MAKING VOICES HEARD

EVOLUTION AND TYPOLOGY OF VOICE INTERFACES

Literature Surveys

Making Voices Heard
Literature Surveys: Evolution and Typology of Voice Interfaces

Research and Writing **DEEPIKA NANDAGUDI SRINIVASA, SHWETA MOHANDAS**
Review and Editing **SAUMYAA NAIDU, PUTHIYA PURAYIL SNEHA, PRANAV MANJESH BIDARE**
Research Inputs **SUMANDRO CHATTAPADHYAY**
Copyediting **THE CLEAN COPY**
Illustration **KRUTHIKA N.S.**
Report Layout and Design **SAUMYAA NAIDU**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Background

The availability of multiple modes of interaction such as voice and gesture makes devices accessible to a wide variety of people. Voice interfaces (VI), in particular, create a level playing field for those who are limited by single-language, text-based interfaces.

Schnelle-Walka defines VIs as "user interfaces using speech input through a speech recognizer and speech output through speech synthesis or prerecorded audio".[1] In essence, VI technologies involve two processes: one converting the language to code that a computer understands, and converting the computer language back to a language that the human understands. Considering that the predominant means of input for VIs is speech, they are also known as natural language interfaces.[2]

2. Tracing the evolution of VIs

Before Siri and Alexa, we had 'Audrey', created by Bell Laboratories' Harry Fletcher and Homer Dudley, who are considered the pioneers of VIs for their groundbreaking research on speech synthesis and human speech modelling.[3] In 1952, Audrey was used for number recognition through spoken input.[4] A decade later, IBM's 'Shoebox' could not only recognise digits from zero to nine but also comprehend 16 words.[5]

In 1992, AT&T Telefonica developed a speech-to-speech prototype, VESTS (Voice English/Spanish Translator), which relied heavily on spoken language translation.[6] VESTS, a speaker-trained system that could process over 450 words, was exhibited at the Seville World's Fair in Spain. VIs have come a long way from these early prototypes to modern voice assistants, such as Alexa, Siri, Cortana, and the Google Assistant, which are now accessible to consumers worldover.[7]

One of the main reasons for the proliferation of VIs today is that since 2012 smartphones come with a built-in VI. According to a 2018 PwC survey, consumers issued voice commands most commonly on smartphones from among a plethora of voice-enabled devices.[8] Mobile phones now operate almost like 'shrunken

1 Schnelle-Walka, D., "I Tell You Something," *Proceedings of the 16th European Conference on Pattern Languages of Programs - EuroPLoP '11*, 2011.

2 Miller, L., "Natural Language Interfaces," *Journal of the Washington Academy of Sciences 80*, no. 3 (1990): 91–115, accessed on 3 June 2020, www.jstor.org/stable/24531256.

3 Bhowmik, A. K., *Interactive Displays: Natural Human-Interface Technologies* (John Wiley & Sons, Incorporated, 2014).

4 Carbone, C., "Audrey, Sibyl, and Alice in the Technical Information Libraries," *STWP Review 9*, no. 1 (1962): 14–15, accessed on 19 June 2020, www.jstor.org/stable/43091178

5 "IBM Shoebox," *IBM Archives*, accessed on 2 November 2021 https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

6 IBM Archives, IBM Shoebox. Retrieved from https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

7 Tank, N., "Voice User Interface (VUI) – A Definition," *Bot Society Blog*, 2018, https://botsociety.io/blog/2018/04/voice-user-interface/

8 "Consumer Intelligence Series: Prepare for the Voice Revolution," *PwC Survey*, 2018. https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html

desktops' because of their inherent operational versatility. However, the reduced screen size is the primary structural limitation of these devices. To overcome this limitation, voice has become an important input to complete tasks without having to use the touch function or type on their phones.[9] Hence, developers have now integrated cloud-based voice technologies into devices – as in the case of Amazon Echo and Google Home as well as through open-source initiatives such as Mozilla's Deep Speech, which is an open-source speech-to-text engine.[10]

3. Features of VIs

In the early 90s, researchers identified the five basic elements[11] of voice processing technologies:

