2022.10-31
Version 1.1

Interoperability of Conversational Assistants

A NEW APPROACH

The Open Voice Network
Architecture Work Group of the Technical Committee
October 31, 2022

Executive Summary

The Open Voice Network (OVON) is an open-source community of the Linux Foundation, dedicated to developing the technical standards and usage guidelines for the emerging world of conversational Artificial Intelligence and, within that, voice assistance.

The future of conversational Artificial Intelligence is a future of diversity – of infrastructure providers and platforms, of enabling technologies, devices, and enterprise use cases. Conversational AI – and within conversational AI, voice assistance -- will be a primary interface to the internet, personal transportation, smart environments (domestic and enterprise), and the immersive digital world of gaming and augmented and virtual experiences.

To reap the greatest economic and social value of conversational assistance, we believe that voice must work like the web, enabling users to access any voice-enabled content destination regardless of platform. Interoperability – the sharing of dialogs between conversational assistants and agents of different technological parentage – is an essential capability to reach the value of what will become a Worldwide Voice Web. It must also be worthy of individual and enterprise user trust; readers will note references here to OVON work in data privacy, data security, and ethical use.

This paper is a step on the path to conversational assistance interoperability. In this paper, we assert that the proper model for conversational assistance interoperability is a human one – and suggest that humans generally resolve questions with one another through a process of either ***mediation*** or ***delegation.*** Using this framework, we identify the jobs to be done for priority constituencies, the architectural patterns that must be resolved, lessons learned from other work, and a path forward to an ecosystem of standards-based OVONICA – Open Voice Network Interoperable Conversational Agents.

ABSTRACT

This is a publication of the Open Voice Network (OVON, www.openvoicenetwork.org), a non-profit industry association operating as an open-source community of the Linux Foundation. It asserts that, to realize the full economic and societal potential of conversational assistance, conversational assistants must not only ***mediate*** human-to- assistant conversations – e.g., host the conversation and obtain relevant information to fulfill user intents – but also ***delegate*** conversations to other assistants, and in doing so, pass textual, acoustic and contextual data, as well as privacy and security controls.
To meet this vision, the Open Voice Network proposes an approach to **interoperability *between* conversational assistants** – specifically, the sharing of multi-layered dialogs between assistants and assistants of differing infrastructures through the standardization of communication protocols by which autonomous assistants and assistants collaborate to achieve a common goal.

This paper also suggests areas for further research, as well as next steps for the proposal and testing of communication protocols that may be built from existing technologies and universally accepted standards.

At the core of this work is this firm belief: we are in the early days of conversational assistance, and the future will be one of stunning diversity -- a multiplicity of voice-enabled user end points, content and infrastructure providers, voice-enabled destinations (measured, perhaps, in the billions), industry ecosystems such as transportation and smart homes, and organizational/enterprise value propositions.

This paper is the creation of the Architecture Work Group of the Technical Committee of the Open Voice Network. The Open Voice Network is a trademark of The Linux Foundation. Other trademarks referenced in this report are the property of their respective owners.

This version is a revision of the original publication of this document, version 1.0. It has benefited from numerous comments from internal and external reviewers, for which we are very grateful. We welcome additional comments, which can be sent to our public email list, whitepapers@lists.openvoicenetwork.org.

OPEN VOICE NETWORK
Voice worthy of user trust

TABLE OF CONTENTS

PREFACE

The Open Voice Network (OVON) was founded by *users* of conversational assistance. Its purpose is to develop and drive adoption of technical standards and usage guidelines that will make the emerging world of conversational assistance **worthy of user trust**.

This document represents the work of but one of several OVON technical work groups – all under the aegis of the Open Voice Network Technical Committee. Separate, yet related initiatives are at present exploring issues of data protection and data security for conversational assistance, assistant-agent destination and location services, voice-centric authentication, and synthetic voice. Other envisioned work streams await resources.

To date, the work has been developed on parallel tracks. As we go forward into Q4 2022, the streams will begin to merge – informing and being informed by each other.

We look forward to collaborating with you.

SECTION ONE: RATIONALE, SCOPE, AND DEFINITION

1.0 Introduction

This section speaks to the Open Voice Network's reasons for pursuing interoperability of conversational assistants, the boundaries of our work, and a 1.0 definition of conversational assistant interoperability. Note that, for the purposes of this paper, the terms "assistant" and "agent" both refer to conversational AI user interfaces that are perceived to be a single conversational actor, operate on behalf of the user, and operate in accord with a conversational

platform. More detail about the difference between assistants and agents will be provided in later sections.

1.1 Intended Readership

The first version of this White Paper was published for public evaluation and criticism on 12 August 2022. This version represents a revision based on reviewer comments.
It seeks to address three global audiences, each with a stake in the future of voice:

- Enterprise content creators and communicators (business-to-consumer and business-to-business)
- Stakeholders and decision makers within the voice technology industry, including potential architectural partners
- Current and prospective participants and sponsors of the Open Voice Network.

1.2 Purpose of this White Paper

This White Paper seeks to:

- Identify the architectural elements required to facilitate interoperability of conversational assistants
- Prioritize these elements for standardization by the OVON Architecture Work Group
- Foster discussion with the constituents noted above in section 1.1.

In this White Paper, we seek to remain neutral with respect to competing architectural directions, and to reveal the underlying concepts/issues that exist regardless of architectural choice.

This document is intended to be read alongside the Open Voice Network Technical Master Plan. In the short term these two documents may contradict one another as ideas are developed and tested. Such conflicts will be resolved as they emerge.

This paper replaces a previous document entitled 'Architecture Design.'

1.3 The Open Voice Network and Interoperability of Conversational Assistance: Why and What

The Open Voice Network (www.openvoicenetwork.org) is an open-source community of the Linux Foundation dedicated to the development of the standards and usage guidelines for the emerging world of voice assistance.

It enjoys the regular participation of more than 200 volunteers from 13 nations and 5 continents. The OVON community includes participants from leading voice platforms and infrastructure providers; large enterprises that use voice technologies in customer service and operations (from industries such as retail, healthcare, telecommunications, and financial services); marketing and consulting firms; and more than 40 independent voice development and services companies.

As a technology-neutral, nonprofit organization, the Open Voice Network occupies a unique and strategic position within the voice technology industry. Our sponsors and participants witness a growing diversity of underlying voice technology (speech recognition, language understanding, dialog modeling), conversational design paradigms, labels to represent semantic content, and endpoints (from smart speakers to voice-enabled web pages and immersive games). They also see the rapid growth and evolution of voice value propositions, especially for organizations and across vertical industries.

This diversity points to a significant total available market (TAM) for conversational AI as both an interface and a source for data-fueled insights. Looming before us – as first envisioned by Dr. Monica Lam and colleagues at Stanford University – is an *interoperable* Worldwide Voice Web (WWVW) (Lam et al., 2021)., where the spoken word offers an open, standards-based interface to billions of voice-enabled media, enterprise, website, transportation, smart IOT environment, and metaverse destinations.

Given the economic and societal value that will be created by an interoperable Worldwide Voice Web, the Open Voice Network commissioned the Architecture Work Group of its Technical Committee to research and recommend architectural options for open, standards-based interoperability for conversational assistance.

Using a technologically and architecturally neutral eye, the Architecture Work Group reviewed existing technology standards and existing and proposed voice architectures and messaging protocols. The Work Group ascribed to the *Harvard Business Review's* "Jobs to Be Done" methodology (Christensen et al., 2016)., and examined in detail the current and future needs of four key voice constituents:

- Consumers of voice-enabled experiences
- Enterprise content providers
- Technical innovators of dialog systems
- Voice system infrastructure providers.

Through these efforts, the work group identified the opportunities and challenges of conversational assistant interoperability, and a series of foundational concepts.

We are deeply grateful for the contributions of the many Open Voice Network participants and sponsors referenced on page 55. These individuals gave freely of their time and intellect to develop this paper and the plans to take it forward.

The Open Voice Network:

- Believes that conversational assistance will realize its economic and societal value when it is interoperable, like telephony or the worldwide web (WWW).
- Sees in today's rapidly growing enterprise investment (Frost & Sullivan, 2022) in conversational assistance the seeds of a rich ecosystem of independent, purpose- and brand-specific voice assistants, an ecosystem that will bring desired content and experiences to the users of general-purpose, proprietary conversational assistants.

- Recommends the development and industry-wide adoption of standardized communication protocols between conversational assistants operating on different platforms.

- Asserts that conversational assistants must both *mediate* dialogs (e.g., act as a host, and acquire desired information to fulfill the user intent) and *delegate* dialogs (e.g., act as an initiator of communication, and bridge to other assistants so that the user intent can be fulfilled.) In addition, conversational assistants must also be identified and authenticated to serve as a *destination* of a delegating assistant or agent.

- Envisions a model of conversational assistance interoperability in which autonomous assistants and agents *collaborate to achieve a goal together*. In so doing, they will use standardized ways to share information with each other and operate at various, negotiated levels of trust and information sharing. This will lead to the development, test, and proposal of standards for:
    - The way in which a spoken assistant name can be used to find an associated assistant. OVON is neutral on how users indicate with which voice assistant they wish to interact. The Speech Recognition/NLU of the host agent must be able to detect when the user wants to connect to a different conversational assistant and which conversational assistant the user wishes to connect with.
    - The way in which basic linguistic information is shared between dialog assistants
    - The way in which immediate linguistic context and history is shared between dialog assistants
    - The way in which control is handed over between dialog assistants
    - The way in which dialog assistants negotiate trust.

In addition, the Open Voice Network

- Will neither develop nor recommend the standardization of platform components or the format of content used to configure these components.
    - We do not believe that a standard methodology for expressing and describing conversational interaction is possible or desirable. Attempts to describe how conversations can or should be modeled will quickly become outdated as conversational interaction innovates and evolves.

In the coming months, the Open Voice Network will develop, test, demonstrate, and propose to existing standards bodies a set of standardized communication protocols that will enable assistant-to- assistant voice interoperability.

1.4 Going Forward: Your Participation

In keeping with the practices of the Linux Foundation, the Open Voice Network works in an open, communal manner. We seek contributions from every corner and region of the conversational AI ecosystem.

Readers of this paper are invited to

- Propose corrections and additions to this paper through the technical mailing list whitepapers@lists.openvoicenetwork.org. We are especially interested in criticism and comment, regarding:
    - Concepts that are not clearly defined
    - Requirements that you feel are missing
    - Requirements that you feel are unnecessary
- Contribute usage scenarios (similar to those in the paper) that challenge the concepts within this paper, and perhaps point us to new or revised requirements for the interoperability of conversational assistance.

You are also welcome to join the weekly interoperability development meetings of the Architecture Work Group of the Open Voice Network's Technical Committee. One-hour sessions are scheduled Tuesdays at 17:00 CET, 11:00 Eastern, and 08:00 Pacific. Conferencing details are found at https://openvoicenetwork.org/calendar.

1.5 A User's Vision: An Interoperable World for Today and Tomorrow

Today: Pat Shops by Voice as Humans Do

Pat has finished the day's labors and is driving home. It's late. A request to Pat's automobile voice agent starts a music playlist that brings back memories of school days – which, with a start, reminds Pat that a long-time friend from school, now in town for a conference, will be coming to dinner tomorrow.

What to prepare? Perhaps a favorite seafood stew. Crab, mussels, shrimp, snapper in a tomato broth with wine? Pat asks the automobile-based voice agent to connect to the grocer that knows Pat's preferences – and, in an instant, Pat is ***directly connected****, and speaking to the grocer's conversational assistant.*

A recipe is identified. A shopping list is created -- items, quantities – based upon availability of overnight shipments from seafood suppliers. Wines are recommended (a Zinfandel and Sauvignon Blanc) and chosen.

