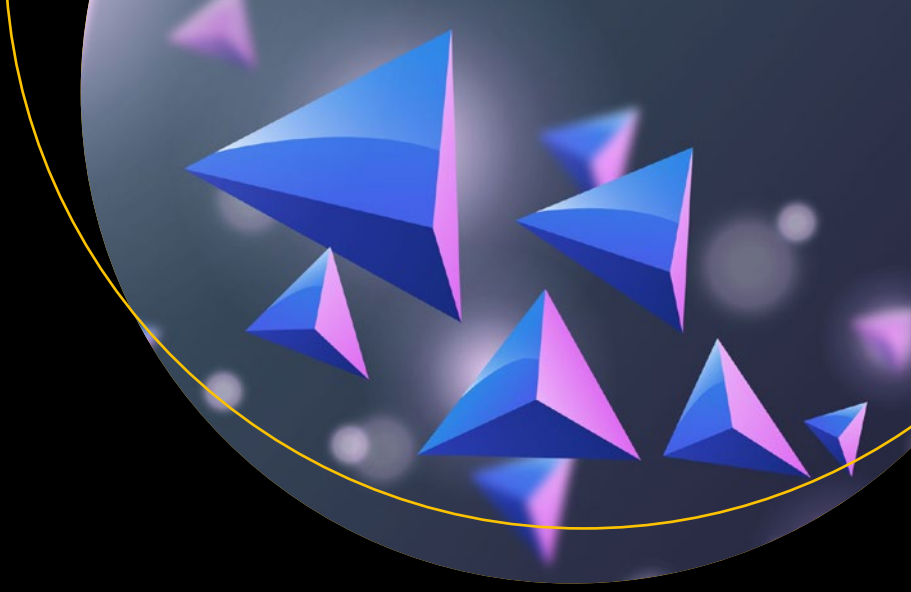

Voice User Interface Design

Moving from GUI to Mixed Modal Interaction

Ritwik Dasgupta

Apress®

Voice User Interface Design

Moving from GUI to Mixed Modal Interaction

Ritwik Dasgupta

Apress®

Voice User Interface Design: Moving from GUI to Mixed Modal Interaction

Ritwik Dasgupta
Hyderabad, Telangana, India

ISBN-13 (pbk): 978-1-4842-4124-0 ISBN-13 (electronic): 978-1-4842-4125-7
https://doi.org/10.1007/978-1-4842-4125-7

Library of Congress Control Number: 2018966797

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Smriti Srivastava
Development Editor: Laura Berendson
Coordinating Editor: Shrikant Vishwakarma

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit http://www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-4124-0. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

Table of Contents

About the Author

Ritwik Dasgupta works as a UX designer with Microsoft, India. He works on the Cortana team for Windows 10, assistant-enabled devices, and iOS and Android apps. He received his Bachelor's of Architecture degree from NIT Calicut and his postgraduate degree in Industrial Design (MDes) from IIT Delhi.

About the Contributor

Akshat Verma completed his masters in new media design from the National Institute of Design. He has actively worked on voice user interfaces, interaction design, Voice UX, context-aware computing, user interface design, and experience design using technologies to create new and engaging experiences on screens and beyond.

He is currently working in AVP Innovation Design & Technology at the Newzstreet Media Group and looks after product development and identifying potential strategies for fulfilling business revenue opportunities with the product updates and features.

He has specialized his focus area on voice-based UI systems to create new experiences, having already worked to create successful voice interactions on Amazon Alexa and Google Home technologies in the Indian markets.

He previously worked with the global strategy team at Honeywell (HTS) on voice recognition technology while exploring the areas of context-aware computing and was part of the core team that designed India's first audio e-learning platform called I-Radiolive.com.

About the Technical Reviewers

Simonie Wilson has worked in speech and voice user interfaces for 20+ years. Her career in Computational Linguistics has taken her from big companies like Microsoft and GM to startups, contracting, and back again. With a masters from Georgetown University, Simonie has participated in numerous conferences and workshops and holds a patent in dialog design. Her current focus is on usability and best practices for these systems and the tools used to build and tune them.

Kasam Shaikh is a certified Azure architect, global AI speaker, technical blogger, and C# Corner MVP. He has more than 10 years of experience in the IT industry and is a regular speaker at various events on Azure. He is also a founder of DearAzure.net. He leads the Azure India (azINDIA) online community, the fastest growing online community for learning Microsoft Azure. He has a concrete technical background with good hands-on experience in Microsoft technologies. At DearAzure.net, he has been organizing online free webinars and live events for learning Microsoft Azure. He also gives sessions and speaks on developing bots with Microsoft Azure cognitive and QnA Maker service at international conferences, online communities, and local user groups. He owns a YouTube channel and shares his experience over his web site at https://www.kasamshaikh.com.

