assistants (VAs) like Google Assistant, Alexa and Siri have become household names [2]. VAs might be accessible via an app on devices like mobile phones, tablets, smart speakers, or more [3]. Over time, these VAs have seen remarkable advancements in terms of their comprehension and fulfillment of user intent. VAs are available in various form factors and modalities [4]. For example, VAs could be either a native application in the phone or an installed application, or be embedded in smart speakers that may or may not have a screen, or be integrated in the car. The primary mode of interaction of humans with VAs continues to be via the voice interface. The increasing popularity of this form of interface with technology has resulted in an increased research towards its user experience.

An important aspect of VAs is their usability. A literature review was conducted to evaluate the usability of VAs using the ISO 9241-11 framework [5]. The study evaluated VAs across 3 dimensions – effectiveness, efficiency and satisfaction. The study concluded that there is scope to improve the effectiveness and efficiency of VAs in getting the requested job done. It also found challenges that need to be addressed to improve the usability of VAs. These challenges are the need to improve the accuracy of understanding user requests and responding comprehensively to them, the need for better natural language understanding and responding, the need to incorporate contextual understanding in interactions, and the need to have a channel to communicate with humans about what they are doing wrong and how to correct it. There is also literature about VAs being a sub-set of conversational agents or chatbots, and have been designed to understand and respond in the way humans interact [6].

While VAs are voice-forward, they can communicate with humans using various modalities over and above voice (e.g., images, videos, text, hardware motion). Voice is the primary modality that users use when it comes to Vas [5]. VAs employ Automatic Speech Recognition and Natural Language Understanding to understand the utterance issued by a human and respond to it. Examples of voice commands issued by users to a VA are asking for the news or weather, or asking the VA to play music, or commanding the VA to perform specific tasks like securing the home alarm or setting the timer for 10 minutes. Another modality provided in conjunction with voice in VAs is text. Some VAs allow user input in the form of text (e.g., a Google Assistant app in an Android phone), allowing users to type their query instead of having to use their voice. This feature can be especially beneficial where the VA is not succeeding at accomplishing the task at hand because of its failure to understand the voice-issued request in the first place due to various reasons like poor ASR due to background noise, or request for non-popular content that has a high chance of failing at natural language understanding (NLP), and so on. Another modality VAs offer is images and videos. VAs like Alexa and Google Assistant are embedded in multi-modal devices like Echo Show and Google Hub that have a screen, providing the opportunity to show images in response to questions to the VA, or to display videos that goes along with the verbal response of the VA or even simply "quiet" suggestions on the screen. Some VA-embodying devices are also able to respond to non-verbal queues, like taps, facial recognition, presence detection, and so on.

This paper dives deeper into the multi-modal aspect of VAs, enabling VAs to respond to both visual and verbal inputs from users. The paper shares unique insights, challenges and future

opportunities with the integration of vision and verbal language. In the context of this paper, for simplicity, we assume multimodal VAs refer to VAs with visual and voice I/O capabilities.

How Multi-Modal Voice Assistants Work

Multimodal VAs that are commonly available include VAs in the phone (e.g. Siri in iPhone), VAs integrated into the car (e.g. native car VA like BMW's Intelligent Personal Assistant or Google Assistant in Chevrolet) and VAs integrated into multimodal smart speakers (e.g. Alexa in Echo Show or Google Assistant in Google Nest Hub Max) [2]. These VAs can take input in the form of voice commands (e.g. "what's the latest news?"), touch (e.g. clicking on a news card that is displaying on the VA's screen) or both (e.g. asking the VA to play your playlist, and then selecting one of your five playlists displayed on the screen as a way to disambiguate the playlist you are referring to). The output, too, can be either voice-only (e.g. streaming the latest news), touch (e.g. displaying certain cards in response to an input) or both (e.g. a TTS asking user to choose one of the five playlists displayed on the screen).

In order to understand the challenges associated with interactions on a multimodal VA, it is important to understand the end-to-end working of a VA. Let's assume a hypothetical VA, Duna, for this purpose. As seen in Figure 1, a user might provide a multimodal input to the VA via both spoken word ("Hey Duna, play music by this artist") and touch (i.e. pointing to or touching one of the 3 artists on Duna's screen).