Pat sighs with relief. And then – right before completed list is transferred to Pat's mobile grocer app for authentication, pick-up scheduling, and payment -- Pat remembers that the school friend has a chronic medical condition (in this case, diabetes.) Pat quickly asks the grocery assistant if the seafood stew will be healthy or unhealthy. The assistant, in turns, informs Pat that it is ***directly connecting*** *to a conversational capability developed specifically to recommend food choices for individuals with diabetes.*

From the conversational capability, Pat learns quickly that the seafood stew will be a healthy choice for dinner with the school friend. A quick request to the conversational capability, and Pat is sent back to the grocery assistant – and the shopping list is confirmed.

Tomorrow: Pat Plans Travel by Voice as Humans Do

Pat plans to travel to an international conference and stay for the weekend that follows. Pat needs a visa, airline reservations, and information about attractions for anticipated free time. Pat's personal conversational agent is named "Butler," and was developed by a third party using a mix of proprietary and open technologies.

Pat begins trip planning by pressing a button on a smartphone to begin conversation with Butler. Butler is local to Pat's phone and has permission to access Pat's personal data. Butler invokes a voice passport authentication capability to verify that the speaker is Pat and uses Pat's smartphone for second-factor authentication.

Pat informs Butler of the conference in a major city on another continent. Pat has been asked to deliver a keynote address. A visa may be required to attend. Butler uses a voice destination and location service (similar in purpose to a Domain Name System) to discover the conversational assistant of the passport-visa office of the destination country.

The visa conversational assistant provides Butler with the latest visa requirements and guidance on how (and when) to apply; Butler forwards the guidance and application URL to Pat's digital account and schedules a reminder to ensure its submission. At the same time, Butler has explored flight options by searching among highly rated travel services.

Butler selects AirWithFlair, a specialty conversational agent and capability for artists and musicians., and mediates a connection to obtain a list of travel options (and prices) for Pat. Butler - knowing that Pat is a loyal customer of the Royalty Hotel chain - delegates the dialog to the Royalty conversational assistant, which recognizes Pat and immediately identifies a location, room type, and pillow choice for the potential conference visit.

1.6 Architectural Aspirations

The following architectural aspirations guide our thinking and our actions:

1. We recognize that the material and financial resources required to create, operate, and maintain conversational AI and voice assistant systems are significant. Our architectural direction will allow for sharing the cost of those resources, but we will not ourselves specify how that needs to occur.
2. There are several competing architectures for control interfaces between conversational AI platforms, assistants, and agents. OVON will evaluate architectures, produce recommendations, and remain neutral with respect to the choice until and unless making a choice becomes required for completing our specifications.
3. Each of the dominant proprietary platforms within the existing voice market has an installed base of platform-specific applications. At the same time an ecosystem of third-party voice content creators has emerged outside the proprietary platforms. To maximize the value of our work and speed global adoption, we will seek to establish a platform-neutral approach to interoperability – while adopting (as noted above) common concepts and ideas.
4. The lifespan of a standard is much longer than the lifespan of the physical components subject to the standard. The specifications and guidance which we will produce may be informed by existing design patterns but must not be limited to those patterns.
5. In the immediate future, we aspire to:
    a. Outline the different architectural approaches that could support interoperability.
    b. Outline the advantages and disadvantages for each with respect to the OVON constituents and the Jobs to Be Done.
    c. Recognize that several viable architectural approaches to interoperability may coexist and that OVON needs to remain responsive to the directions taken by the market.
    d. Identify platform-neutral standardization that will enable both the new emerging third-party ecosystem and the applications resident on proprietary voice platforms.

6. Our architectural principles include the principle that, whatever happens, the human should be in charge. This includes the ability of the human to cancel the interaction at any time.

> *When you're exploring the internet, do you shut your browser down to open a new web page?*
>
> *Why, then, would you want to do that with voice?*
>
> Royal O'Brien, Linux Foundation, July 2022

1.7 Boundaries

The Open Voice Network defines the boundaries of this work using the "Jobs to Be Done" approach (Christensen et al., 2016)., singling out four groups of individuals for whom interoperability will create incremental and sustainable value. Identification and clarification of the needs and desired experiences of each group allows the OVON to define the work required from the Open Voice Network to fulfill those needs and desires.

Though Open Voice Network-delivered standards will benefit the voice industry at large, our work suggests that four distinct groups of individuals are central to the question of interoperability, shown in Figure 1.7: (1) consumers of enterprise content who wish to use voice enabled technologies, (2) enterprises that provide voice enabled experiences, (3) technical innovators of dialog systems who wish to create systems for enterprises to use, and (4) voice systems infrastructure providers who wish to create and pursue commercial opportunities within a free and open environment.

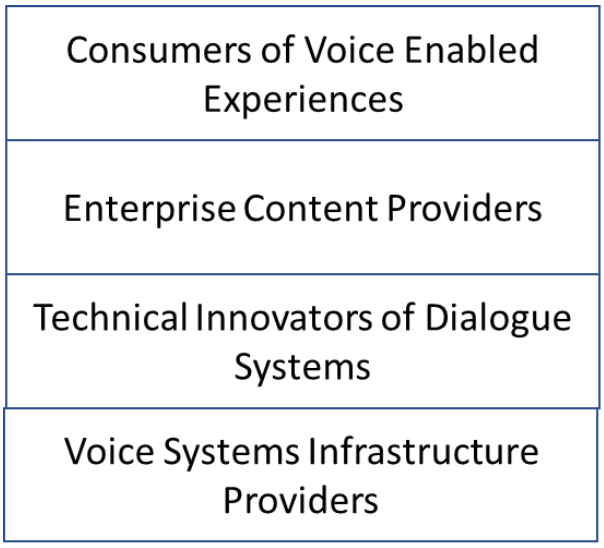


Figure 1.7 Benefactors of conversational assistant interoperability; the value ecosystem

Together, these groups constitute a *tiered ecosystem* that will flourish in a world of conversational assistant interoperability.

Consumers of Voice Enabled Experiences are at the top of the tier. These are the end consumers of voice enabled experiences. OVON-developed standards will enable them to engage in voice experiences in web environments that currently do not provide voice experiences. The "Jobs to be Done" that we are supporting for this group include the following tasks:

- *Help me continue my voice task as I switch voice devices (home, phone, auto)*
- *Make it possible for me to use all global voice platforms and web / IoT / metaverse-based voice-based destinations with a single voice application or service*
- *Enable me to give consent to sharing, storing, processing of my data.*
- *Help me avoid restrictions on vendor choice.*

Enterprise Content Providers rank next. These are the businesses and organizations that offer applications on proprietary, general-purpose consumer platforms, or do not deliver voice experiences to their customers outside the organizational security firewall. Together, enterprise

content providers represent a significant third-party content and voice-centric services ecosystem. “Jobs to be Done” for enterprise content providers are:

- *Make it possible for me to use all global voice platforms and web / IoT / metaverse-based voice-based destinations with a single voice application or service*
- *Help me grant all potential constituents free and unfettered access regardless of the constituent's "home" voice platform*
- *Help me ensure that all my potential constituents can freely and directly connect with my voice applications and services, regardless of home platform*
- *Help me be compliant with applicable data protection policies for all dialog data (text, acoustic, semantic) shared between my applications, services and constituents.*

Technical Innovators of Dialogue Systems are the technology-centric and consulting organizations that will create customer service, web, or operational environments for enterprise content providers. These innovators develop the different technology solutions that work together to enable voice experiences. They will build the pieces of the modularized future we envision. “Jobs to be Done” for technical innovators include:

- *Help me (re-)use voice-based core business processes for transactions*
- *Help me (re)-use existing messaging protocols for sharing of dialog, data and controls between platforms, agents and assistants.*
- *Help me give customers the choice to protect their voice data*
- *Help me secure my customer's voice data*
- *Help me (re-) use specifications /processes for voice authentication, attestation, and authorization*
- *Help me get to content and consumers without dependence upon proprietary platforms*

Of note: while there are “Jobs to Done” for each of the constituent groups – and each plays an important role in an interoperable ecosystem – all will rely on the work of the Technical Innovators. These individuals will use Open Voice Network standards to do their work; the OVON expects to provide significant support to assist them in accomplishing their jobs to be done, each of which require a specific experience. **These are the experiences we will create for the technical innovators.**

Technical Innovator Jobs to Be Done	Corresponding Technical Innovator Experiences
Help me continue my voice task as I switch voice devices (home, phone, auto)	Implement OVON standard protocols and specifications to enable business processes for interaction
Help me (re)-use existing messaging protocols for sharing of dialog, data and controls between platforms, agents and assistants.	Select and plug and play messaging protocols
Help me give customers the choice to protect their voice data	My customer elects to protect their voice data
Help me secure my customer's voice data	I match security level per customer's level of data protection
Help me (re-)use voice-based core business processes for transactions	I learn OVON Standards so that I know how it should work
Help me get to content and consumers without relying on proprietary technology	I know where to go and what to do to manage proprietary platforms

Voice Systems Infrastructure Providers provide technologies used by content creators and technical innovators to deliver voice-based content and services to consumers of voice-based experiences. In an open, standards-based, interoperable world, they will enable a stable voice ecosystem. Technical innovators will rely on these open infrastructure offerings to create voice technology solutions that will be used by enterprise content providers. “Jobs to be Done” that we are supporting for voice system infrastructure providers are the following:

- *Help me create an accessible environment*
- *Help me create an environment where seamless interaction is commonplace*
- *Help me create an environment where scalability is empowered*

Two additional jobs to be done were identified during working group consultation with a major infrastructure provider:

- *Help me maintain ownership of data*
- *Make sure that when my infrastructure is used (work is being done by me), that my work gets recognized.*

1.8 Interoperability Defined for Conversational Assistance

At the most basic level, *interoperability* means "the ability of two or more systems or components to exchange information and to use the information that has been exchanged." IEEE Standards Information Network/IEEE Press (2000). The OVON adapts this definition to the conversational assistance ecosystem as follows:

Interoperability is 1) the ability of conversational assistants of diverse parentage to collaborate seamlessly to achieve a goal together, using standard ways to share information with each other, operating at various levels of trust and information sharing, and operating independently of the physical devices being used; and 2) the ability of conversational AI-driven content, assistants, and service providers to reach target audiences across the diversity of platforms, general purpose assistants, and physical access devices (endpoints).

Conversational assistants can be fully or partially interoperable.

1.9 Core Requirements for Interoperability of Conversational Assistants

Conversational assistants must:

- Be enabled to fulfill user intents through mediated and delegated communication.
    - In ***mediated assistance***, the user interacts with a single conversational assistant. The mediating conversational assistant *hosts* the conversation; it never cedes control of the conversation. To fulfill user intent, the mediating assistant may obtain information from third party sources or introduce an application resident on its platform.
    - In ***delegated assistance***, the user interacts with one or more independent conversational assistants using standardized messaging protocols. The hosting conversational assistant initiates the conversation; it passes control to a delegate assistant (or assistants).

- Be enabled to serve three roles to fulfill user intents: as 1) **a *host*** of applications that **mediates** content to fulfill user intent, 2) ***an initiator*** of a connected dialogue that

delegates fulfillment of intent to one or more independent assistants, or 3) **a *destination*** that is delegated to, and as such, assists the initiator in fulfilling user intent.