1. **Voice coding**: the process of compressing the information transmitted through the voice signal to transmit or store it economically in systems of a lower capacity.
2. **Voice synthesis**: the synthetic replication of voice signals to facilitate the transmission of information from machine to human.
3. **Speech recognition**: the extraction of information that is there in a voice signal to control the actions taken by the device in response to spoken commands.[12]
4. **Speaker recognition**: the identification of voice characteristics for speaker verification. This process ensures that the speaker is verified through their voice characteristics.
5. **Spoken language translation**: On recognising the language the person is speaking in, the translation of a message from one language to another. Through this process, two individuals who do not speak the same language can communicate.[13]

Voice output is of two distinct categories: pre-recorded speech and synthetic speech.[14] Pre-recorded speech is natural speech that is recorded and stored for future use. In contrast, synthetic speech employs natural language processing (NLP) for the automatic generation of appropriate natural-language responses or

9 Breen, A., et al., "Voice in the User Interface," in *Interactive Displays: Natural Human-Interface Technologies*, ed. Bhowmik, A. K. (John Wiley & Sons, Incorporated, 2014): 107.

10 Lawrence, H. M. "Beyond the Graphic User Interface," In *Rhetorical Speculations: The Future of Rhetoric, Writing, and Technology*, ed. Sundvall, S., (Logan: University Press of Colorado, 2019).

11 Rabiner, L. R., "Voice Communication Between Humans and Machines –An Introduction," in *Voice Communication Between Humans and Machines*, ed. D. B. Roa and J. G. Wilpon (The National Academies Press, 1994), https://doi.org/10.17226/2308.

12 Rabiner, "Voice Communication between Humans and Machines."

13 "Voice User Interfaces," *Interaction Design Foundation*, https://www.interaction-design.org/literature/topics voice-user-interfaces.

14 Candace Kamm, "User Interface for Voice Applications", in *Voice Communication Between Humans and Machines*, eds. David B. Roe and Jay G. Wilpon (The National Academies Press, 1995), 428-429.

output in the form of written text.[15]

NLP involves the conversion of textual information into speech and vice-versa, which enables a device to discern and process natural language data. The system then processes this data by standardising text inputs and splitting it into words and sentences. Then, the device can ascertain the syntax of the input provided. NLP comprises two main natural-language principles:[16]

1. **Natural language understanding (NLU)**: NLU is a branch of NLP that deals with reading comprehension, synonyms, themes, and lexical semantics. It is used to construct the responses of VIs through responses of VIs algorithms.[17]

2. **Natural language generation (NLG)**: The first step of NLG involves processing relevant content from databases. This is followed by sentence planning, which involves the formation of natural-language responses through text realisation. As a consequence, the NLG process delivers a meaningful and personalised response, as opposed to a pre-scripted one.[18]

Synthetic speech employs NLP for its characteristically high 'segmental intelligibility' – or its ability to understand each segment of speech. However, pre-recorded speech outputs tend to be preferred by all for their human voice and pronunciation characteristics. These characteristics exist on the condition that the pre-recorded speech maintains the delicate balance between natural prosody[19] and the recorded elements. Since it successfully maintains the quality of natural speech, the natural prosody of pre-recorded speech output is higher than that of synthetic speech.[20]

4. Types of VIs

A plethora of developers are creating VIs that can perform various functions, thereby giving a wide array of definitions to similar interfaces. Interactive voice response (IVR), voice channels, voice bots, and voice assistants are variations of voice-based customer service solutions.[21] Although these terms are sometimes used interchangeably, some authors opine that there are nuanced differences that

15 Androutsopoulos, I., *Exploring Time, Tense and Aspect in Natural Language Database Interfaces*, (John Benjamins Publishing Company, 2002).

16 "AI – Natural Language Processing", *Tutorials Point* accessed on 11 November 2021, **https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm**.

17 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*, 24 December, 2019, **https://tinyurl.com/mrxp9emu**.

