



VoiceBox

Interacting with Intelligence

Voice Interface for the Internet of Things

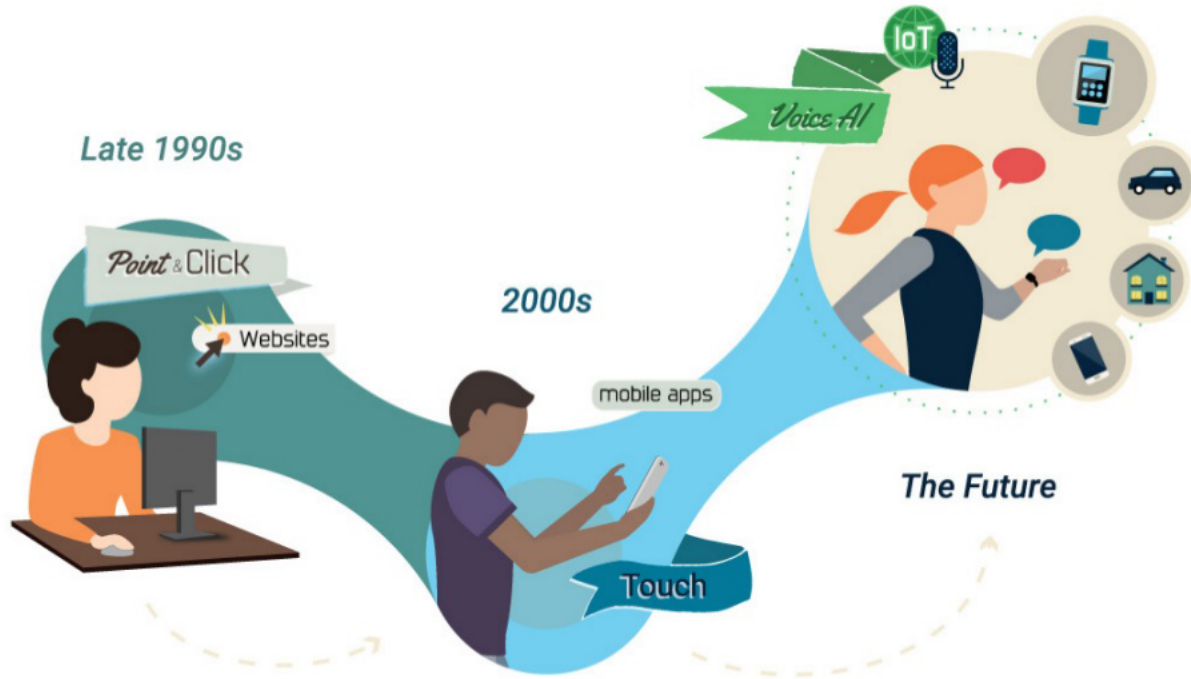
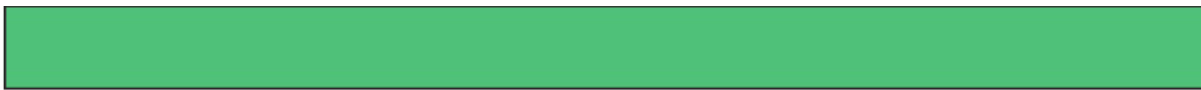
The internet is changing...again. Just as typing and clicking were replaced with tapping and swiping, our voices will soon replace our hands as the principle means of connection. This interaction won't resemble the crude spoken commands of the past where specific phrases activated a limited set of functions. The new voice interface will understand *you*, the way *you* speak, the phrases *you* choose, and the intentions *you* have. It will anticipate and collaborate. And it will be wherever you need it—on your computer, your phone, throughout your home, and in your car.

Mike Kennewick,
President
Voicebox

Dr. Philip Cohen
Sr. VP of Technology
Voicebox

As an ever-widening array of products—from thermostats and lightbulbs to dishwashers and automobiles—are connected to the “Internet of Things,” consumers will demand a natural, convenient means to control them. You might ask your alarm clock to play a digest of the news, instruct your coffee maker when to brew a new pot, or tell your car how warm to heat the driver’s seat.

This white paper presents an overview of the changes to come, the progress thus far, and the technical challenges that remain.




Voice AI

A seamless voice-interface will require a much higher level of voice recognition and understanding than is currently available. We call this capability Voice Artificial Intelligence, or Voice AI. People will use Voice AI to make more complicated requests, receive more precise answers, and relate to digital assistants that can anticipate their needs and intentions.

Voice AI will combine the advanced understanding of spoken words and phrases with visual information. Improved microphone arrays and imaging systems will process speech as well as the body language, mannerisms, and expressions that provide context to conversation. These semi-intelligent systems will learn each user's distinctive voice, personality, and preferences. In doing so, the computers of science-fiction's past will finally become a reality

Voice AI and the Internet of Things

The devices we use every day are increasingly connected to the internet. Washing machines can remind you to order more detergent...or do it for you. Cars can send you a message when they need an oil change. You can tell your oven to turn on as you drive home from work. Many of these interactions will be accomplished through speech.