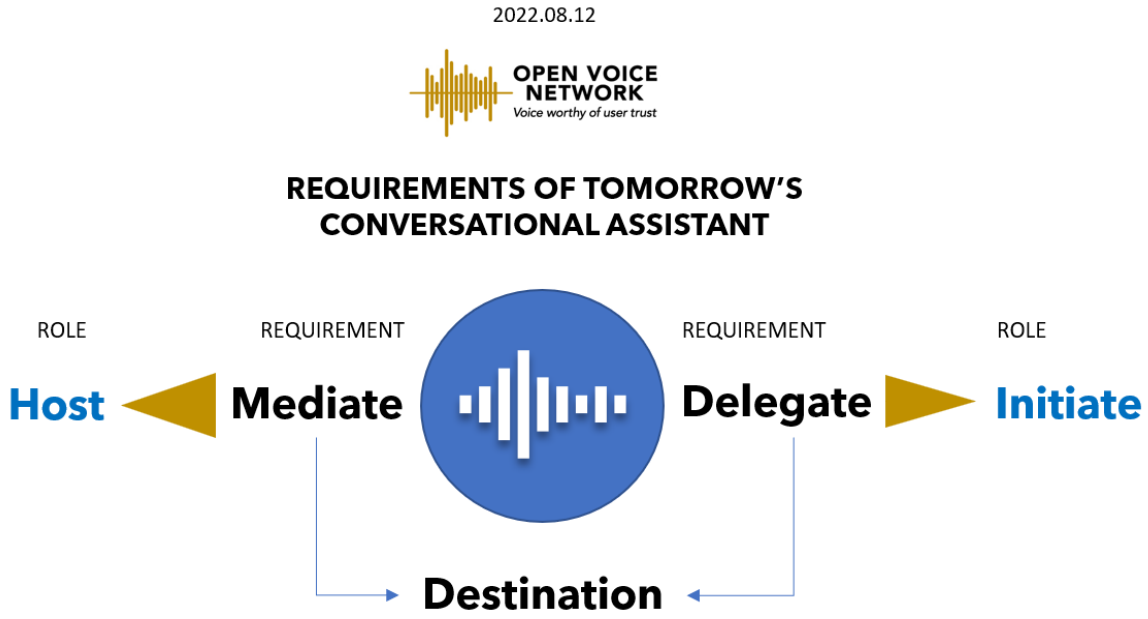


Figure 1.9: Core Requirements for Conversational Assistants

As an initiator and as a destination, an assistant must offer six types of interoperability, as shown in the grid below.

Type of Interoperability	Use	Example	Benefit
Share dialog fragments *Dialog fragment* = portion of a dialog pattern, a piece of interactive fragment, a chunk of interaction specification	Conversational assistant uses fragments from one or more other conversational assistants to achieve a dialog goal – user thinks is talking with one assistant	*A delivery service conversational assistant (with Pat's consent) uses the location collection dialog fragment from a payment conversational assistant to collect Pat's address information. The fragment may also be identified so it can be reused.*	Application developers save time by reusing dialog fragments from other conversational assistants.
Share data *Shared data* = information which the user wants to maintain across several sessions.	A conversational assistant specifies how data is shared with another conversational assistant.	The value in the Amount Due slot in the Shop conversational assistant is copied into the Payment Amount slot in the Payperson conversational assistant.	User controls how personal data can be shared with other conversational assistants, saving users from reentering the data value.

Share conversational context *Conversational context* = information which is shared between a user and a conversational assistant over a (tbd) period of time. With the user's permission, conversational context may be shared with other conversational assistants. Conversational context is a subset of personal data that includes data about the conversation itself, including the history of previous utterances in the conversation.	Relevant context (of the user, of the dialog) established on one assistant is shared with another conversational assistant.	The context (history) of actions in the Shop voice assistant is shared with the Payperson voice assistant. The Payperson conversational assistant is aware of information collected and generated by the Shop voice assistant.	Enables conversational continuity *conversational continuity* = the principle of making sure that all details in one voice assistant are consistent with the details in another voice assistant. "Conversational continuity" includes several concepts, (1) data integrity -- if I say I want to pay $50 to Target the value of "payment amount" slot stays $50 when the conversation is handed off to the next agent (2) "semantic integrity", so the next assistant gets a "payment amount" slot it knows what that is, even if it's native term for that slot is "amount to pay", (3) the agents' style of speaking, grammar, and tone of voice

Type of Interoperability	Use	Example	Benefit
Transfer control *Transfer control* = determine which conversational assistant is invoked and becomes active. Determination of which assistant has "the floor," which assistant has the permission to speak.	Transfer control from one conversational assistant to another	While shopping for groceries through a conversational assistant, Pat switches to the Payperson voice assistant to pay for the groceries.	Matches the natural thought processes of human users.
Share components (This is an aspirational goal) *Components* = components of a conversational assistant, including but not limited to ASR, TTS, NLP, NLG.	The developer specifies which components are used within a dialog.	A voice developer may desire to use sophisticated natural language processing rather than a simple natural language processor provided by the conversational assistant being used. Other examples of component sharing include (a) using different ASRs for speakers of American English and British English, (b) performing natural language translation between national languages such as English/German or English/Mandarin, (c) providing access to algorithms that perform explicit query or implicit query (a.k.a. voice search).	Developers can develop custom assistants through the use of standards-based, best-of-breed technologies.

Type of Interoperability	Use	Example	Benefit
Share endpoints *Endpoint* = hardware or software through which the user speaks and listens to conversational assistants	User controls where input may come from or output may be directed to.	The user may begin a request from a device with a public microphone/speaker and later request to continue with a private device. The user may ask for output to be directed to a different conversational assistant or a different device, such as a printer, or may ask for input to come from some shared data.	User can access any conversational assistant using any endpoint

1.10 Interoperability Intentions of the Open Voice Network

The Open Voice Network seeks to enable mediated and delegated communication by conversational assistants within an ecosystem of general purpose, consumer-centric conversational assistants, and independent, purpose- and brand-specific conversational assistants. To make this happen, OVON will develop messaging protocols and a destination management and services system, to standardize the way in which:

- a spoken assistant name can be used to find an associated assistant
- basic linguistic information is shared between dialog assistants
- immediate linguistic context and history is shared between dialog assistants
- control is handed between dialog assistants
- dialog assistants negotiate trust

SECTION TWO: A FOUNDATION FOR ANALYSIS -- THE TECHNOLOGY OF CONVERSATIONAL ASSISTANCE

2.0 Introduction

The Open Voice Network grounds its research into conversational assistant interoperability in current operative definitions of conversational Artificial Intelligence, as well as the normative components and architecture of contemporary voice assistance. In this section, intended for both subject matter expert and lay audiences, we share our understanding of those components and architectures.

2.1 Conversational AI and Voice

Voice assistance is but one form of conversational Artificial Intelligence, or conversational AI.

Citing *Interaction.com* (2022): "Conversational AI is the set of technologies behind automated messaging and speech-enabled applications that offer human-like interactions between computers and humans.

"Conversational AI can communicate like a human by recognizing speech and text, understanding intent, deciphering different languages, and responding in a way that mimics human conversation." (Haas, 2020).

Conversational AI solutions can be offered through text and voice modalities. Voice solutions – referred here broadly as *voice assistance* – can be delivered as an audio-only interface, or one in which voice assistance works in a complementary way with screen-based information. The latter is often described as *multi-modal* voice assistance.

This paper speaks primarily to the sharing of voice-based elements within a dialog. However, the Open Voice Network recognizes clearly the importance of multi-modal voice assistance (especially in an enterprise context) and will extend future editions of this work to multi-modal interoperability.

2.2 Conversational Assistants and Platforms

The Open Voice Network draws a clear distinction between conversational *assistants* and conversational *platforms.*

Conversational *assistants* function as conversational user interfaces. They engage in conversation and are identifiable by users as conversational actors. A conversational assistant typically has an assigned identity, an identifiable voice, and/or a name: for instance, Alexa, Siri, or Magenta assistants. The user recognizes a given conversational assistant as a singular entity and refers to that assistant with a pronoun (*he, she, it, they*) or by an assigned name. The conversational assistant may also refer to itself using a name or a pronoun.

Conversational assistants often respond to queries in the form of web searches. The response typically describes how the assistant locates the delivered content. For instance, one might ask, "Assistant, what is the weather like in Brooklyn, NY today?" To which the assistant responds, "Okay, here is something I found on weather.com . . . "

A conversational platform is the software to which the assistant connects, and that interprets the input from the user. The assistant responds through an endpoint, which is typically a piece of hardware, such as a smart home display, a television remote control, an automobile smart dashboard, a smartphone, or a smart speaker.

There are endpoints and platforms currently associated with proprietary interfaces, such as Amazon Alexa or Google Assistant. They handle wake words ("Hey, *Google*…!"), audio coding and streaming, and graphical events and content.

If a given platform provides generic access to content, it may have a host voice assistant, and give the user the option to ask for a *delegate assistant.* In other words, the host can provide a gateway to delegate content.

With the exception of assistants written in the VoiceXML standard (McGlashan et al., 2004)., delegate assistants are typically authored for a given platform by third party companies.

Examples of End Point-Platform-Assistant Pairings

End Point	Platform	Host Assistant
Echo Dot	Amazon	Alexa
Google Home	Google	(Hey) Google
Magenta Mini	Deutsche Telekom	Magenta

Names and brands may be the property of others.

2.2.1 Conversational Assistants and Agents

For the purposes of this paper, the terms “assistant” and “agent” both refer (as noted above) to conversational AI user interfaces that are perceived to be a single conversational actor, operate on behalf of the user, and operate in accord with a conversational platform. Both assistants and agents have an addressable name, continuity of knowledge, and a bounded identity.

Both assistants and agents have the ability to fulfill user intents, mediate and delegate dialogs, and serve as a destination. **Agents** will also operate independently on behalf of the user. In the worldwide voice web, there will be specific-purpose *conversational assistants and agents,* and *general-purpose conversational assistants and agents.*

2.3 Platforms and Content: Existing Standards

Most conversational systems separate the *platform* from the *content* that is executed on that platform. In *Web Models* - a common structure - the content itself is typically either static or active web content. The platform is roughly analogous to a browser; the content is stored on web servers and accessed by standard web protocols (e.g. HTTP) (Fielding, 2022).

A set of related standards proposed by the W3C Voice Browser group has successfully created standardized formats for describing content for spoken dialog, speech recognition grammar,

semantics, text-to-speech markup and pronunciation (McGlashan et al., 2004). This content can be interpreted by a W3C compliant voice browser.

The W3C voice-browser group standards have served the telephony community well but have not yet been adopted by the leading general-purpose conversational platforms, nor by the many creators of text-centric chatbot platforms. One notable exception is *speech-synthesis markup*, which has been widely adopted (McGlashan et al., 2004).

This new wave of voice assistant platforms also uses a *web-content model* to support third-party dialog content. The formats for such content are mostly published but remain proprietary. Examples include Amazon Skill Definitions and Google Actions. Content for this new wave of platforms utilizes standard web protocols such as HTTP or OAuth (Fielding, 2022), but also strongly promotes or mandates proprietary infrastructure and authoring tools.

2.4 Conversational Architecture

Below is a schematic view of a generic conversational assistant system.

The Open Voice Network bases its explorations of voice assistance interoperability upon this understanding of system components and architecture.