CHAPTER 1

Introduction to VUI

This is 2019. The year becomes significant when we start talking technological advancements and their effects as we move forward. Every year, we see something new, something that has the potential to change technology forever. But as American fiction author William Gibson puts it aptly, "The future is already here; it is just not very evenly distributed." The year acts as a milestone, a benchmark for the immense amount of effort for the entire civilization to reach to this point, and shows where we are headed in the near future.

Voice User Interface (or VUI) is an interaction model where a human interacts with a machine and performs a set of tasks at least in part by using voice. For example, "Hey Siri, tell me today's headlines" is a simple VUI command where Siri identifies and "tells" the user the news as output. In a similar manner, IVR (Interactive Voice Response) systems are widely used in the banking and travel industries. These systems are primarily dependent on voice biometrics for identifying the users and choosing the set of tasks that the user wants to complete using voice as a primary interaction mode.

The explosion of VUI has come about at the same time that major companies have started experimenting with fluid cross-device experiences. We live in a time where Alexa aims to become our go-to shopping assistant, Google is our search assistant, and Cortana is our work assistant. Imagine using an travel booking web site to book a flight. Once the flight booking is completed and the travel details are confirmed, the

R. Dasgupta, *Voice User Interface Design*, https://doi.org/10.1007/978-1-4842-4125-7_1

various assistants set automated reminders on your phone to remind you to catch your flight or to show you the traffic conditions before catching your flight so that you may reach the airport on time.

But voice recognition is not a new technology.

When Did It All Start?

An experimental device designed by IBM in 1961, the *Shoebox* was an early effort at mastering voice recognition. The machine recognized 16 words spoken into its microphone and converted those sounds into electrical impulses. It was first demonstrated at the 1962 World's Fair in Seattle by its developer, William C. Dersch of the Advanced Systems Development division. The name given was *Shoebox,* owing to its small size. This was the beginning of two new technologies—Automated Speech Recognition (ASR) and Natural Language Understanding (NLU). This dealt with only the first part—voice recognition. For a pure voice-user interface, the machine needed to generate a human voice. This was experimented on even earlier, as early as 1939.

The Voder by Homer Dudley (Bell Telephone Laboratories, Murray Hill, New Jersey) was the first device that could generate continuous human speech electronically. In 1939, Alden P. Armagnac wrote in *Popular Science* magazine about this speaking device. It was created from vacuum tubes and electrical circuits, by Bell Telephone Laboratories engineers. It was meant to duplicate the human voice. To manufacture conversation, the machine operator employed a keyboard like that of an organ. Thirteen black and white keys produced all the vowels and consonants of speech. Another key regulated the loudness of the synthetic voice, which came from a loudspeaker. A foot pedal varied the inflection so that the same sentence may state a fact or ask a question. About a year's practice enabled an operator to make the machine speak.

Time magazine wrote on January 16th, 1939, that Bell Telephone demonstrators made it clear that Voder did not reproduce speech, like a telephone receiver or loudspeaker. It created speech via an operator

who synthesized sounds to form words. Twenty-three basic sounds were created by a skilled operator using a keyboard and foot pedal. Two dozen operators trained for a year.

The VUIs were interactive voice response (IVR) systems that understood human speech over the telephone in order to carry out tasks. In the early 2000s, IVR systems became mainstream. Anyone with a phone could book plane flights, transfer money between accounts, order prescription refills, find local movie times, and hear traffic information, all using nothing more than a regular phone and the human voice.

So, how does this put "today's" technology into perspective?

Technologies like voice interaction, augmented reality, and virtual reality, among others have been present or been researched for a relatively long time. What makes the current offerings exciting is that they are finally widely commercially available, and we have a need for designers and engineers who can take up the challenge to develop scenarios to solve everyday problems for the user.

This is very similar to when GUI became the norm for human-machine interaction, where we felt the need for designers to clear up the clutter, simplify the data, and present the users with flows and solutions that were easier to grasp. Let's take a TV remote as an example. It can be extremely difficult to operate one when we have 20-30 buttons on the device and it becomes difficult for a person to comprehend what all the buttons do. Without good design, technology is difficult or even impossible to use.

We need to realize that we are in the next era of VUIs—the era of digital assistants. At present, there are many things that a digital assistant can do well by voice, but there are still many things it just cannot do.

Era of Digital Assistants

We are gradually getting more and more dependent on digital assistants like Siri and Alexa to get information or do tasks. But there are two types of assistants—one that uses only text to interact with us, which includes