VAs can be activated using their specific wake words [7]. For example, users of Amazon's VA say "Alexa" while those of Apple's VA say "Hey Siri" to invoke the respective VA. In this example, the user says "Hey Duna" to alert the VA to start processing the words following the wake-word "Duna" for interpreting user intent. 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) [7]. 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]. In this example, the user provides an ambiguous contextual reference "this". At the same time, the user provides a touch input via screen, indicating the artist that they are referring to on the VA's screen. In this case, out of the 3 artists on the screen, the user is referring to artist at position 1. This information is also provided to the NLU layer so that it can be used in conjunction with the voice cue to provide a complete interpretation of the user's intent – to play music by the artist Ed Sheeran.

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 [8]. 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. The output is also displayed on the screen in the form of a list of songs by Ed Sheeran, and the playback of music starts after the TTS ("Here's some music by Ed Sheeran"). See Figure 1 for details.

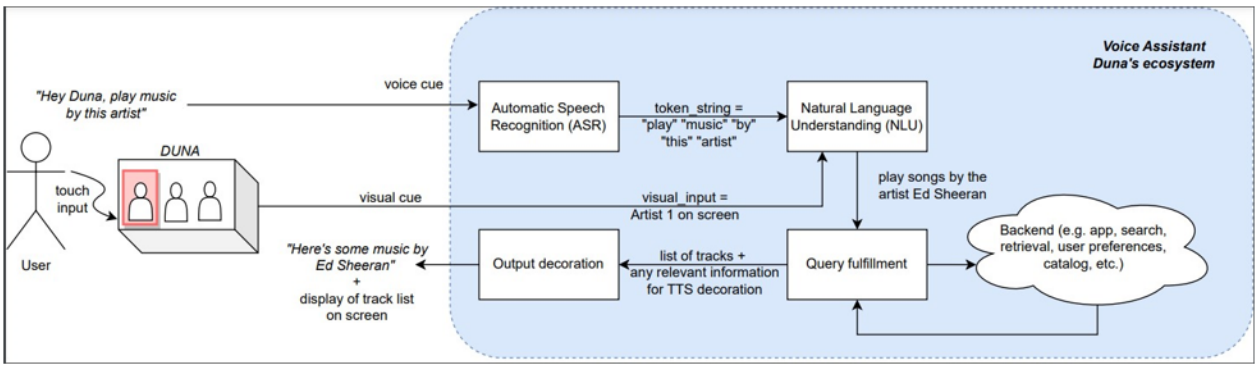

Figure 1: Hypothetical Voice Assistant Duna's Ecosystem

Insights and Challenges with Multi-Modal Voice Assistants

Combining voice and visual input-output mechanisms into a VA can unlock new capabilities for the VA, but can also present unique challenges.

Visual Cues

A user's verbal query to a VA can be enhanced if fused with the contextual information available via visual cues [9]. In case of a multi-modal device, this can refer to users issuing voice commands while providing visual cues to the VA via a camera or other means. A combination of voice and visual inputs should increase the chances of the VA understanding the user intent accurately [10]. For example, if there is a music event like a concert happening, the user can simply point to any artifact associated with the concert (e.g., hoarding, pamphlet, social media posts) and ask the VA questions about "it" without necessarily having to disambiguate that "it" refers to the music event. A challenge that arise in these scenarios is that if the visual input is too complex, the VA might find it hard to extract the right, and meaningful, context from the visual inputs to enrich the verbal query of the user [8]. For example, continuing with the music concert example, if the user simply points the visual input interface of the VA to a pamphlet about a music concert, but there are multiple pamphlets surrounding the pamphlet of interest about other events, including but not limited to music concerts, the VA might find questions like "when is it happening" to be challenging, because it is hard to disambiguate which of the N concerts whose information is in front of the user across pamphlets, is being referred to by the user when they asked the question to the VA.

Ambiguity of User Intent

While multi-modal interactions with a VA enrich the inputs available to the VA to better understand and serve the user intent, there could be ambiguity introduced because of the multi-modal aspect of VAs [11]. In such scenarios, the VA needs to accurately interpret the user's intent out of the multiple possible interpretations because of the combination of the two modalities. For example, if a user simply points the VA to a technology magazine and issues a verbal query to "read that part about AI" from the magazine, it is ambiguous to figure out if the user is referring to certain parts of the article that you might currently be reading or if they want to read another article about AI in the magazine. Another example is the user asking the VA to play Evermore by Taylor Swift while pointing to a smart television. It is hard to disambiguate if the user meant for the music to play on the VA (i.e. the device that the VA being conversed in is embedded) or the smart television the user pointed to.