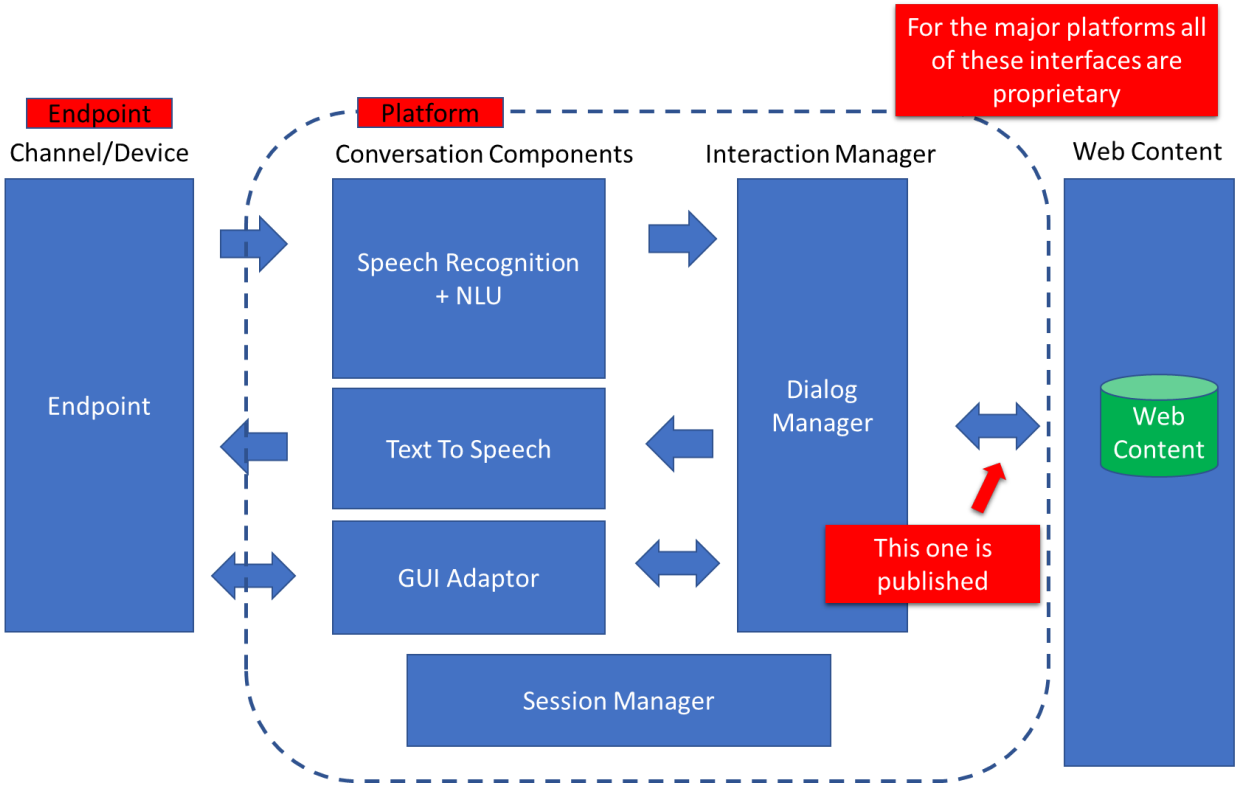


Figure 2.4 Typical Components of a Conversational Assistant

Figure 2.4 illustrates the schema of traditional conversational architecture, along with the various components that are collectively enlisted to enable two-way communication with the user.

The architecture typically operates as follows, as per the above diagram, from left to right:

The user begins (far left) with an **endpoint**, which can be located on any piece of digital hardware. Today's endpoint options include conversational assistant implementations on devices ranging from smart speakers, smartphones, desktop or laptop computers, and smart home systems, to automobile dashboards, television remote controls, and kitchen appliances – the list is endless. *Endpoints* provide a software or hardware interface between the user(s) and the **Conversational Platform**, which includes all the components within the encircled figure above. An example of a *simple endpoint* is an old-fashioned telephone (often referred to as the

plain old telephone system, or POTS). A *sophisticated endpoint* might be an immersive VR environment.[1]

Endpoints engage with the user via audio, text and visual media. Platforms interpret content that embodies voice assistants and enables them to interact with users via the endpoint.

The diagram above depicts three conversational components for exchanging information between the endpoint and the interaction manager. Let's explore these as listed from top to bottom in Figure 2.4.

The most traditional form of user-assistant communication involves **Speech Recognition Technology,** where the device interprets the user has uttered. Speech recognition technology (speech-to-text) has grown more accurate and sophisticated in recent years. Previous generations of this technology enlisted a very narrow spectrum of comprehension; at each dialog step, they required application-specific grammar to define the wording of 'allowable' spoken responses. Such grammars could be described using the W3C GRXML standard (McGlashan et al., 2004). Newer generations of speech-to-text engines use general purpose language models that can be shared across applications and contexts. This, of course, represents a considerable step forward. However, there are still no open standards for language model description at the present time, which limits interoperability.

Empowering this process, **Natural Language Understanding (NLU)** capability - a branch of Artificial Intelligence - enables the comprehension of the meaning of a user's human speech and clarification of underlying intentions.

The second category of conversational components - **a Text-to-Speech System (TTS) -** serves to convert language text into audible speech and can be implemented in software or hardware products. Normal language text may contain symbols, numbers and abbreviations that are converted to spoken words presented to the user via a speaker. The quality of a speech synthesizer is gauged by its similarity to the human voice and by its ability to be understood clearly.

[1] Standards exist for establishing audio connections between endpoints and platforms, for example in POTS telephone systems or VoIP connections. Smart speaker endpoints currently use interfaces that are specific to the platform. These may be published but they are not interoperable with other platforms.

The third category of conversational components, the **Graphic User Interface (GUI) Adapter,** comes into play in those unique instances where the user requests rich media content, such as buttons, images, and formatted text, to be visually displayed *on* the endpoint's screen. The endpoint will generally only be able to display said content if it receives specially formatted input from the platform. Likewise, the platform will only be able to "understand" user GUI input if it is first "translated" into a form that platform can understand.

As defined by the World Wide Web Consortium (W3C) Voice Interaction Community Group https://www.w3.org/community/voiceinteraction/, **Dialog Manager** is a component that receives semantic information derived from user input (via speech recognition + NLU), updates the dialog history, its internal state, then decides upon subsequent steps to continue a dialog and provides output.

The Dialog Manager gets its content and logic from the **Web Content** layer. The content might be hard coded or be dynamic via interactions with external APIs (for example it could output an address using a third-party address normalization API).

The Dialog Manager **also** passes information to the **Session Manager** and retrieves information from it. The Session Manager is a governance body within the voice architecture. It governs and administers conversational sessions with the user. This means controlling starting and stopping points, identifying and verifying the speaker, verifying the conversational assistant, managing connectivity among multiple conversational assistants, and, significantly, maintaining the *context* necessary for an active conversation between a user and a bot/assistant. A unique session ID is generated for each session, allowing the context of the conversation to be maintained over time.

2.5 The Diversity of Today's Conversational Assistance

The popularity of general-purpose consumer conversational assistance (Amazon Alexa, Google Home, Deutsche Telekom, Samsung Bixby, Baidu Xiaodu, and others) (Frost & Sullivan, 2022) has often obscured the growth of other parts of the conversational assistance ecosystem.

Firms such as Microsoft, Deutsche Telekom, Nuance (now a division of Microsoft), SoundHound, PeopleReign, Cerence, RASA, Mycroft.ai, Redfox.ai (and many others) (Frost & Sullivan, 2022) now offer enterprises the tools and services to develop conversational assistance solutions for devices, smart environments, personal transportation, customer service and employee support.

Enterprise investment in conversational AI worldwide is expected, by 2025, to reach a level double that of where it was four years prior; we're now seeing per year annual growth rates of roughly 22-25 percent (Frost & Sullivan, 2022).

Analysts foresee the bulk of future growth in these areas:

- Customer service and employee support solutions – many of which have grown through the years from call center automation. These are often classified as interactive voice response (IVR) systems.
- Personal and autonomous transportation
- Smart environments, from homes to manufacturing facilities
- Hands-free environments, from surgical suites to warehouses

SECTION THREE: AN APPROACH TO VOICE ASSISTANCE INTEROPERABILITY

3.0 Introduction

Section Three describes the Open Voice Network approach to the interoperability of voice assistants, and the current intent fulfillment patterns under investigation.

3.1 Modeling Interoperability on How People Communicate

The Open Voice Network acknowledges that a standard methodology for expressing and describing conversational interaction may be achievable -- but may not achieve the objectives of the “Jobs to Be Done’’ identified above, particularly for the Technical Innovators.

Attempts to describe (and standardize) how conversation can and should be modeled will quickly become outdated as the field of conversational interaction innovates. As an example, the latest generation of Voice System Infrastructure Providers and Technical Innovators put aside the widely adopted W3C VoiceXML and Speech Recognition Grammar standards in favor of creating new design paradigms based on less constrained speech-to-text and NLP technologies.

We assert this: human languages can be considered the ultimate interoperability standard. People have internal hidden motivations and knowledge. They use language to align their understanding of the world with tasks to be achieved. In some conversations people will have enormous levels of trust and shared understanding prior to the interaction. In others they will be complete strangers and may not trust each other.

We envisage that successful interoperability approaches will support **a model of autonomous agents collaborating to achieve a goal together using standard ways to share information** with each other**, operating at various levels of trust and information sharing**.

3.2 Interoperability: Mediation and Delegation

The modeling of voice assistant interoperability according to human communication leads us to identify two types of interaction: ***mediated*** and ***delegated.***

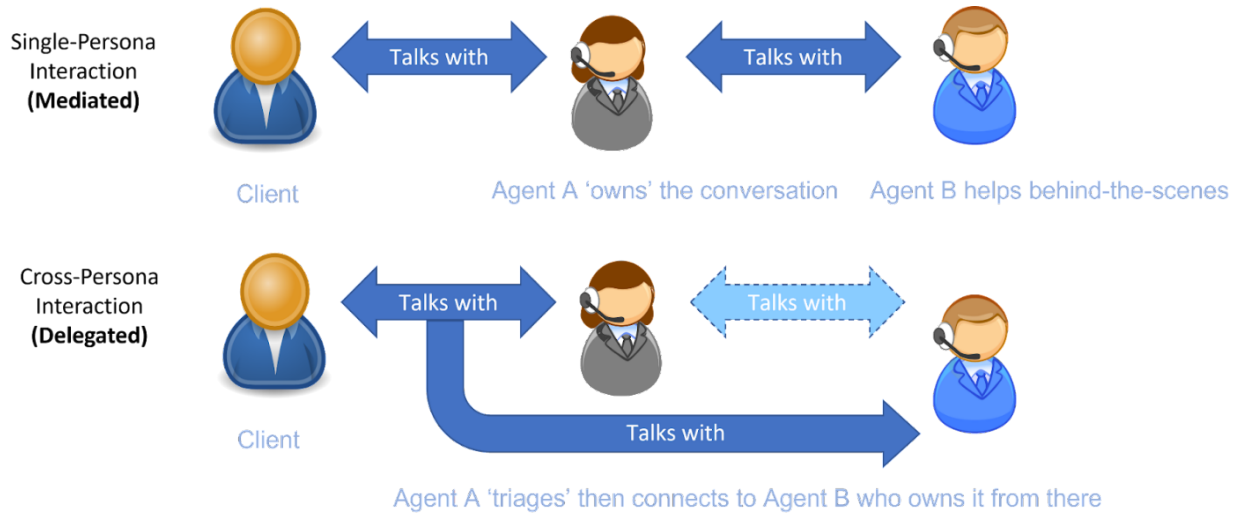


Figure 3.2.a Example of mediated and delegated communication.

Figure 3.2.a shows two examples of how people inter-operate to achieve a conversational task, in this case between a 'user' and a helping ' assistant'.

In the top example of ***mediated*** communication, the user interacts with a single assistant. assistant A owns the conversation but may talk with others (e.g., assistant B) behind the scenes to achieve the goal. The client may be aware of the existence of assistant B but does not interact directly with them. Assistant A is the ***host*** of the communication.

The second example is one of ***delegated*** communication. To fulfill the intent of the user, assistant A passes control of the conversation to assistant B. Assistant A may monitor the ongoing interaction or may drop-out of the conversation altogether. In this example, assistant A is the ***initiator*** of the communication.

