OPEN VOICE NETWORK
Voice worthy of user trust

2022.11.08
Draft Version 1.0

Interoperability Architectural Patterns - Initial Thoughts

The Open Voice Network
Architecture Work Group of the Technical Committee

November 8, 2022

OPEN VOICE NETWORK
Voice worthy of user trust

- TABLE OF CONTENTS

Chapter 0 - Preliminaries

0.1 Preface

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 document describes the basic architectural principles and requirements for achieving this interoperability. It builds upon an initial whitepaper "Interoperability of Conversational Assistants" found at https://drive.google.com/file/d/15GEequrmi7gscHDoWy9X9pNZoLAK7OJ0/view?usp=sharing (hereafter, "the whitepaper") and is part of a set of documents produced by OVON covering all aspects of interoperable voice agents, including but not limited to: security, privacy, ethical use, and discovery & location. These can be found at https://openvoicenetwork.org/white_papers .

The authors of this document are members of the OVON Interoperability Architecture Working Group; we can be reached at whitepapers@lists.openvoicenetwork.org.

0.2 Introduction

The purpose of this document is to capture and describe the requirements for the subsequent protocols and interface standards that OVON will produce, to foster an interoperable voice ecosystem. These requirements are intended to be as independent from existing technologies and solutions as is possible, with a view toward future innovations which are likely to overcome limitations of existing technologies.

To accomplish this goal, the Open Voice Network Architecture Working Group has been meeting weekly as a forum for gathering diverse viewpoints, representing technology innovators and content providers. We have used a combination of directed brainstorming and strawman technical review as a means to elicit ideas and evaluate their suitability within the framework of our mission to enable "*Voice worthy of user trust*."[1]

This initial draft is being released in an admittedly incomplete state, as a way to enlarge the range of feedback and ideas relevant to our task of capturing and identifying the requirements for the protocols and interfaces that we are attempting to define. Your input is welcome at whitepapers@lists.openvoicenetwork.org or at any of our meetings, a schedule of which can be found at https://openvoicenetwork.org/initiatives.

0.3 How This Document is Structured

- The next section below lists a few definitions of terms that we employ in this document. A more complete list can be found in the glossary of the whitepaper.
- The first chapter describes the assumptions and preconditions for achieving interoperability among voice agents as well as the scope of OVON efforts with respect to those assumptions.
- Chapter 2 describes the basic patterns of interaction, which we are calling "native, mediation, channeling, and delegation."
- Chapter 3 illustrates the mechanics of interaction between agents.
- Chapter 4 outlines the additional attributes and topics that we are exploring.
- Topics listed in chapter 4 will be developed independently; we anticipate releasing an updated chapter about every two months; check back frequently for updates.
- Content will also be updated based on reviewer & implementation feedback.

0.4 Definitions and Conventions

An *Agent* or a *Conversational Assistant* is a generic term for a digital participant in a conversation in a voice-driven system.

A human participant will always be referred to as *human* or *user*.

[1] https://openvoicenetwork.org/

A *client*, *conversational endpoint* or simply *endpoint* is a generic term for the physical device that enables a human user to interact with one or more conversational assistants. The client or endpoint may itself offer additional capabilities including those often associated with an assistant, but this is explicitly neither presumed nor required in our examples which follow.

For illustration, we have invented (presumably) fictitious names for devices and agents. *BTC* is our pet name for a generic smart speaker (Big Tin Can); other names we hope will be self-explanatory. Any accidental use of existing trademarks will be rectified upon being so notified.

Chapter 1 - Assumptions and Scope

1.1 Assumptions and Preconditions

Definitions for this section:

Assumption refers to mental models of the world and in particular limits placed on those models upon which OVON standards will be created.

Precondition refers to the real-world state that must exist at the moment when a system which was constructed in compliance with OVON standards attempts to perform its mission.

Assumptions pertaining to OVON and its work:

1. We assume that OVON's work, consisting of standards, policies guidelines, and other guidance intended to foster an open voice ecosystem, does not extend to "policing" or "enforcement" of said policies and standards.

2. Taking assumption 1 above into account, we assume that we need only define standards which apply in expected operating conditions; in other words, we assume that all system components will behave cooperatively and ethically.

Assumptions and preconditions pertaining to voice systems and their components:

3. Only one system component (an "agent," see definitions) will "speak" to the human user at a time. For brevity, we call this condition "having the floor."