Voice and Visual Input Synchronization

It is challenging to integrate real-time visual data with voice interactions in a way that it appears in perfect synchrony to users. Being able to achieve this lockstep integration is challenging. An example is when a user might be interacting with a VA while the VA's visual screen elements need to change per the voice interactions, like choosing what items to add to the cart for shopping via a VA. Another class of problems is for the VA to be able to locate a specific object correctly in a cluttered visual scene and associating that object with the user query to provide the right response [10]. For example, pointing to a car in a parking lot and asking the VA "what model is that car?" without mentioning what car they are referring to, or referring to a book in a bookshelf and asking "is there a French translation of this book?" without mentioning the book name.

Mismatch in Contextual Relationships

Any interaction can introduce hierarchical relationships. However, when it comes to multi-modal interactions, there is a higher level of complexity introduced because of multiple modalities conveying information at varying levels of context. For example, a user start talking about the style of the painting they are standing in front of and pointing to, and then proceed with voice interactions, within the broader context of this visual cue about the painting. In order for a model powering the VA to be able to respond coherently, the model needs to be able to understand these contextual and hierarchical relationships. This involves not just understanding contextual interpretations within a modality (e.g. if the user says "it is a perfect symbol of the cubism era", "I particularly like the choice of colors in it", knowing what "it" means in every sentence from the voice input), and also across modalities (e.g. if the user says "but I think this color could have been muted more", knowing what "this" means from the visual input).

Temporal and Spatial Synchronization

Another aspect of synchronization in addition to the voice – visual one described in Section 3.3 is time and space synchronization to provide accurate experiences to users [12]. For example, if a user is interacting with a VA in the car, and they ask the VA to "remind me to buy milk and eggs when I pass by a grocery store", the VA needs to be able to spatially map out the user's trajectory as well as estimate the time it will take to progress on this trajectory to accurately predict the time to remind the user about making a stop. Being able to link the user's future location with time is a non-trivial challenge.

Multilingual Multi-Modal Inputs

Section 3.3 above mentions the challenges with being able to synchronize voice and visual inputs in order for the VA to provide accurate responses. These challenges are multiplied when the inputs, both voice and visual, are multilingual. Differences in linguistic and visual structures can make integrating of voice and visual inputs challenging. The complexities with respect to differences in grammar, script, phrases and idioms, and cultural references can further make accurate interpretation by the VA challenging.

Mitigation Strategies for Challenges with Multi-Modal Voice Assistants

This section provides potential mitigation strategies for some of the challenges mentioned in Section 2. These mitigation strategies require a user-centric design approach, coupled with advanced technical solutions to overcome the challenges and provide a good user experience.

Visual Cues: Increasing Accuracy of Interpretation

When interpreted unambiguously, visual inputs can enrich the understanding of a user's query and context to provide accurate responses. This requires development of advanced visual algorithms that can detect and classify objects, make sense of complex scenes, understand context, and more. Object detection algorithms, such as Faster R-CNN and YOLO, can accurately detect and classify objects in real-time [13]. Algorithms that perform semantic segmentation to partition an image into multiple segments and assigning each segment a semantic label, can help segment complex images to identify the context of objects within them. Some examples of such algorithms are DeepLab and PSPNet [14].

Instance segmentation, which involves detecting and segmenting objects in a video, such as Mask R-CNN and Panoptic FPN, can also be used to detect objects and their context within the video [2,7]. ML models that can extract high level features from visual inputs and correlate them with the user's voice queries can also improve the accuracy of interpretation. Finally, Natural Language Processing (NLP) techniques can be used to generate responses capturing the most relevant aspects of the visual context.

Ambiguity: Resolution with Cross-Modal Grounding

Ambiguities arising because of inputs from multiple modalities can be resolved with cross-modal grounding mechanisms. Multi-modal grounded learning can be used with both visual and voice aspects to allow a VA to communicate, ground and learn from language [15]. Gaze tracking can be used to understand what the user is focusing on and use that context to better interpret queries [16]. Similarly, any other gestures like pointing, turning head, and others can be used in addition to the voice query for eliminating ambiguity and improving accuracy of interpretation. In order to be able to recognize specific objects in a clutter in order to use them in conjunction with the user voice query, models that can comprehend and process fine grained attributes of objects like colors, shapes, specific attributes, should be used.