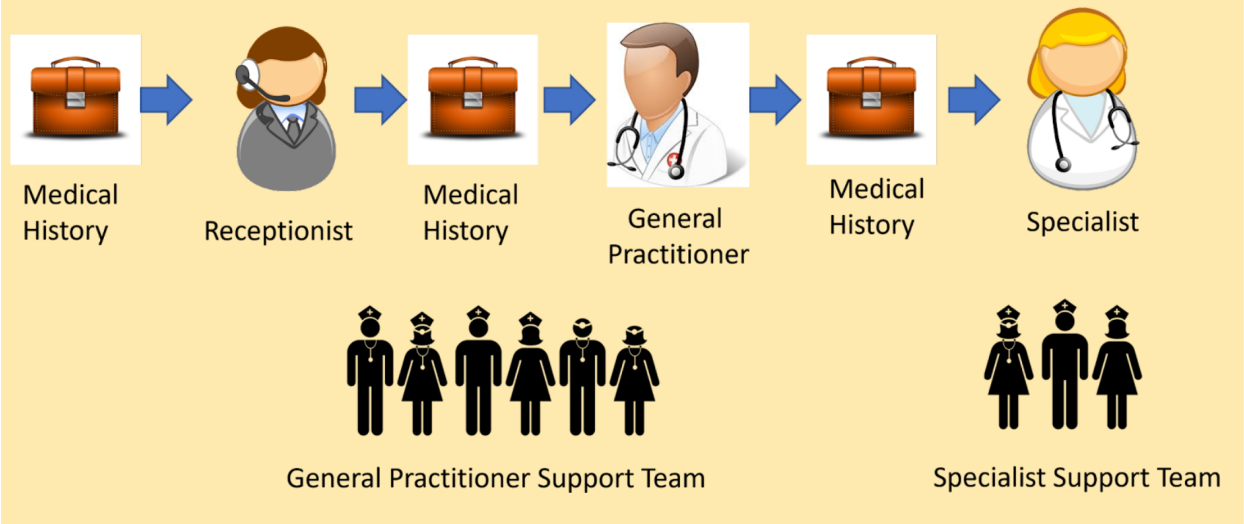
In this example of ***delegated*** communication, assistant A may need to talk behind the scenes with assistant B to establish whether assistant B is indeed the appropriate (or best) destination for delegation prior to their actual act of delegation.

To properly fulfill all user intents, assistant A will need to **both mediate and delegate conversations**. In some cases, the client may not need to communicate with assistant B resulting in a fully mediated conversation. For example, assistant A may place the user 'on hold'

if this were a telephone conversation, or the dialog may be paused if it is an electronic channel such as chat.

In addition, assistant A may also be asked to serve as a delegate **destination** – to receive a request from assistant C to fulfill a user intent.

Analogous to this, Figure 3.2.b illustrates another aspect of interoperability between people. In situations where considerable trust exists between the collaborating assistants then brief referrals (in language) can be made along with the passing of a case-history.



A brief referral plus a detailed case history is passed to the next specialist. Each specialist interprets their role based on a referral request and the case history

Figure 3.2.b. Example of a delegated communication in which context (here, a cast history) is also shared.

The Open Voice Network proposes that, to the extent it is possible, models of intra-assistant and inter- assistant collaboration follow similar patterns to those followed by human assistants.

Key features of such a model would include:

- Each assistant has an identifiable identity and knowledge boundary.

- Each assistant will both mediate a communication (serving as a host assistant) and delegate a communication (serving as an initiating assistant) depending upon the user intent. Assistants will also serve as a destination of a delegated communication.
- A desired scheme for interoperability will support collaboration as a mixture of mediated and delegated patterns.
- Delegation between assistants will be performed using brief requests.
- Levels of trust will vary between clients and assistants and this will affect the level of information sharing across boundaries.
- Where high levels of trust exist and there is shared understanding of how to represent knowledge, delegations and mediations will be supported by shared history and context.

3.3 From the Human Model to Standards: OVON Interoperability Objectives

The OVON architecture seeks to establish **standardized communication protocols** between dialog assistants running on diverse platforms.

It does not, however, seek to standardize platform components or the format of content used to configure such components. OVON envisages a world in which conversational systems will continue to evolve in complexity and function.

As such the OVON architecture celebrates diversity in the following areas:

- The diversity of underlying technology (i.e., speech recognition, language understanding and dialog modeling)
- The diversity of endpoints [e.g., smart speakers, smart phone apps, dumb phones (POTS), audio-enabled web pages etc.
- The diversity of conversational design paradigms
- The diversity of labels used to represent semantic content.

3.4 Architectural Patterns Under Investigation

In response to the Open Voice Network's Jobs To Be Done review, and informed by the vision above, the Open Voice Network Architecture Work *Architectural Patterns and Specifications: Delegation Overview*

OVON Group identifies several architectural patterns that, when combined together, could enable seamless intent fulfillment for users of different conversational assistants.

The architectural patterns that are being explored are:

- **Dialog Delegation and Give-Back (additional detail to follow in Section 3.4.1)**

  A standard, distributed approach to the delegation and take-back of tasks from one platform to another. This is modeled on the idea of intercommunicating human assistants.

- **Delegation Request and Conversation Context (Section 3.4.2)**

  A standard, extensible method of passing context between assistants. This is closely related to the delegation and take-back mechanism. This is modeled on the idea of maintaining case-history during sustained interactions.

- **Dialog Event Payload (Section 3.4.3)**

  A standard, extensible format for the passing of linguistic events between components. This is expected to consolidate work done over many years by the linguistic and dialog engineering community.

- **Dialog Component Interfaces (Section 3.4.4)**

  Standard component interfaces, built on top of the linguistic event format, starting with dialog interaction managers. This is expected to align as closely as possible with existing proprietary approaches.

- **Discovery and Location (Section 3.4.5)**

Standard patterns and mechanisms to allow assistants to publish and discover services from one another.

- **Sharing and Protection of Data (Section 3.4.6)**

  A strategy for sharing data elements among conversational assistants. It also postulates a constraint enforcement mechanism that enforces privacy constraints on shared data.

These six different design patterns could be used separately or in conjunction with each other. They could also be adopted as part of other interoperability initiatives such as the Stanford University Open Voice Assistant Lab (OVAL) model discussed later in this paper.

Additional design patterns are in review by the Open Voice Network Architecture Work Group.

3.4.1 Dialog Delegation and Give-Back

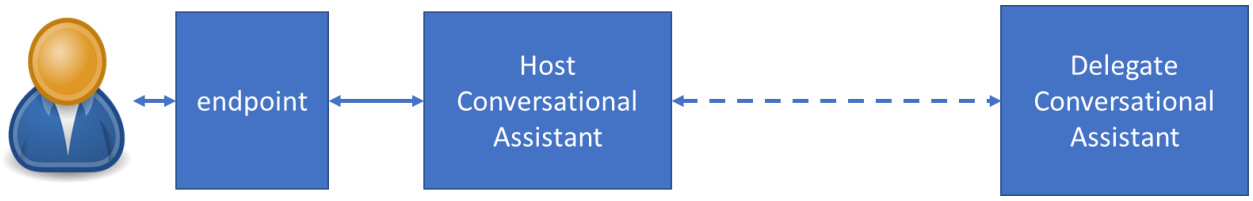


Figure 3.4.1 Delegation request

Figure 3.4.1 illustrates **a typical delegation request.** The user's request is transferred from the user to a conversational assistant (which we will call the *host*) which examines the request. If the host determines that it cannot process the request itself, then the host negotiates (the

dotted line in figure 3.4.1) with a second conversational assistant (the *delegate*) for possible processing.

The outcome of the negotiation will be an acceptance by both agents that the delegated request is accepted.Some establishing or authenticating information will typically accompany the request to provide context about the initiating assistant and requested task.

Through this mechanism, the delegate conversational assistant will assume control, and then fulfill the user's intent. In the most basic version of this scenario, the delegate assistant will then declare the task as finished. It may also ask the host assistant to reassume control.

3.4.2 Delegation Request and Conversation Context

In order to negotiate the delegation described above, the initiating assistant will frame its request so that all potential destination assistants can understand it. An establishing context for the dialog may also be necessary.

A delegation request from an initiating assistant requires these capabilities:

- Request: what the host assistant would like the potential delegate assistant to do: This might be a verbatim user request or a request generated by the host assistant to summarize what is needed at this point in the conversation.
- Context: conversational history, plus other parameters (TBD).
- Other data

The process of delegation may begin with an initiating assistant (assistant X) passing part of all of the user utterance as its 'request' to a Destination and Location Service or directly to a destination assistant (assistant Y). For example, one might say, "assistant X…" (wake word) ". . . ". . send me to assistant Y . ." '. . so that I might accomplish x, y and z".

The potential destination assistant Y will determine if and how it can respond to the request.

The Architecture Working Group is exploring the use of a standardized format to support the negotiation of delegation of requests in a standard form.

3.4.3 Dialog Event Payload

The OVON Architecture Working Group has identified three primary *levels* that can be used in communication between the different components of a dialog system. For example, one or all of the following levels could be used to communicate a request from one assistant to another.

- Level 0. An audio stream or file containing a natural language request.
- Level 1. A text string containing a natural language request
- Level 2. A structured semantic representation of a natural language request.

There is no architectural requirement for the delegation request to be built directly from the user's language. The host delegate could synthesize a delegation request in its own natural language. When level 2 communication is used, both conversational assistants must have a common understanding of the semantic representations. In addition, we do not anticipate a requirement for a delegation request to be built directly from the initiating assistant's language. The initiating delegate could synthesize a delegation request in its own language – natural or internal. This would also be of value in situations in which multiple natural languages are present, with either a multilingual human user or a multilingual assistant. The mechanisms for handling multilingual delegation are not yet specified.

The OVON Architecture Working Group recognizes that delegation requests are only one of a number of contexts in which linguistic information is passed between dialog system components in an interoperable environment. Section 3.4.4 will identify several dialog component interfaces which will use linguistic information.

The group proposes to define a standard format for a dialog event, which unlike other non-linguistic events which happen at a point in time, span a certain time-frame. We call this the 'dialog event payload'. This extensible format would support the representation of aspects of a linguistic dialog event in one or more parallel representation formats. Time spans could be whole utterances, phrasal units, or slices of an utterance evolving as a stream over time.

In addition to the three layers noted above (audio, text, and semantic representations of speech) the following may also be represented in the event payload:

- a user ID, standard for Identifying users

- a platform ID, standard for identifying the platform
- a conversational assistant, standard for identifying the conversational assistant
- a session ID, standard for identifying session
- inflection, standard symbol sets for annotating stress and intonation of speech
- pronunciation, standard symbol sets for phonemic or phonetic representation
- affect, standard symbol sets for coding of emotions or affect of the speaker
- a dialog act, standard symbol sets for representing dialog acts (e.g., asking a question).

Other richer semantic representation formats will likely emerge over time. The set of supported layers will grow over time as proprietary formats are proposed for adoption.

The Open Voice Network may propose a standard format for each of these layers. At present, we envision the identification of specific layers as either mandatory or optional for specific uses and contexts.

Regarding the Open Voice Network approach to standardization, the development of each of these layer schemas will draw heavily from formalized or de facto industry standards. These could include formal standards (e.g., TEI, UTF, W3C Pronunciation representation, etc.) or broadly adopted specifications from industry leaders.

The dialog event payload may possibly also be expressed using the Extended MultiModal Annotation Language EMMA (Johnson, 2009).

In an early proof of concept demonstration, the OVON Architecture Work Group successfully transmitted level 1 (text) messages between Mycroft, Magenta, and Genie conversational assistants. [MOAD summary.docx - Google Docs].

3.4.4.1 An Illustrated Example: Pat Goes Shopping

Let's return to the illustration of shopping with Pat, as presented earlier in section 1.5. Below, figure 3.3.4 breaks down the requests of Pat's dialog with two conversational assistants, Shop (a home shopping assistant) and Payperson (a payment assistant). Note the three layers of interfaces and their interaction, as depicted here.