   Discussion: Unlike visual interfaces where a human is able to discern and choose among

multiple visual objects being presented simultaneously, the human auditory system has more challenges discerning among multiple concurrent audio dialog streams emanating from a single device (the "endpoint"). However, recent research[2] [3] suggests that there may be ways to allow multiple streams to co-exist, using a principle known as the cocktail party effect[4]. While it may be possible for the system to deliver multiple simultaneous audio streams, e.g. speech plus music plus alarms, we will continue to assume only one speech output at a time for the purposes of creating standards.

4. The system architecture for which we are producing specifications allows for independent creation of content and generation of audio streams. This implies that the audio characteristics of the audio delivered to the human (voice, pitch, volume, natural language, etc.) are potentially created or modified independently from the content generation process. In particular, technologies such as ASR and TTS capabilities are becoming stand-alone services and OVON standards must not interfere with system design or innovation in this regard.

5. Irrespective of the limitation of a single agent having the "floor", multiple agents may be able to listen simultaneously to any human user interaction or to each other. OVON interoperability architecture standards must not preclude the implementation of such a solution, even as use of this capability will be subject to guidelines issued by other working groups, including but not limited to privacy, security, and ethical use.

6. Agents may wish or need to share dialog history, data, and context. OVON interoperability standards must provide for this capability, subject to guidelines issued by other working groups, including but not limited to privacy, security, and ethical use, but will neither mandate how such data and context is to be used nor require agents to share anything.

7. Irrespective of any shared dialog history, data, and context, agents may wish to maintain their own context. OVON interoperability standards must not preclude such choices.

8. The voice ecosystem for which we are creating standards can be multilingual even within a single interaction, irrespective of whether such a configuration is technically feasible today.

[2] https://www.jneurosci.org/content/40/34/6613
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3834087/
[4] https://en.wikipedia.org/wiki/Cocktail_party_effect

9. Agents may wish or need to simply convey requests or responses without interpretation, for any reason. OVON standards must neither mandate nor preclude such choices.

10. Interaction between agents is not necessarily symmetric, with respect to authorizations, trust, control, data, etc.

11. Interaction of any kind between agents, with the exception of speech generation (see assumption 3), can be nested with no theoretical limit to the depth of nesting. Any patterns of interaction between two agents (e.g. A and B) can be repeated between B and C, between C and D, and so on until an agent that can perform the given task has been reached. OVON standards will allow for the existence of nesting but will not mandate its implementation or any particular mechanisms for managing nesting.

12. The human user maintains ultimate control over execution. OVON standards must accommodate the existence of mechanisms for interruption / cancellation of voice operations, without specifying either the user interface for triggering such mechanisms or the implementation of the mechanisms. OVON guidance will provide recommendations for such cases, including the principle of "help keep the user out of trouble", but as stated above, OVON itself cannot enforce such principles.

1.2 In Scope and Out of Scope

In full respect of our mission to produce standards for interoperability that promote innovation without unduly placing constraints on any particular implementation, we have set our scope accordingly along considerations that impact interoperability.

The following basic requirements are In Scope for OVON interoperability standards:

- How does the human user acknowledge trust in an agent
    - And as a corollary, how does an agent know that it is authorized (trusted) to perform a given task
    - However any mechanisms for storing or remembering that trust are out of OVON scope
- How does an agent find another agent
    - Note however that the algorithms internal to one agent for *drawing inferences* (e.g. when the user did not specify a named agent) and *choosing a particular agent, e.g. B,* are out of scope
- How do agents determine that they can trust each other

    - Recall that this is not necessarily symmetric
- How do agents share context and other data securely
- How do agents negotiate having the "floor"
- What interaction is necessary when an abnormal condition occurs, including a human request to interrupt, cancel, abort, "go back", undo, etc.
    - This includes mechanisms of some sort for "signals" between agents, including time-outs
    - However that the mechanism by which agents *recognize the user's desire* is out of scope

The following system design considerations are Out of Scope for OVON <u>interoperability</u> standards

- Handling of requests by native functions, including how an agent knows that it is able to perform a given task
- Formats of vocal commands and responses including wake words
- Persistence of context within agents
- Which level of data abstraction "should" be used for interactions
- Which ASR, dialog manager, or TTS "should" be chosen in any given situation
- How an agent handles multilingual commands and output
- How an agent draws inferences
- How an agent presents and chooses among multiple choices for fulfilling an implicit request ["voice search" function]
- Considerations with respect to "good" user experience, including any attempt to define or describe "good"
- How to detect unethical or inappropriate agent behavior and what action to take as a result
- Enforcement of any standards, policies, or guidance

Note that many of these items will be addressed in *other* OVON initiatives such as Privacy, Security, Ethical Use, and Discovery & Location Services.