To effectively control such a complex array of devices, Voice AI will have to understand what people mean beyond the literal interpretation of the words they say. "Can you get out of the car?" is often a request to exit a vehicle, although it could be a true question if, for example, the car was parked in a tight spot.

Voice AI must be able to decipher such ambiguities in word-choice and sentence structure—what's called natural language understanding. This will require a large and ever-growing knowledge base of phrases, and a statistical association that links them to their common meaning.

Such a system will also need to recognize and respond to a person's varied emotional states such as happy, sad, angry, or nervous. It will need to understand how people interact in groups and their personal relationships (Bob is Jim's father and is angry, Hannah is Jim's wife and is supportive). Only by understanding human beliefs, desires, and intentions can Voice AI serve as a collaborative assistant, one that understands what was said and responds in a manner the user expects.

In practical terms, a true digital assistant should be continuously present and available across a variety of platforms (computer, phone, car, home). It must know where and how people are accessing the system, and respond within that particular platform's capabilities. Information that could be presented visually on a screen device would have to be presented vocally on a screenless device like a smart speaker.


The system would need to know when they are being addressed, when a response is appropriate, and when to offer suggestions proactively. All this is easy to say. But how do we go about creating it? The challenge of Voice AI can be approached with three innovative technologies—semantic parsing, knowledge, and reasoning.

Semantic Parsing

Semantic parsing is a form of translation. Spoken phrases and sentences are translated from their literal words to their actual meaning. The result is an accurate "meaning representation" of what was said. This process of translation allows any system to understand more complex speech patterns.

As an example, to translate or "parse" the question, "What is the second most populous state that borders Ohio?" the system must:

--understand it is looking for a set of states.



--translate "most populous" into an expression that creates a sorted list of states based on population.

--limit this list to states that border Ohio.

--and finally, choose the second entry from this list.

By translating the question into a logical form, and then executing that form against a database, a semantic parser will be able to provide the answer as "Michigan."

Such a system can answer complex questions such as: "What is the capital and population of the states that border both Ohio and New York?"¹ But the benefits of semantic parsing are not limited to geography. Someone searching for a restaurant could consult a Point-of-Interest database with the question, "Find me the best French restaurant within walking distance of Key Arena that has off-street parking."

Semantic parsing is not a new idea², but no system is commercially available that will truly understand utterances by translating them to accurate meaning representations. And no system currently benefits from the other critical elements of Voice AI: Knowledge Graphs and Deep Neural Networks.

Knowledge

A capable semantic parser, able to create meaning representations using natural language understanding, can provide extremely accurate responses when paired with well-structured database. Human beings, however, often speak in ways that are not grammatically correct, and often data is only available from poorly organized sources.


In the case of unstructured data, statistical language processing helps, but must derive a meaning representation

answers must be found using knowledge bases or knowledge graphs³. Such resources contain both factual world knowledge and the ontological information used to organize that data.

1-The answer is Harrisburg, Pennsylvania, whose population is 49,188 (in 2013).

2-See Pereira, & Warren (1982) and Woods (1972).

3-The term "knowledge graph" has become commonplace, but refers to an implementation method (i.e., as a graph). The term "knowledge base" refers to the collection of facts, rules, ontologies, etc. that may be encoded in a graph.



The main difficulty in utilizing such resources is the wide range of methods used to structure them. To solve this problem, Voice AI could translate an utterance with the same semantic parser used to encode the sentences in the data source and organize a structured knowledge base. This requires a system that can learn the relationship between question and answer regardless of the specific meaning representation of the question.

In addition to large scale knowledge resources, a digital assistant may need to access personal information. A user's individual data—members, friends, coworkers, devices, daily activities, likes and dislikes, messages and social networks—would have to be handled securely and only with permission. If such access is granted, however, a digital assistant could over time learn a user's interests and preferences, and provide a more individually customized service.

Voice AI should also learn by being told. One should be able to tell an assistant, "I don't like XYZ airlines," and the assistant should avoid scheduling trips on that airline. A user preference such as, "I don't want text messages when I am in a meeting," should also be learned and respected. In this way, a digital assistant should be "advisable"⁴, in addition to learning about people from their daily activities.

Reasoning

People engage in conversation as active participants, striving to understand not only the words being spoken, but the plans and intentions behind them. Each participant's job is to be a collaborative conversant, to recognize plans and intention, and to help reach a successful conclusion.

For example, when faced with a kitchen-disaster, a person may exclaim, "The rice is boiling over!" They say this not merely to state the obviously, but as a plea for assistance. In normal interactions, people infer what the other person wants—in this case, to turn down the heat or remove the rice from the stove—and act accordingly.⁵

The Voice AI system of the future must engage in *collaborative dialogue*. Plan recognition and inference are essential abilities for any digital assistant and have widespread commercial applications. Consider the following interaction that takes place after a user has heard a song they enjoy:

4-See McCarthy, J. (1958)



"What band is that?"