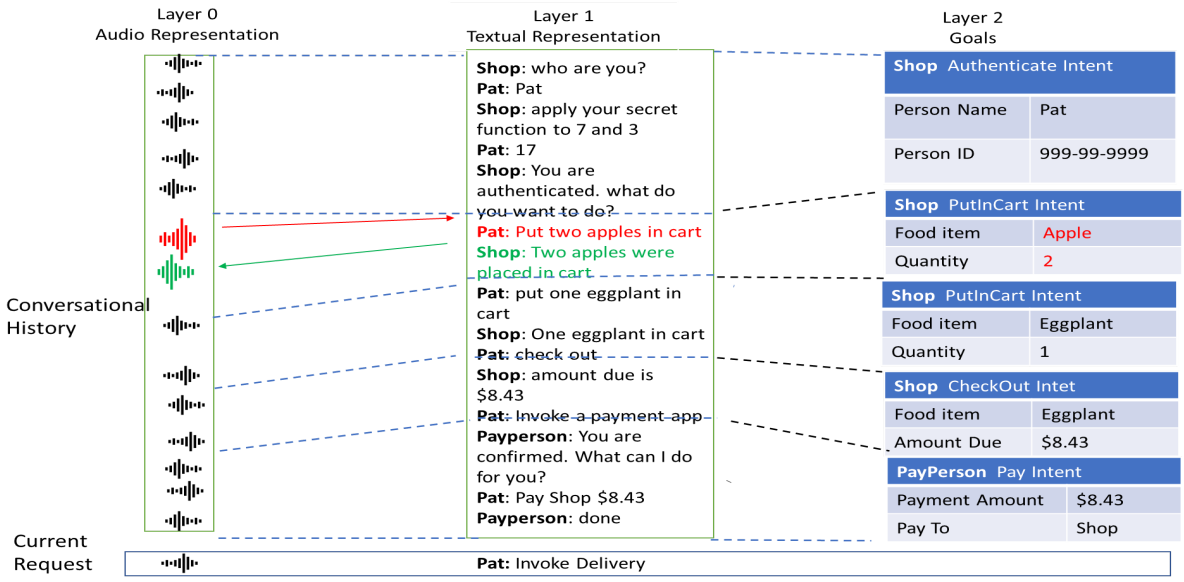


Figure 3.3.4: Current request and a subset of the conversational history of a Pat shopping dialog.

- ***Layer 0: Audio.*** *An audio stream or file containing a natural language request. Voice is represented in computers as a sequence of bits, often called a wav file. Above, we show electronic representations as wave icons.*

- ***Layer 1: Text.*** *A text string containing a natural language request. Above, the red icon, representing Pat's utterance, is converted into the text: "Put two apples in cart."*

- ***Layer 2: Semantics.*** *The structured semantic representation of a natural language request, here shortened to "goals." Semantics (meaning) of a user goal is often represented using intent and slot container structures, which represent goals/actions to be performed and the parameters defining the actions. Above, Pat's text expression is converted into the Intent, PutInCart, with the parameters "Apples" as the value of Food Item and "1" as the value of Quantity.*

To complete the example, the dialog manager generates the text message, "Two apples were placed in cart," which is converted by the TTS into a wav file (depicted by the green waveform graphic above) *and then presented to the user.*

3.4.4 Dialog Component Interfaces

Figure 2.4 shows the interfaces between typical components in a dialog system. The Open Voice Network does not currently intend to specify interoperability standards between all of these components, preferring instead to focus on those that are needed for interoperability between agents hosted on different platforms and infrastructure.

Having said this, the Architecture Working Group does envisage that the standardization of a dialog event payload as described in section 3.4.3 will be of general use to such an effort.

The dialog component interfaces that are considered a priority for interoperability are as follows:

- Delegation request and negotiation between agents.
- Delivery of interaction events from the user to the agent.
- Delivery of agent output for rendering to the user.
- Maintenance and communication of dialog history and context.

The Working Group envisages that the dialog event payload format will play a strong role in all of these interfaces.

3.4.5 Discovery and Location

An open, worldwide voice web will allow users to fulfill an intent through a mediated or delegated connection to *any* content source or conversational assistant, regardless of platform parentage. *We believe that conversational assistance can and should work like the web.*

Users will need to both *find* (e.g., the connection to a specific, named destination) and *discover* (e.g., the exploration of information and destination options within a topic or category) content in the worldwide voice web.

To meet this need, the Open Voice Network has initiated research into the development of what may be termed *Discovery and Location Services.* The services may include:

- a DNS-like service for the identification of available conversational assistants (by name and address)

- a standardized approach to metadata representation and organization

- the use of search engines, browsers, and aggregators that examine meta information about conversational assistants that provide the conversational assistant's address
- the use of natural language requests sent from one assistant to another which the receiving assistant can use to decide if it can fulfill the request.

A request to a Discovery and Location service from an initiating assistant will likely include linguistic information with the user's request and contextual information.

The response from Discovery and Location services(s) and an identified destination assistant will include a digital address, configuration information, and privacy/security constraints.

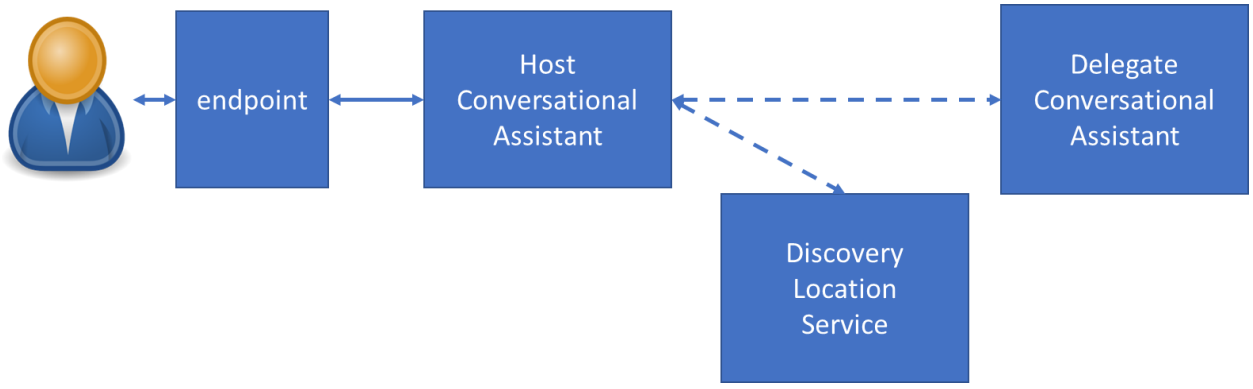


Figure 3.3.5: Schematic of the role of an envisioned Discovery and Location Service.

The schematic above illustrates **the role of an envisioned Discovery and Location Service** in the process of delegation. In a future worldwide voice web with millions, if not billions of destinations, a Discovery and Location Service (or Services) would enable an initiating assistant to find and connect with a destination assistant. In practice, it would determine which conversational assistants should be invited to the delegation and identify the digital addresses of the potential destinations.

3.4.6 Sharing and Protection of Data

An open, standards-based worldwide voice web of mediated and delegated dialogs will require a new and systemic approach to data protection. The Open Voice Network seeks to enable secure end-to-end connection between users and conversational assistants. This may include the encryption of messages, authentication of users as they initiate and progress through delegated dialogs, and authentication of conversational assistants in both mediated and delegated dialogs.

Work in these areas is underway.

Earlier this year, the Privacy and Security Work Group of the Open Voice Network Technical Committee published for public review the guiding white paper Privacy Principles and Capabilities Unique to Voice.

From a perspective of a human user of voice assistance – and anticipating an OVON assertion of both mediated and delegated interoperability – the Privacy Principles white paper reviews the current landscape of regional and national privacy regulation and legislation, identifies primary human user risks and potential harms, and details four principles that the OVON believes should guide the development and implementation of voice assistance:

- **Transparency:** proactive communication – through an easily accessible and readily available interface -- to the individual user of data collection practices, general data usage, and data sharing policies.

- **Consent:** the expectation that the individual user must be allowed to give explicit, unambiguous agreement to the collection and processing of personal data.

- **Limited Collection and Use:** a restriction of the collection and analysis of raw and processed voice data – beyond that necessary for immediate dialog functionality – to stated and creator consent-given purpose.

- **Control:** the ability of the individual user to easily access, rectify, suppress, limit, oppose, and transport the data created by the individual user.

Under development now by the Privacy and Security Work Group are two additional papers:

- ***Data Security Specific to Voice,*** a review of current data security issues within voice assistance, and of data security issues inherent within multi- assistant voice interoperability. This paper is scheduled for publication in August 2022.

- ***Voice Assistance and the Privacy of Organizational Data,*** an exploration of why, when, and how organizations of all types (and especially enterprises) may protect proprietary data from undesired sharing and platform collection in both mediated and delegated interoperability. A first draft of this paper is scheduled for publication in the fourth calendar quarter of 2022.

In parallel, the Ethical Use Task Force of the Open Voice Network published for public review the white paper Ethical Guidelines for Voice Experiences.

From the perspective of a human user of voice assistance, the Ethical Guidelines white paper identified voice-specific issues of ethical and moral concern, discussed rights that must be respected and values to be promoted, and addressed preventative measures that could be taken to protect users from voice-specific harm.

The paper also outlines a voice-specific ethical framework of five principles:

- **Compliance**: the acknowledgement of, and adherence to, ethical principles, standards, guidelines, and existing laws and regulation.
- **Transparency:** open and clear communication – in an easily accessible, understandable, and explainable user interface – regarding the collection, usage, and sharing of user and user-created data.
- **Privacy Protection:** not only adherence to the regulation and legislation that governs personal data of all types (General Data Protection Regulation (GDPR) – Official Legal Text, 2019) - especially voice data (Frost & Sullivan, 2022) - but the prioritization of data security and the vetting of the third parties that may, even with transparent consent, handle and process voice data.
- **Inclusivity**: working at all times to allow all to be heard – across languages, dialects, genders, ages, ethnicities, types and levels of disability.
- **Accountability:** maintenance of, and adherence to, highest ethical standards throughout the voice development, implementation, and operational value chain.

A next step for the Open Voice Interoperability Initiative (see Section Six, below) will be to translate the principles of these four papers into tangible, implementable technical guidelines and specifications.

SECTION FOUR: LESSONS FROM OTHER INTEROPERABILITY INITIATIVES

4.0 Introduction

As noted in Section One, the coming world of voice will be one of diversity, one marked by a multitude of voice assistants and agents. At present, however, the realm of general-purpose consumer-facing conversational assistants is one of singular, proprietary platforms that do not interact with one another.

As noted, the Open Voice Network envisions a future in which proprietary walls to content and connection are lowered, and users may seamlessly move from one assistant to another according to intent – a future in which voice operates broadly like the web, and not like apps on a mobile platform.

This section provides an overview of two important interoperability initiatives outside the Open Voice Network. We applaud –and are learning from – both. We also anticipate some level of adoption and adaptation of the concepts described here, as our work progresses.

4.1 Amazon Voice Interoperability Initiative (VII)™

The Amazon Voice Interoperability Initiative (VII) enables the deployment of multiple conversational assistants on a singular endpoint device. Assessed with the Open Voice Network framework, it offers *limited delegation – e.g., it does not allow the delegation of a conversation, but allows the user to leave one assistant and engage with another.*

In a VII implementation, each assistant is activated via its own 'wake word,' enabling users to talk to the assistant of their choice in a secure manner by simply saying its name. For example, a user can interact with a specialized conversational assistant (such as Fridge) for specific refrigerator interactions, and leave the specialized conversational assistant and switch to a general-purpose conversational assistant (Amazon Alexa) for general purpose interactions. In this example, the Fridge conversational assistant would be designed with the vocabulary and skills specific to refrigerators and would bypass the vocabulary and skills that Alexa understands. Likewise, Alexa would not need to deal with the vocabulary and skills for the refrigerator.

The assistants interoperate with each other via the Multiagent experience (MAX) toolkit. The toolkit provides the MAX Library and a Sample Application that demonstrates the interoperability of Alexa and a second independent voice assistant. The MAX Library facilitates interoperability between voice assistants according to the guidance provided by the Voice Interoperability Initiative (VII) Multi-Agent Design Guide.