Chapter 2 - Basic Patterns of Interaction

The Open Voice Network Architecture Work Group has identified several architectural patterns that, when combined, could enable seamless intent fulfillment for users of different conversational assistants and agents.

We call these patterns native, *mediation*, *channeling*, and *delegation*.

2.1 Interaction Patterns

The illustrative cases which follow involve the user interacting with up to two agents, which we will call agents A and B. For simplicity, we are limiting these descriptions to two agents; in the general case, there could be multiple agents involved in the conversation. Note that we are not suggesting that any particular agent needs to operate in all the modes described here.

To illustrate these cases, we will describe the sequence where the user wishes to know the price of a kilo of apples. In these sequences, the user initially interacts with Agent A via an endpoint which merely provides audio capabilities (microphone and speaker). For clarity, the endpoint is not included in any of the following descriptions.

Case 1 - Native: Agent A may use internal services or programmatic interfaces to fulfill a user request, without using the services of another conversational agent.

The user asks Agent A “how much do apples cost”; Agent A:

- interprets the request,
- consults an internal or external service known only to the publisher of Agent A,
- vocalizes the answer to the user.

As there is no interoperability with another conversational agent occurring in this case, it is outside OVON scope.

Case 2 - Mediation: Agent A, acting as a user, has a **conversation** with Agent B behind the scenes using dialog – semantic or linguistic – interfaces to achieve a goal and return to the user. E.g. make an appointment. Agent B does not **directly** interact with the user, does not interpret **user** utterances or formulate output responses. Agent B could potentially be a human! The user of Agent A might or might not know that this is happening. In addition, Agent A may choose to share user input with Agent B, and share Agent B output with the user, but such considerations are **out of scope for OVON**.

The user asks Agent A “How much do apples cost?”; Agent A:

- interprets the request,
- decides that it cannot fulfill the request itself,
- identifies and locates an external source for obtaining the response (Agent B),
- holds a private conversation with Agent B,

- vocalizes the response to the user.

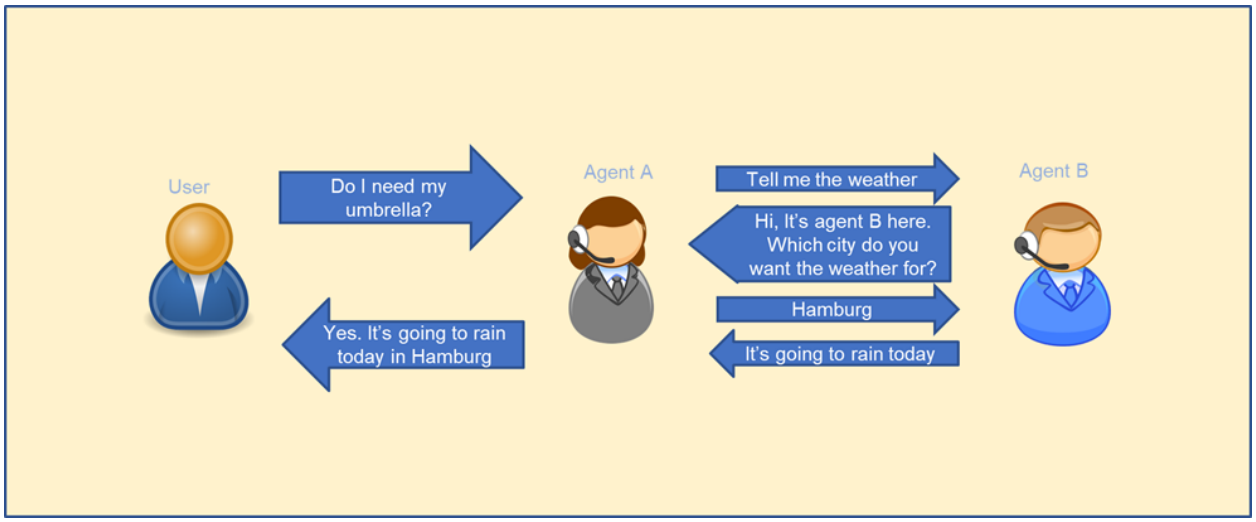

Figure 1: *Mediation* – Agent A acts as a dialog user to another agent to fulfill a goal for the user. Agent A stays in full control of the dialog with the user.