"The group is Coldplay."



"When is their next concert?"

"Their next concert is in Seattle on March 25. Would you like me to get you tickets?"



"Yes."

"Sorry, the March 25 concert is sold out. There is a concert on March 24 in Portland. Would you like to go to that one?"



"Sure."


"Ticket prices range from 35 to 100 dollars. Here is the seating chart."



In this example, the assistant understands the user's goal of attending a Coldplay concert. This plan, however, has a precondition—the availability of tickets.

When the assistant learns the closest concert has been sold out, they not only explain the situation, but provide an alternative—the nearest concert for which tickets are still available. The resulting success of this plan recognition is the screen in Figure 2.

If the user purchases tickets, the assistant could collaborate even further and offer transportation and lodging suggestions. If the user declined to purchase tickets, the assistant could provide information on the next time Coldplay is scheduled to perform in the user's preferred venue.



Although this example may appear simple, no commercial system available today can perform such inferences and act on the user's behalf to facilitate their goals and plans. A primary capability of Voice AI must be plan recognition and, more generally speaking, collaboration. Human dialogue is a special case of collaborative action, with all participants jointly committed to understanding one another.¹

This shared responsibility can be seen in subtle “back channel” feedback (head nods, “uh-huh”, etc.) the listener performs to ensure the speaker that their communication has been received and understood. The absence or delay in such feedback, for example due to bad phone reception or a satellite delay, can result in a communications break-down. Voice AI systems that aspire to natural dialogue, rather than simple question-answering interactions, will need to both recognize and generate the conversational behaviors that people expect.

Summary

The next generation interface to the Internet of Things will be powered by Voice AI. This semi-sentient, collaborative assistant will understand the meanings of what is said through semantic parsing and the use of large-scale knowledge resources. Voice AI will infer the plans behind the utterances, and take action to facilitate those plans. Given the unprecedented ability to collect data about people, activities, business, etc., this type of plan-based reasoning will be increasingly common. In a commercial context, such plan-based interpretation of language will provide the basis for a much more powerful and satisfying type of interaction, one that can support goal fulfillment such as voice commerce.



References:

Allen, J. F., A plan-based approach to speech-act recognition, PhD Thesis, Dept. of Computer Science, University of Toronto, 1979

Allen, J. F. and Perrault, C. R., Analyzing intention in utterances, *Artificial intelligence* 15(3), 1980, pp. 143-178,

Clark, H. H. and Wilkes-Gibbs, D., Referring as a collaborative process, *Cognition* 22(1), 1-39.

Cohen, P. R., On knowing what to say: Planning speech acts, PhD Thesis, Dept. of Computer Science, University of Toronto, 1978

Cohen, P. R. & Perrault, C. R., Elements of a plan-based theory of speech acts, *Cognitive Science*, 3(3), 1979, pp. 177-212. Reprinted in *Readings in Artificial Intelligence*, Nilsson, N., Grosz, B. J., & Nash-Webber, B., Morgan Kaufmann Publishers, 1981

Cohen, P. R. & Levesque, H. J., Teamwork, *Noûs*, 25(4), 1991, pp. 487-512. Reprinted in *Knowing, Reasoning, and Acting: Essays in Honor of Hector J. Levesque* volume 16, pp. 137-156

Cohen, P. R., Perrault, C. R., & Allen, J. F., Beyond question answering, in *Strategies in Natural Language Processing*, Ringle, M. (ed.), Lawrence Erlbaum Publishers, 1982, pp. 246-274

Grosz, B. J. & Sidner, C. L., Plans for Discourse, in *Intentions in Communication*, Cohen, P. R., Morgan, J., Pollack, M. E. (eds.), MIT Press, Cambridge:MA, 1990.

McCarthy, J. Programs with common sense, Symposium on Mechanization of Thought Processes. National Physical Laboratory, Teddington, England, 1958.

Pereira, F. C. N. & Warren, D. H. D., An efficient easily adaptable system for interpreting natural language queries, *Computational Linguistics* 8 (3-4), 1982, pp. 110-122.

Rich, C., and Sidner, C. L., COLLAGEN: When agents collaborate with people, *Proceedings of the First International Conference on Autonomous agents*, 1997, pp. 284-291.

Schmidt, C.F., Sridharan, N.S., and Goodson, J.L. "The plan recognition problem: An intersection of Psychology and Artificial Intelligence." *Artificial Intelligence*, 11, 1978, 45-83.

Woods, W. A., Kaplan, R. M. and Nash-Webber, B. L. The lunar sciences natural language information system: Final report. BBN Report No. 2378, Bolt Beranek and Newman Inc., Cambridge, MA. 1972. Available from NTIS as N72-28984.