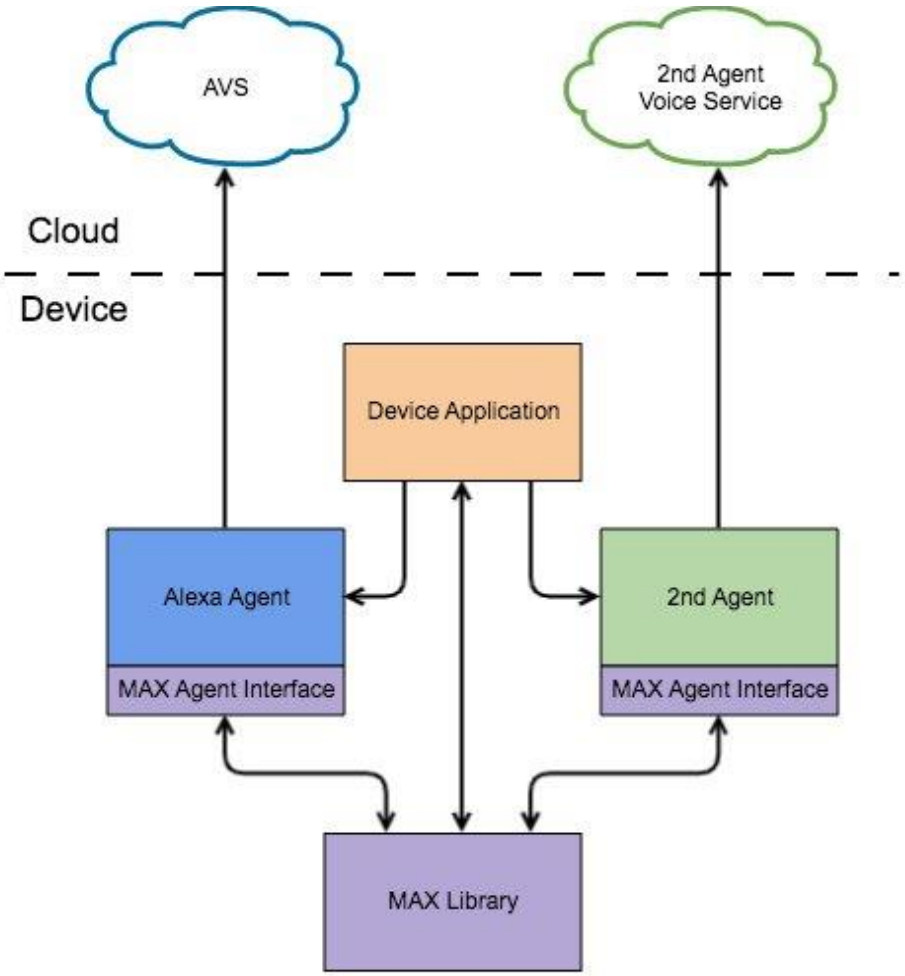


Figure 5.1.1: Amazon Voice Interoperability Initiative

Within the VII guidelines, an assistant can transfer a user to a second assistant when it cannot directly fulfill a user request and is aware that the second assistant on the device can likely fulfill that request. No data or context is passed between assistants during a transfer, and the customer repeats their request directly to the second assistant without needing to say the wake word.

In addition, assistants interact with the device via Universal Device Commands. *UDC*s are commands and controls that a customer may use with any compatible assistant to control certain device functions, even if the assistant was not used to initiate the function. Examples

include changing the volume of the device's output speaker or stopping a sounding timer or music initiated from assistant B when assistant A is in control.

Comparison to the Open Voice Network proposal

- The focus of Amazon's VII is on multi- assistant devices, i.e. several assistants accessed from one device.
- Delegation is limited to the assistants registered to the device. The Open Voice Network envisions a limitless number of potential voice-enabled destinations.
- Interoperability and assistant "discovery" is handled by the MAX toolkit and is limited to assistants registered and running on the device. In contrast, OVON allows for interactions between assistants potentially running exclusively on the cloud.
- In VII, assistants are discouraged to talk directly or pass information and context to each other. The user will need to repeat the queries to the second assistant.
- VII demands the presence of two running middlewares to mediate assistant interactions (the process running the MAX library and the device application and UDCs).

4.2 The Stanford Open Voice Assistant Laboratory (OVAL) Model

The Stanford University Open Voice Assistant Laboratory, under the direction of Professor of Computer Science Dr. Monica Lam, has proposed an open-source model for voice interoperability that enables mediation and partial delegation to multiple independent devices and services. (Lam et al., 2021). (Full disclosure: Dr. Lam is a valued advisor to the Open Voice Network, and OVON is a financial supporter of OVAL.)

The Stanford OVAL model also follows the "Standardized Open Single Platform" model. The Stanford OVAL model gives developers the ability to collaboratively create "articles" through a device and services knowledge base known as "Thingpedia." This empowers users to move from an initiating assistant to voice-enabled devices (i.e., a smart light bulb or smart factory sensor) and various services (a smart home system, a restaurant or retail website, or media properties

such as Twitter or radio content). OVAL has demonstrated this model with several partners, using the OVAL “Genie” open-source conversational assistant as both mediator and initiator.

A strength of the Stanford OVAL model is that it allows the chaining and management of different articles combined in a single utterance (i.e., “When weather goes over 100F tweet out ‘oooh it’s hot’”). This also allows users to access multiple devices/services in a single request, although interoperability is restricted to the destinations, services, and use cases identified in Thingpedia.

Another significant strength of the OVAL model is that it enables users to enroll and save their devices/services. This allows users to interact with these devices/services securely and privately (without the need to authenticate separately and repeatedly).

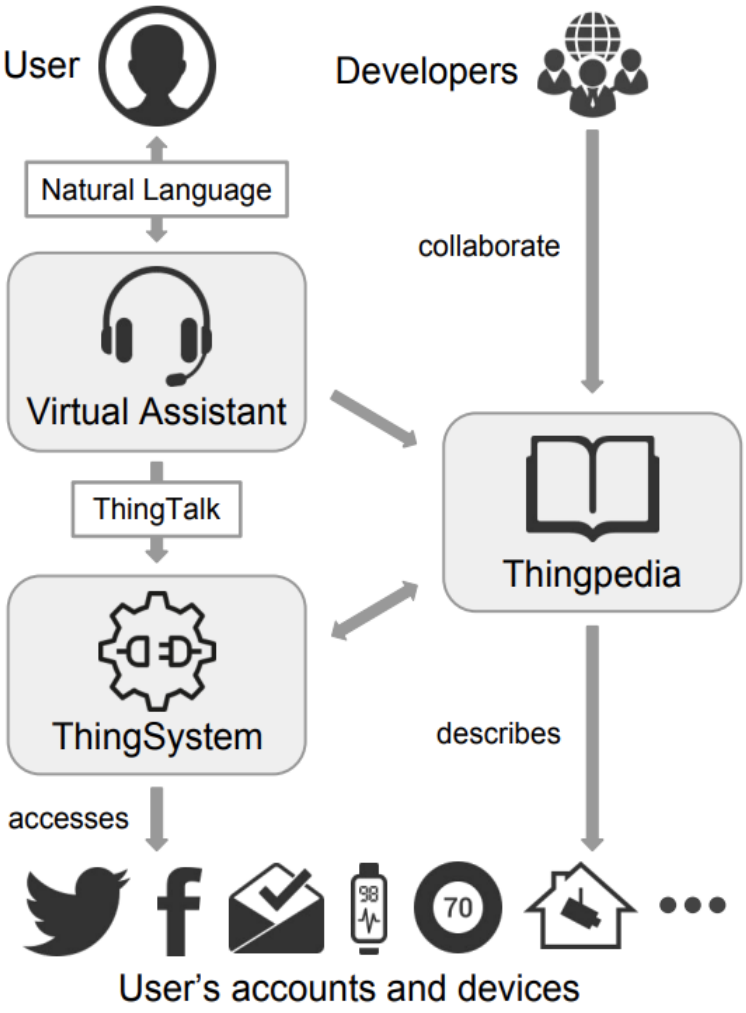


Figure 5.2.1 Stanford OVAL architecture

The three core elements of the OVAL model:

ThingPedia

- ThingPedia is the knowledge base of what can be done by each device/service.

ThingSystem

- ThingSystem stores all device/service credentials for individual users allowing the platform to maintain privacy and security for the user. It runs the ThingTalk code that is provided by ThingPedia.

ThingTalk

- ThingTalk is the programming language that lies at the heart of OVAL platform. It gives the platform the power to connect IoT devices, web services, and database queries that are specified in ThingPedia.

Observations:

1. We perceive the OVAL model to be primarily one of single- assistant mediation.

2. If the OVAL model (with knowledge base Thingpedia and programming language ThingTalk) were widely adopted, it would deliver the capabilities of the Discovery and Location services identified above as a requisite for interoperability.

3. As the ThingPedia service expands, we fear that its ability to deliver discoverability may become strained. Currently Thingpedia has several hundred distinct “skills.” We must learn plans (UX, governance, etc.) for the scaling of the knowledge base to incorporate millions (or billions) of future skills.

SECTION FIVE: FURTHER STUDY AND NEXT STEPS

This paper is but the first step on an important but challenging journey toward mediated and delegated voice assistance interoperability.

Visible next steps for the Open Voice Network Architecture Work Group include the following:

- The review, discussion, revision, and confirmation of the concepts shared in the Work-in-Progress content in the Appendix to this white paper.

- The identification and assertion of messaging protocols that will allow independent voice assistants to mediate, delegate and serve as destinations for dialogs. With the encouragement of the Open Voice Network Steering Committee, the Architecture Work Group will first pursue adopting or adapting both existing standards and technologies that are now broadly used within the voice industry.

- Continued research as to mediated and delegated interoperability with existing and emerging voice and conversational AI services, including:
    a) interactive Voice Response (IVR) systems inside corporate data security firewalls
    b) voice-enabled conversational bots
    c) voice-enabled web applications
    d) conversational AI implementations within enterprise software and processes
    e) enterprise-focused Conversational AI platforms.

- The development, with the OVON Voice Registry Work Group of the Technical Committee, of a standards-based approach to discoverability, findability, and location services. We anticipate publication of a separate document on this issue in the months ahead.

- The expansion of OVON messaging protocols to enable the sharing of multi-modal content.

SECTION SIX: OPERATIVE VOCABULARY

Architectural pattern: a model of an important feature that occurs in all implementations of a system.

Component: an identifiable part of a voice assistant or agent. A component provides a particular function or group of related functions.

Context: Information extracted from *n* prior utterances of the current conversation. This could include some or all of the following: information that has been input to, output from, or inferred in Conversational Processors, and the information state of the Dialog Manager. Also known as 'Conversational Context.'

Conversation: a joint activity in which two or more assistants (human or automated) use linguistic forms and non-verbal signals (i.e., gestures) to communicate to achieve an outcome that meets a shared goal.

Conversation Event: a conversation event signals shifts in the conversation that may be acted upon. Such an event may occur at the beginning or ending of a Conversational Session, completion of a Conversation Processor, decoding of Conversation Information, changes to the state of a Conversation Endpoint, or changes to the status of a Conversation Stream. Any component with access to the system is allowed to generate a Conversation Event.

Conversation Facilitator: a component that coordinates communication between two or more Dialogue Systems and/or Processors during the course of one or more Sessions. This allows dialogue Systems and associated Processors to collaborate regardless of technology being used. Examples of Conversation Information include semantic, lexical, syntactic, and prosodic features.

Conversation Information Layer: an abstraction of a type of information in a Dialog System. A layer may be a specific type of acoustic, linguistic, non-linguistic, or paralinguistic features. Examples of layers would be Cepstral features, Phonemes, Intonation Boundaries, Words,

Phrases, Turn Boundaries, Syllabic Stress, Discourse Move Type and specific Semantic representation schemes.