Case 3 - Channeling: Agent B provides dialog services to Agent A to achieve a task. The interaction could be single- or multi-turn. Agent B interprets utterances and formulates utterances, but Agent A may modify the "character" of the utterances delivered to the user. The user might or might not know that this is happening.

The user asks Agent A "how much do apples cost"; Agent A:

- interprets the request,
- decides that it cannot fulfill the request itself,
- identifies and locates an external source for obtaining the response (Agent B).

Agent B then:

- reinterprets the user request if necessary,
- formulates questions or ongoing responses to Agent A which agent A may modify or enhance when interacting with the user,
- responds to subsequent user inputs that are passed to it via agent A.

This continues until Agent A decides to terminate the interaction or Agent B informs Agent A that it has no more responses to give or questions to ask. Agent A remains in control of the dialog throughout the process.

One situation where channeling might be used is when Agent A and Agent B use different natural languages (such as English and German). Agent A is responsible for translating requests and responses from Agent B.

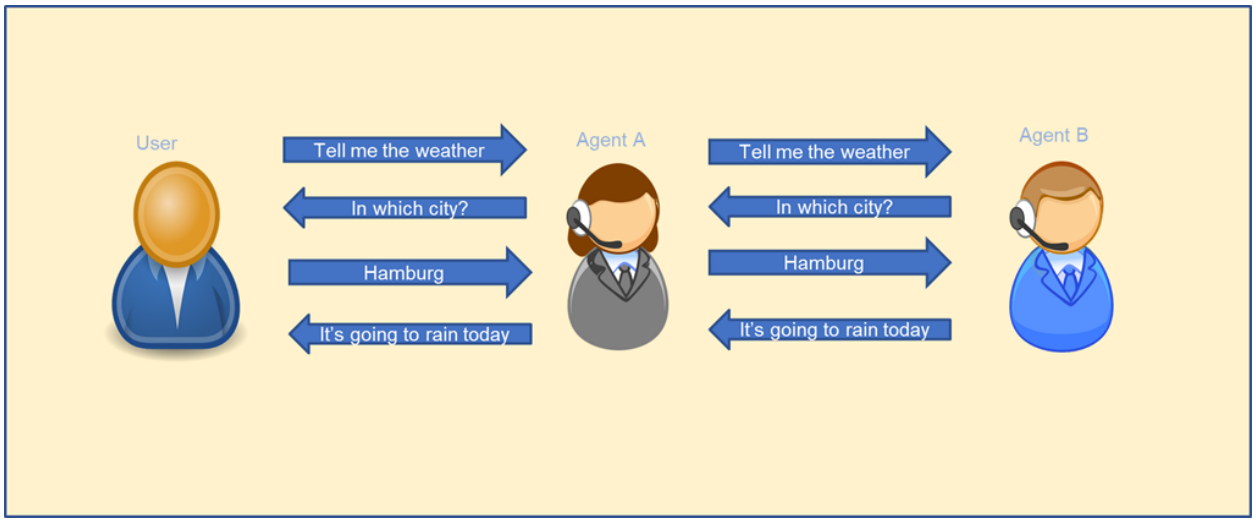

Figure 2: *Channeling* – Agent A passes user requests on to Agent B, receives the answers and passes them back to Agent A. Agent A might modify the responses but the conversation is driven by Agent B.

Case 4 - Delegation: Agent A passes control and management of the dialog to Agent B, along with a negotiated amount of context and dialog history. The user may or may not be aware [assumption to be validated] of the transfer but may have requested it. [it is good practice that the user is aware of the transfer of agency but this is outside of the scope of OVON standards.]

The user asks Agent A “how much do apples cost”; Agent A:

- interprets the request,
- decides that it cannot fulfill the request itself,
- identifies and locates an external source for obtaining the response (Agent B),
- may notify the user that further interaction will be managed by Agent B.

Agent B then:

- reinterprets the user request,
- continues the dialog with the user,
- vocalizes the response to the user.

At the end of the exchange, Agent B remains in control of further dialog.