Voice and Visual Input Synchronization via Alignment Methods

In order to achieve synchronization across voice and visual cues as inputs, alignment methods like the hidden Markov model and dynamic time warping can be used. The hidden Markov model can be used to model relationships between visual and voice cues and estimate alignment between the audio waves and visual frames.

Dynamic time warping on the other hand can be used to create alignment of data sequences by stretching or compressing them in time in order to match them more closely.

Contextual Mismatch Resolution

In addition to the techniques mentioned in Sections 3.2 and 3.3, another opportunity to reduce contextual errors is by enabling users to provide inputs via touch in case of ambiguity in interpretation. For example, if a user points the VA to a magazine with multiple artists on the cover page and asks the VA to "play his music", the VA might display all the male artists it was able to recognize on the cover of the magazine and ask the user to pick the one they were referring to. Finally, contextual information from previous interactions, whether voice-only, visual-only or multi-modal, should be taken into consideration to increase the chances of a successful query interpretation.

Temporal and Spatial Synchronization Techniques

Models that can track and predict user movement with time should be developed in order to predict future locations. Geographic information systems (GIS) data should be used to enhance spatial reasoning and understanding of spatial contextual cues from user [17,18]. This approach can help in accurately interpreting and correlating spatial and temporal information from both voice and visual inputs in a noisy environment. Sophisticated time-series analysis might also be able to help infer patterns in user movement to predict future interactions [19].

Multilingual Multimodal Support

The first step in simplifying multimodal inputs is to leverage machine translation models to convert all spoken language into a common language for processing. The accuracy of this translation is critical to ensuring high quality interpretation. Developing cross-lingual embeddings that allow mapping of not simply words, but semantic understanding space can help in cross-lingual understanding [20]. Another alternative is building language specific models and then combining the outputs to form a single unified response [21].

Architecture and Design Considerations for Multimodal Voice Assistants

Along with the mitigation strategies shared in the previous section, there are additional architectural and design considerations that must be adhered to in order to provide a seamless experience to users.

Contextual Memory

Along with being able to use voice and visual cues to interpret a user request accurately, it is important to create a dynamic contextual memory framework that "remembers" previous interactions to maintain context across sessions. For example, if a user tells a VA once that they prefer the Taylor Swift version of the song This Love, the VA should always play the Taylor Swift version and not the Maroon 5 version of that song.

Attention Mechanisms

It is important to design attention mechanisms that dynamically focus on the most relevant voice and visual inputs during multimodal interactions. This ensures that the model's interpretations shift as the user's focus shifts.

Cross-modal Learning Transfer

Implementing transfer of learning across modalities can enhance overall understanding of the VA across modalities. For example, if the VA has learned via voice modality that the user hates country music, transferring that learning to the visual modality can ensure that if a user were to point to a magazine and ask the VA to add the artist on the magazine cover to their library, the VA will interpret that out of the two artists on the cover, the user likely meant the non-country one.

Human-centric Design

Designing a VA that allows seamless voice and visual interactions while also ensuring low latency, high degree of ease of use and honoring of user preferences for privacy is challenging. Ensuring the VA is designed in a human-centric way, especially for multi-modal interactions is important. At its best, this means that the user doesn't have to care about the language they use or how much detail they specify in their interactions with the VA or whether they choose voice or visual modality to convey their intent to the VA, the VA understands correctly.

Future Research Areas

This paper explores the top existing challenges in multimodal VA interactions and how to mitigate them. It also provides direction to architectural and design considerations. However, there are several areas that are still not explored as thoroughly in this space. Some top areas for future research include the following – dynamic contextual hierarchies to ensure the most relevant cues are prioritized, methods to integrate multimodal data while honoring user privacy settings especially for visual information, adaptive learning technologies for the multimodal VA to adjust to how the user is interacting with it, ethical concerns about usage of visual cues to interpret user intent and natural multimodal generation that is similar to natural language generation but generates multimodal cues instead of simply language ones.

Conclusion

Multimodal interactions with a VA can vastly enhance the inputs available for the VA to contextually and accurately predict user intent and fulfill it satisfactorily. However, multimodal interactions also introduce complexities and challenges. These challenges can be mitigated to varying degrees in order to create a rich and engaging experience for users. There are areas that require further research in order to continue enriching multimodal interactions.