18 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

19 Lauren Applebaum, et al. (2015) note that "Prosody, the intonation, rhythm, or 'music' of language, is an important aspect of all natural languages. Prosody can convey structural information that, at times, affects the meaning we take from a sentence." In "Prosody In a Communication System Developed without a Language Model," *Sign language and linguistics vol. 17*, no. 2 (2014): 181–212, doi:10.1075/sll.17.2.02app

20 Kamm, "User Interface."

21 Caile, C., "Keto or Atkins? IVR or Voice bots?" *Nuance*, 2019, accessed on 4 January 2022, **https://whatsnext.nuance.com/enterprise/voice-bots-and-ivr-similarities/**

set them apart.[22]

4.1. Interactive voice response (IVR)

IVR systems are one of the oldest VIs in public use. These do not require a smartphone and are still used in several domains. Corkey and Parkinson (2002) define IVR as "a telephone interviewing technique in which the human speaker is replaced by a high-quality recorded interactive script to which the respondent provides answers by pressing the keys of a touch telephone (touch-phone)."[23] The recorded scripts used single voices,[24] combinations of male and female voices,[25] combinations of many female voices speaking in different languages,[26] or synthetic voices.[27]

4.2. Chatbots

The terms voice bots, chatbots, and automated conversational interfaces are used synonymously. They are enhanced by AI, NLP, and machine learning.[28] The term 'voice bot' is shorthand for 'voice robot'.[29] Here, voice is the primary medium of input.[30] They use automated speech recognition (ASR) technology to convert input into text. 'Chatbot' has a wider connotation, as it allows people to provide inputs in the form of text, gesture, touch, and voice. In this section, we use the term chatbot in the context of voice-enabled chatbots. The chatbot's output may be in the form of written text or voice, for which it uses text-to-speech (TTS) technology.[31] Voice chatbots can be further classified into two major categories: task-oriented (declarative) chatbots and data-driven (predictive or conversational) chatbots.[32]

22 Ghanchi, J., "Chatbots vs Virtual Assistants: Right Solution for Customer Engagement," *Medium*, 22 October 2019, accessed on 4 January 2022, **https://chatbotsjournal.com/chatbots-vs-virtual-assistants-right-solution-for-customer-engagement-17fd1b06f152**; Joshi, N. (2018, December 23). "Yes, Chatbots and Virtual Assistants are Different!" *Forbes*, 23 December 2018, **https://www.forbes.com/sites/cognitiveworld/2018/12/23/yes-chatbots-and-virtual-assistants-are-different/#6b41450b6d7d**

23 Corkrey, R., Parkinson, L., "Interactive Voice Response: Review of Studies 1989–2000," *Behavior Research Methods, Instruments, & Computers* 34 (2002): 342–353, **https://doi.org/10.3758/BF03195462**.

24 Piette, J. D., Weinberger, M., and McPhee, S. J., "The Effect of Automated Calls with Telephone Nurse Follow-Up on Patient-Centered Outcomes of Diabetes Care: A Randomized, Controlled Trial," *Medical Care* 38 (2000): 218–230.

25 Baer, L., Jacobs, D. G., Cukor, P., O'Laughlen, J., Coyle, J. T., and Magruder, K. M., "Automated Telephone Screening Survey for Depression," *Journal of the American Medical Association*, 273 (1995): 1943–1944.

26 Tanke, E. D., and Leirer, V. O., "Automated Telephone Reminders in Tuberculosis Care," *Medical Care* 32 (1994): 380–389.

27 Meneghini, L. F., Albisser, A. M., Goldberg, R. B., and Mintz, D. H., "An Electronic Case Manager for Diabetes Control," *Diabetes Care* 21 (1998): 591–596.

28 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

29 Middlebrook, S., and Muller, J. "Thoughts on Bots: The Emerging Law of Electronic Agents," *The Business Lawyer* 56, no. 1(2000): 341–373, accessed on 12 June 2020, **www.jstor.org/stable/40687980**

30 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

31 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

32 "What Is a Chatbot?", *Oracle*, accessed on 21 June 2020, **https://www.oracle.com/solutions/chatbots/what-is-a-chatbot/**.

a. Task-oriented chatbots

Task-oriented chatbots, also referred to as 'linguistic-based' or 'rule-based' chatbots, are devices that employ VIs that focus on a single purpose.[33] Due to this characteristic, they are considered to lack flexibility of functionality. They generate automated, conversational responses using NLP and logic. The functions of these chatbots are fairly limited, and hence they are used for specific purposes. A common example of these chatbots is interactive FAQs.

b. Data-driven chatbots

Data-driven chatbots, also known as machine-learning or AI chatbots,[34] are enhanced with AI, NLP, NLU, and machine learning to deliver personalised and meaningful responses. They are considered more interactive and contextually aware than rule-based chatbots, as their functioning is more complex and predictive.[35] This is because they learn the individual preferences and consequently create a profile of the person based on the data received.