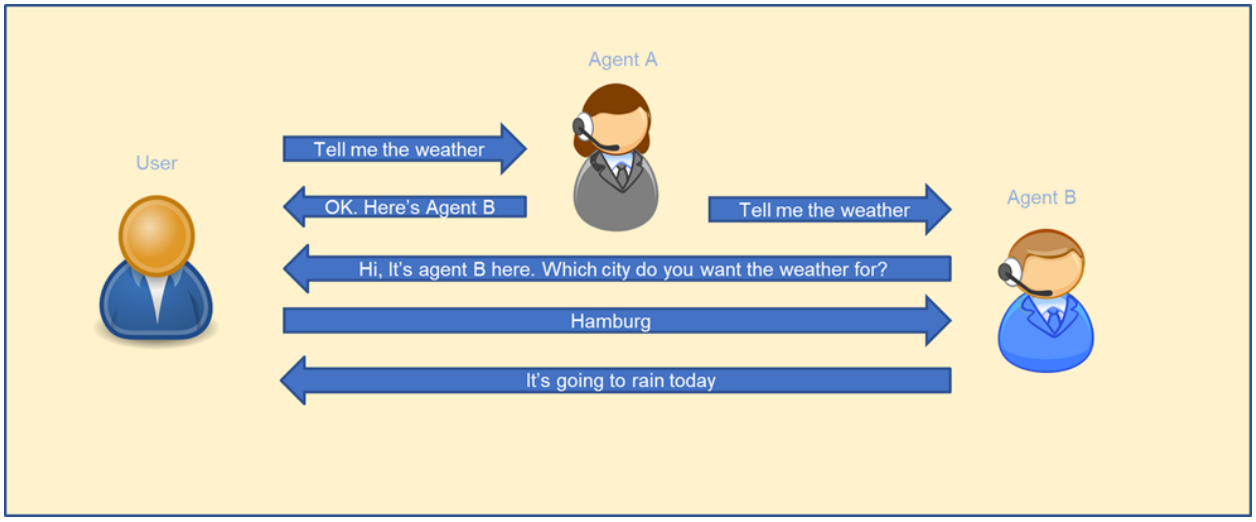

Figure 3: *Delegation* – Agent A passes user requests on to Agent B but then cedes control of the dialog to Agent B who may or may not return it later.

2.2 Informing the Human User

In all cases above, the human user may or may not be aware of the involvement and eventual transfer of control to different agents. The choice of informing the human user or not of the involvement of additional agents is left to the discretion of the producers of conversational assistants. OVON Interoperability Standards will take no stand on whether the human user is notified of the involvement of additional agents in any particular conversation, but other OVON initiatives may deliver guidelines with respect to best practices in this area.

Chapter 3 - Illustration of Interactions

The following examples are described here for illustration purposes only, as command syntax and interpretation are out of scope for OVON interoperability standards.

3.1 Common Steps [illustrations only]

Below is a list of steps that might occur during a multi-Agent interaction.

The following initial steps are out of scope for OVON interoperability standards:

1. *Before this scenario, we assume that the User has activated an Endpoint Client* (smart speaker, IOT device, mobile app, phone, website, etc.) *and that the Endpoint Client has been connected to an Agent, hereafter Agent A, via some connection mechanism*. [Note: the connection between an Endpoint Client and an Agent can potentially have been initiated from either side.]

2. *User* issues a request to *Agent A*

3. Agent A interprets the request, and resolves any implicit aspects of the request [See Chapter 3.2]

4. Agent A decides, potentially informed by user settings, trust levels, privacy concerns, and other factors, that it needs to involve another agent.

The following steps are potentially within the scope of OVON interoperability standards, at least partially:

1. *Agent* A uses a mechanism to determine which Agent[s] it will need to involve.

2. Agent A negotiates with these other Agents, determines the best Agent[s] with whom to interact and negotiates which pattern the interaction will follow – mediation, channeling, delegation. During this negotiation Agent A may provide certain elements of context and dialog history. [See Chapter 3.3]

3. Agent B (and C, etc.) explicitly accepts the request from Agent A to be involved in the conversation, along with confirmation of the pattern of interaction [See Chapter 3.4]

4. Agent A provides Agent B (etc.) with additional elements of context and dialog history, along with information about control flow. [See Chapter 3.5]

5. The dialog with the user continues, according to the type of interaction [step 5 above] as described in Chapter 2.

3.2 Explicit versus Implicit Requests

This chapter is supplied for illustration purposes only; OVON Interoperability Standards will take no position on how an Agent resolves implicit references in human speech.

The following is a non-exhaustive and evolving list of possible user command patterns, noting that the actual syntax will depend on implementation decisions and is thus out of scope for OVON interoperability standards.