Conversation Processors: conversation information is encoded and/or decoded by one or more Conversational Processors. Conversational Processors may also take as input the output from another Conversation Processor. A Conversation Processor may generate Conversation Events and Conversation Streams. Example conversational processors include Automatic Speech Recognizers (ASR), Natural Language Processors (NLP), Dialog Managers, Text to Speech Synthesizers (TTS), etc.

Conversation Session: a particular conversation that consists of two or more Conversation Streams (see below) generated by two or assistants through one or more Conversational Endpoints. Sessions may be persistent, but they will often have a start-point and an end-point in time determined by one of the assistants or some other external event.

Conversation Stream: each Conversational Endpoint generates one or more Conversation Streams based upon the capabilities of the Endpoint and the preferences of assistant. A Conversation Stream is associated with a particular assistant and may include any media type including text, audio, video, and application UI events.

Conversational assistant: a digital participant in a conversation. This may be an application with a consistent persona, such as Amazon Alexa, Google Assistant, the Target Google Assistant Action, a Facebook Messenger chatbot, or an IVR system at a bank or a human. conversational assistant is a common term utilized in Dialogue System research and university level instruction; the term often is used to describe a human participant in a conversation. For clarity of reference, however, the OVON will use the term "user" to identify a human participant. (See "user" below.)

Conversational AI: the set of technologies to enable automated communication between computers and humans. This communication can be speech and text. Conversational AI recognizes speech and text, understands intent, decipher various languages, and responds where it mimics human conversation. In some cases, it is also known as Natural Language Processing.

Conversational Context -See 'Context,' above.

Conversational Delegation: the passing of dialog layers and control between one Conversational Assistant and another to fulfill a user intent. The first assistant in the delegation sequence is the *initiating assistant*; the second is the *destination assistant*.

Conversational Endpoint: assistants conduct conversations using conversational endpoints; these may be a phone, mobile device, voice speaker, personal computer, kiosk, or any other device that enables an assistant to participate in a conversation. Endpoints may be referred to elsewhere as a "device" or a "channel."

Conversational Information Packets: information that relates to a specific period of time. Packets form the input and output of Conversation.

Conversational Mediation: the hosting of a dialog by a Conversational Assistant. In conversational mediation, the host assistant may fulfill a user intent by itself; it may access third-party data sources through API calls; or, it may introduce to the user a third-party application that is resident on the platform of the host assistant. A mediating assistant does not cede control, nor access to the data within the conversation.

Conversational Platform: A group of technologies that are used as a base for one or more conversational assistants; also (see "Platform" below) a business model that harnesses and creates a large, scalable network of users and resources that can be accessed on demand.

Data: (per the Cambridge Dictionary): information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer. Context (see above) is a subset type of the data accessed and used by the voice assistant system.

Dialog Manager (DM): handles the dynamic response of the conversation. It provides a more personalized response based on the action provided by the NLP to send back to the user.

Disambiguate: when the conversational platform hypothesizes two or more possible resolutions to a user utterance, it may ask the user for additional clarification or choose between the various interpretations to decide the user's correct intention.

Entity: a custom level data type and considered a concrete value to associate a word(s) in a query. It is a part of the structural machine translation. Also known as annotations.

Explicit Invocation: an invocation type where the user invokes the channel, and explicitly states a direct command to accomplish a specific task. The direct authority is to communicate directly to a registered voice application.

Implicit Invocation: an invocation type where the user invokes the channel by describing the target destination rather than naming the target.

Invocation: a part of the construct of the user's utterance during a conversation with a channel. An invocation describes a specific function that the guest wants, and solicits a particular response.

Intent: the identified action that the machine interprets based on the user's query. It is a part of the structured machine translation. Also known as a *classifier*.

Jobs to be Done: An approach to learning what will cause a customer to hire or bring your product or service into their life.

Natural Language Processing (NLP): a service and a branch of Artificial Intelligence that helps computers communicate with humans in their language and scales other language-related tasks. NLP helps structure highly complicated, unstructured human utterances and vice-versa. Natural Language Understanding is a subset of NLP that is responsible for understanding the meaning of the user's utterance and classifying it into proper intents.

Organization: a group of individuals brought together for a specific purpose, including the creation, transaction, and delivery of products or services. Examples would include a for-profit business, a not-for-profit group, or a government agency.

Platform: The collection of components (the environment) needed to execute a voice application. Examples of platforms include the Amazon and Google products that execute voice applications.

Query: user’s word requesting for specific function and expecting a particular response.

Speech-To-Text (STT): conversion of a representation of an utterance from audio to text.Also known as Automatic Speech Recognition (ASR).

Text-To-Speech (TTS): conversion of a representation of an utterance from text to audio. Also known as Speech Synthesis.

Technical Resource: it can be a publisher/developer. It can be a representative of an entity or independent party. Their role is to create an actual listing of the voice application.

Utterance: spoken or typed phrases.

User: a person who interacts with channels.

Web content model: the content is interwoven with the mechanism to access the content.

Web model: the content is separated from the user mechanism to access the content.

SECTION SEVEN: ABOUT THE OPEN VOICE NETWORK

The Open Voice Network (OVON) is a non-profit industry association dedicated to the development of standards for voice assistance transparency, consent, limited collection, and control of voice data that will make using voice technology worthy of user trust. In any reality, virtual or otherwise, we believe personal privacy should be respected as the default. The Open Voice Network operates as an open-source community within The Linux Foundation. It is independently funded and governed with participation from more than 120 voice practitioners and enterprise leaders from 12 countries.

The Open Voice Network community's work is open source. We seek inclusive input and like to share our insights. At present, our work is focused in four areas:

- **Interoperability**, defined as the ability for conversational assistants to share dialogs (and accompanying context, control, and privacy)
- **Destination registration and management**, the ability of users to confidently find a destination of choice through specific requests, and for the providers of goods and services to register a verbal "brand" — similar to the Domain Name System (DNS) of the internet
- **Privacy**, with voice-specific guidance for both the protection of individual user data and that of commercial users
- **Security**, with a focus on voice-specific threats and harms.

Please see our 2022 papers and support the Open Voice Network by visiting [openvoicenetwork.org.](openvoicenetwork.org)

About The Linux Foundation

Founded in 2000, The Linux Foundation is supported by more than 1,000 members and is the world's leading home for collaboration on open-source software, open standards, open data, and open hardware. Linux Foundation's projects are critical to the world's infrastructure including Linux, Kubernetes, Node.js, and more. The Linux Foundation's methodology focuses on leveraging best practices and addressing the needs of contributors, users, and solution providers to create sustainable models for open collaboration. For more information, please visit us at linuxfoundation.org.

The Linux Foundation has registered trademarks and uses trademarks. For a list of trademarks of The Linux Foundation, please see its trademark usage page: www.linuxfoundation.org/trademark-usage. Linux is a registered trademark of Linus Torvalds.

Acknowledgements

This paper is authored by the Open Voice Network, with special thanks to the Architecture Work Group of the Technical Committee. Contributors: David Attwater, serving as Senior Research Scientist; Oita Coleman; Dr., Deborah Dahl, Senior Editor; Bruce Epstein, Co-Moderator of the Architecture Work Group; Pradeep Gopal; Vineet Hingorani; Olga Howard; Carl Jahn; Kiran Kadekoppa; Dr. Jim Larson, Co-Moderator of the Architecture Work Group; Tobias Martens; Dr. Yaser Martinez-Palenzuela; Nick Myers; Shyamala Prayaga, Co-Moderator of the Architecture Work Group; Elizabeth Robins; Dr. Dirk Schnelle-Walka; Nathan Southern; Jon Stine; Vadim Tarasevic; John Trammell; Boris Volfson.

We are grateful for the ongoing support of the Steering Committee of the Open Voice Network: Joel Crabb, Chair; Mirko Saul, Vice-Chair; Ali Dalloul, Bernhard Hochstätter, Doug Rogers, and Christian Wuttke. We wish also to thank and recognize individuals whose guidance, encouragement, and founding vision made this effort possible: Mike McNamara, Dan Cundiff, Kristi Dank, Maria Brinas-Dobrowski and Jay Kline; Ryan Steelberg and Sean King; Dr. Monica Lam and Jimmy Garcia-Meza; Reghu Ram Thanumalayan; Xuedong Huang; Bradley Metrock; Birgit Popp, Ulrike Stiefelhagen, and Anna Leschanowsky; Lawrence Lin.

SECTION EIGHT: REFERENCE LIST

Christensen, C., Hall, T., Dillon, K., & Duncan, D. (2016). Know Your Customers' Jobs to be Done. *Harvard Business Review*, *94*(9), 54–62. https://hbr.org/2016/09/know-your-customers-jobs-to-be-done

European Data Protection Board. (2021, July 7). *Guidelines 02/2021 on Virtual Voice Assistants*. Retrieved August 11, 2022, from https://edpb.europa.eu/edpb_en

Fielding, R. T. (2022, June 1). *RFC 9110: HTTP Semantics*. RFC Editor. Retrieved August 11, 2022, from https://www.rfc-editor.org/rfc/rfc9110.html

Frost & Sullivan. Opportunities in the Conversational AI Market (2022, August); frost.com. Retrieved August 8, 2022 in advance of web publication.

General Data Protection Regulation (GDPR) – Official Legal Text. (2019, September 2). General Data Protection Regulation (GDPR). Retrieved November 8, 2022, from https://gdpr-info.eu

Haas, M. (2020, July 16). *Understanding Conversational AI Tech*. Interactions.Com. Retrieved August 11, 2022, from https://www.interactions.com/blog/technology/conversational-ai-technology/

IEEE Standards Information Network/IEEE Press. (2000). *The Authoritative Dictionary of IEEE Standards Terms (IEEE 100), Seventh Edition* (7th ed.). Institute of Electrical and Electronics Engineers (IEEE).

Johnston, M., Baggia, P., Burnett, D., Carter, J., Dahl, D., McCobb, G., & Raggett, D. (2009, February 10). *EMMA: Extensible MultiModal Annotation markup languag*e. World Wide Web Consortium. Retrieved August 11, 2022, from https://www.w3.org/TR/emma/

Lam, M., Landay, J., & Manning, C. (2021). *Launching a World Wide Voice Web*. Stanford University.

McGlashan, S., Burnett, D., Carter, J., Danielsen, P., Ferrans, J., Hunt, A., Lucas, B., Porter, B., Rehor, K., & Tryphonas, S. (2004, March 16). *Voice Extensible Markup Language (VoiceXML) Version 2.0*. World Wide Web Consortium. Retrieved August 11, 2022, from http://www.w3.org/TR/2004/REC-voicexml20-20040316/

SECTION NINE: CHANGE LOG

Content changes:

1. Updated publication dates and explained that this is a revision of 1.0
2. Changed feedback process to mailing list, from GitHub
3. Section 1.0: Added comment about "agent" vs. "assistant" terminology
4. Section 1.6: Added principles that 1. whatever happens, human is in charge and 2. human needs to be able to cancel at any time.
5. Section 1.7: Replaced references to "avoid" with positive language.
6. Section 1.9: Clarifications to the terms conversational context", "conversational continuity", and "history" on page 22.
7. Section 2.4: "As defined by the World Wide Web Consortium (W3C) Voice Interaction Community Group (McGlashan et al., 2004), Dialog Manager is a component that receives semantic information derived from user input (via speech recognition + NLU), updates the dialog history, its internal state, then decides upon subsequent steps to continue a dialog and provides output." Added link to the Voice Interaction Community Group and removed the reference to McGlashan et al., since that is a different group.
8. Section 3: Overall clarifications

Editorial changes:

1. Corrected typos and cut and paste errors throughout
2. Edited formatting to keep tables from breaking across pages

3. Enlarged Figure 5.1.1 to increase the font sizes
4. Removed references to Appendices, since they are not included
5. Removed duplicate text in Section 3.3 that duplicated text in Section 1.10