References

1. (2022) A Simple Explanation of How Voice Assistants Work And Why You Need One. Artificial Intelligence https://www.europeanbusinessreview.com/a-simple-explanation-of-how-voice-assistants-work-and-why-you-need-one/.
2. Pridmore J, Mols A (2020) Personal choices and situated data: privacy negotiations and the acceptance of household intelligent personal assistants. Big Data Soc 7: 205395171989174.
3. Lewis Silkin LLP (2021) Future of voice assistants: how the VA might overtake the PA. Future of Work Hub https://www.futureofworkhub.info/explainers/2021/4/19/future-of-voice-assistants-how-the-va-might-overtake-the-pa.
4. Rzepka C, Berger B, Hess T (2022) Voice assistant vs. Chatbot–examining the fit between conversational agents' interaction modalities and information search tasks. Inf Syst Front 24:839-856.
5. Dutsinma FLI, Pal D, Funilkul S, Jonathan H Chan (2022) A Systematic Review of Voice Assistant Usability: An ISO 9241–11 Approach. SN Computer Science 3: 267.
6. Voice Assistants. JMIR Publications Advancing Digital Health & Open Science

https://www.jmir.org/themes/1010-voice-assistants.

7. Luiza Stelitano (2021) A quick and easy guide to voice assistants. Artificial Intelligence https://www.miquido.com/blog/what-are-voice-assistants/.
8. Almohsen Ranya (2014) Human Interaction Recognition with Audio and Visual Cues. Graduate Theses, Dissertations, and Problem Reports https://researchrepository.wvu.edu/etd/533.
9. Darda, Chitnis (2020) Voice Assistant: A Systematic Literature Review. ResearchGate 7: 67-72.
10. Hegde Thomson, Serena K Thompson, Mark Brady, Daniel Kersten (2012) Object recognition in clutter: cortical responses depend on the type of learning. Frontiers in Human Neuroscience 6: 1-15.
11. Lee Jaewook, Rodriguez S, Raahul Natarrajan, Jacqueline Chen, Harsh Deep et al, (2021) What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants https://sebas.me/docs/ICMI2021_MultimodalVA.pdf.
12. Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, et al. (2021) Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. Electrical Engineering and Systems Science https://arxiv.org/abs/2107.06592.
13. Chai Wenhao, Gaoang Wang (2022) Deep Vision Multimodal Learning: Methodology, Benchmark, and Trend. Applied Sciences 12: 6588.
14. Keren Ye (2021) Multimodal Knowledge Integration For Object Detection And Visual Reasoning. University of Pittsburg. https://d-scholarship.pitt.edu/41514/13/Keren_Thesis_v7.pdf.
15. Bohdan Ponomar (2022) Multimodal Grounded Learning with Vision and Language. KD Nuggets https://www.kdnuggets.com/2022/11/multimodal-grounded-learning-vision-language.html.
16. Matthew Marge, Carol Espy Wilson, Nigel G Ward (2019) Spoken Language Interaction with Robots: Research Issues and Recommendations. Report from the NSF Future Directions Workshop https://arxiv.org/pdf/2011.05533.pdf.
17. João Paulo Figueira (2020) Map-Matching for Trajectory Prediction. Medium https://towardsdatascience.com/map-matching-for-trajectory-prediction-be307a1547f0.
18. Anita Graser aka Underdark (2018) Movement data in GIS #18: creating evaluation data for trajectory predictions. Free and Open Source GIS Ramblings https://anitagraser.com/2018/12/15/movement-data-in-gis-18-creating-evaluation-data-for-trajectory-predictions/.
19. Edwin Lisowski (2019) The best Forecast Techniques or how to Predict from Time Series Data. Medium https://towardsdatascience.com/the-best-forecast-techniques-or-how-to-predict-from-time-series-data-967221811981.
20. Alvaro Nuez Ezquerra (2018) Implementing ChatBots using Neural Machine Translation techniques. Universitat Politecnica de Catalunya https://upcommons.upc.edu/bitstream/handle/2117/117176/TFG_final_version.pdf?sequence=1&isAllowed=y.
21. Ves Stoyanov (2018) Under the hood: Multilingual embeddings. Engineering at Meta https://engineering.fb.com/2018/01/24/ml-applications/under-the-hood-multilingual-embeddings/.