- Explicit user request to transfer control to Agent B:
    - Examples: “BTC, please connect me to Big Grocer,” “Big Grocer, please connect me to PayHelper.”
- Explicit user request to pass a task (not control) to Agent B:
    - Examples, “BTC, please ask MusicApp to play my playlist,” “TravelHelper, please ask MyHotel if there are any rooms available in Pittsburgh on the 24.th”
    - Note that Agent A is not required to understand the embedded request, for example “BTC, please ask *Météo France* ‘quel temps fera-t-il demain à Marseille’.”
- Implicit request to transfer control to another Agent:
    - Examples: “BTC, please connect me to my usual grocer,” “Big Grocer, I would like to check out now.”
- Implicit request to pass a task (not control) to another Agent:
    - Examples: “BTC, please order my usual pizza,” “TravelHelper, I want to reserve a hotel.”

3.3 Negotiation Between Agent A and Agent B

As part of the negotiation, Agent A sends a request to Agent B asking if it is willing to participate in the conversation. The request has a payload which can include:

1. A representation of the task to be performed as requested by the user, possibly enhanced by Agent A or replaced with another request (at semantic level 0, 1 or 2 – see the whitepaper for description of these levels)

2. Information about what has been happening in the conversation previously (Context and/or Dialog History)

3. Other Information about the user (User Data)

4. Information about Agent A (such as security and privacy)

5. Unique ID for the conversation

6. Additional payload TBD

In response, Agent B may express its capabilities with respect to its ability to participate in the conversation. These may include:

1. Is Agent B authorized to accept this type of request?
2. What semantic level of data can Agent B accept?
3. What semantic level of data can Agent B provide?
4. What types of context and dialog history does Agent B require?
5. Additional considerations TBD

3.4 Does Agent B Accept the Delegation?

An acceptance of participating in a conversation depends on Agent B saying "Yes" to questions such as the following, which are provided as illustration only. Obtaining and confirming this consent is recommended by OVON interoperability standards although the specific considerations might be out of scope for OVON interoperability standards.

1. *"Do I understand the request?"* → i.e., is the meaning explicit or can it be derived from the request context. For example, the user asks "What's the weather ***here***" and Agent A has informed Agent B that the user is in Washington DC
2. *"Can I respond to this request?"* → i.e., do we know what kind of answer or clarifying question is expected. For example, a grocery agent should not be handling weather Requests.
3. *"Do I trust Agent A"* → i.e. can Agent B interact with Agent A given privacy/security Standards
4. *"Do I trust/am I allowed to talk to the end user"* → i.e., can Agent B interact with the user, given its privacy/security standards or market conditions. For example, if Agent B can only give responses to registered users.
5. *"Are my resources available?"*

6. Are the privacy policies satisfied by the request, context transmitted with the request, and data to be shared from agent A to Agent B satisfied?

The response from Agent B to Agent A might be as simple as “yes” or “no”; further exploration of this topic is necessary.

3.5 Handover

Once Agent B has confirmed its consent for participating in the conversation, Agent A will provide additional payload including items such as:

1. Additional context and/or dialog history
2. Additional information about the user and their preferences
3. Information about what to do after transaction is complete (ie: end conversation, delegate back to Agent A, delegate to Agent C)
4. Information about what to do if an error occurs/transaction has to be interrupted
5. And any other information that was negotiated

3.6 Collaboration Among Multiple Agents

We are aware of a special case wherein multiple agents listen simultaneously to the same conversation. For example, if the user is planning a trip, both a hotel agent and an airline agent might be listening simultaneously to dates and destinations, each providing their own output with respect to availability and prices, with the possibility for the user to then say something like “book rooms and flights for <date d> at <time t>”.

This case is not yet described.

Chapter 4 - Open Topics

We are aware of the following topics that have yet to be fully explored to this point.

1. How linguistic data are passed from Agent A to Agent B

2. How linguistic data are passed from Agent B to Agent A
3. How other data, including but not limited to, context and dialog history are shared
4. between Agents, and in particular how privacy and security mechanisms are applied
5. How interpretation and execution are shared among multiple Agents
6. How control flow is managed, both in “normal” and “abnormal” cases, including user control to interrupt, cancel, and “go back”
7. Location services, how does one Agent find another
8. Discovery services, how does an Agent make its availability known

Appendix: How Does a Delegated Agent Interact with a User?

This chapter describes how a user interacts with a delegated agent.

OPEN VOICE NETWORK
Voice worthy of user trust

Delegation Sequence