Some refer to these types of bots as 'virtual assistants'.[36] However, other literature argues that these bots can be distinguished from virtual assistants.

4.3. Virtual assistants

According to scholars, there is no standardised definition of virtual personal assistants.[37] They list several names that other scholars have given to these systems, such as virtual assistants; vocal social agents or digital assistants; voice assistants; intelligent agents; and interactive personal assistants. Virtual assistants such as Siri use the speaker's voice and content and process it to respond in different contexts, like tasks to be performed or an action directed towards the person.[38] Virtual assistants are now increasingly used in several areas of everyday life; some common names are Siri, Google Now, Microsoft Cortana, Amazon Echo, and Google Home. These assistants interact with people in a conversational manner, thereby providing them with a wide range of functionalities.[39]

The conundrum in using the terms 'chatbot' and 'virtual assistant' interchangeably comes from the lack of universally accepted definitions. Some opine that they come under the umbrella term 'chatbots', and, in specific, 'data-driven chatbots'; the opposing view is that a virtual assistant is a completely different branch in the typology of VIs. These dissenting approaches come about because chatbots

33 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

34 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

35 "What Is a Chatbot?", *Oracle*.

36 "What Is a Chatbot?", *Oracle*.

37 Timo Strohmann, et al., "Virtual Moderation Assistance: Creating Design Guidelines for Virtual Assistants Supporting Creative Workshops", *PACIS 2018 Proceedings*, no. 80 (2018), **https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1079&context=pacis2018**

38 Sirbi, K., Patankar, A. J., "Personal Assistant with Voice Recognition Intelligence", *International Journal of Engineering Research and Technology* 10, no. 1 (2017): 416–419.

39 Breen, et al, "Voice in the User Interface.".

are characterised as data-obtaining interfaces.[40] In contrast, 'virtual assistant' is a distinct classification, as it is considered better than a chatbot with respect to understanding the context and the request, proficiency, nature of responses, and the rendering of a personalised experience.[41]

5. The future of VIs

VIs are slowly becoming more accessible as they are being integrated into cheaper mobile phones. The next stage is the development of smart devices for homes that can work with voice assistants, such as Google Home and Amazon Echo. On the business side, voice bots could be used for more complex customer questions. Interestingly, researchers have now also built a prototype linked with Alexa, to provide farmers with a 'smart irrigation voice assistant'.[42] Similarly, a voice application named 'Avaaj Otalo' was launched by UC Berkeley School of Information, Stanford HCI Group, IBM India Research Laboratory and Development Support Center (DSC), an NGO in Gujarat, to help farmers with agriculture-related queries.[43] Lastly, another significant use of VIs, according to Joshi and Patki (2015), is in increasing the safety of the computer system. Passwords set for systems via keyboards can be duplicated. However, when it comes to securing systems via VIs, duplication becomes far more difficult.[44]

6. Conclusion

The reduction in smartphone prices and data, as well as the increase in the functions that they can perform, have enabled the integration of VIs far more complex than IVR systems. One can hope that with further data and research, there will be an increase in not just their variety, but also in their ability to communicate with people who speak different languages.

40 Ghanchi, "Chatbots vs Virtual Assistants."

41 Joshi, "Yes, Chatbots and Virtual Assistants Are Different!"

42 Ramakrishnan, V., "How Mindmeld Is Used to Conserve Agricultural Water (... and Win Hackathons in the Process)", 2019, accessed on 2 November 2021, https://www.mindmeld.com/20190828-how-mindmeld-is-used-to-conserve-agricultural-water.html

43 Patel, N., et al., "Avaaj Otalo – A Field Study of an Interactive Voice Forum for Small Farmers in Rural India." Conference on Human Factors in Computing Systems – Proceedings 2 (2010): 733–742, 10.1145/1753326.1753434.

44 Joshi, P. and Patki, R., "Voice User Interface Using Hidden Markov Model for Word Formation," *International Journal of Computer Science and Mobile Computing* 4, no. 3 (2015): 720-724, https://ijcsmc.com/docs/papers/March2015/V4I3201599a81.pdf.