AI Voice Agents: Use Case & Evaluation Template

Voice Agent Use Cases

Guidance on Use Case Selection

- Focus on high-volume, low-friction phone interactions where users typically seek quick answers or task completion.
- Avoid high-security or sensitive use cases that require multi-step user verification or strict authentication flows (e.g., password reset, account unlock).
- Choose tasks that can be fully completed via voice and do not rely heavily on visual input (e.g. images).
- For use cases that require user input:
    - Recommend use of free text fields and numeric/date time fields
    - Avoid reference fields or fields with lots of choices
    - Avoid dynamic dependencies between fields
- Ideal candidates include:
    - Answering FAQs or knowledge base questions
    - Common HR requests (e.g., Help with downloading payslip, changing benefits, generate employment verification letters, submitting PTO)
    - Creating a new case or request
    - Checking or updating the status of an existing case
    - Simple HR queries (e.g., “How many PTO days do I have?”)

Use Case Details

Use Case Name:

Goal:

- What problem are we solving? Who is the primary persona?

Volume:

- Anticipated query volume based on traffic
- Concurrent calls

Authentication

- Is user authentication required to complete this use case (Y/N)?
- If Y, what type of authentication?
    - SMS
    - Soft PIN
    - Knowledge based questions
    - Authenticator app
- Are the phone fields (of sys_user table) saved in the E.164 format?

Expected Flow

- What might a user say to initiate this flow? Trigger utterances?
- What are the expected steps in the conversation, from start to finish?
- What inputs or information does the voice agent need to collect?
- What actions are triggered during the flow?
- How is the outcome confirmed to the user?

Sample conversation

[User]:

[Bot]:

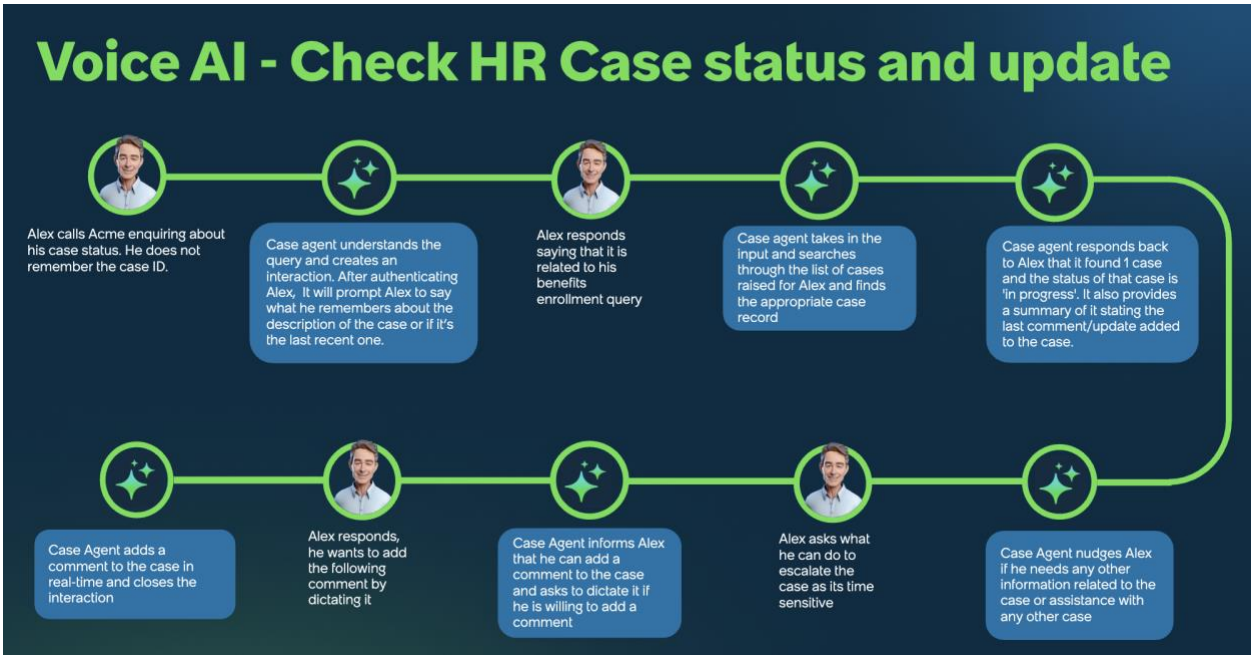
..

.. *(Refer to test data set for case status modeled conversation example)*

Data & Tools

- What data is required to fulfill this use case? (i.e. knowledge bases, user records, etc.)
- Where does this data reside?
- What tools, APIs, or workflows need to be invoked?
- RAG/Search, Notify, Case API, etc.

Flow Diagram



Integration with 3rd Party Systems

- Is integration with external systems required? If so, which ones?

Integration with CCaaS/IVR

- CCaaS Name: AWS Connect/Genesys/3CLogic
- IVR (if diff from ACD): Lex/Dialogflow/calltree/Number direct
- Integration type: Webhook/WebSocket/SIP/gRPC
- Helpdesk Number Separation: Region wise/Department wise/Service wise
- Entry Point: front line?
- Exit Point: conditional branching?
- Relay of information: Query context? Authentication?
- Fallback method: Live agent/Case create/None?

Interaction record

- Transcript
- Recording and summary
- Live agent Context

- CTI integration?

Languages

- Primary Language: English/German?
- Other Languages: Spanish/Japanese/French Canadian/BP?

Expected Outcome / Success Criteria

- What does success look like?
- How will you measure success?
- Are there any fallback rules (i.e. escalate to live agent or create a case under certain conditions)?

Golden Dataset and Evals

- Curated diverse utterances and request response/turns - modeled conversations
- 30 to 50 per usecase
- Validate Agent matching, Tool call success, Task completion rate, Slot/input recognition, helpfulness, Conversation quality, Transcription accuracy, Dialects
  **Sample Dataset for Case follow up OOB usecase:
  https://code.devsnc.com/dev/cbwf-qe-tools/tree/master/tools/voicetestdatagenerator/coval_automation...

Success Criteria

Key Dimensions

There are a few key dimensions to consider when evaluating the success of a voice agent:

- **Was the call successfully handled?** This includes whether the correct voice agent was triggered, the response was accurate and relevant, the right tools or workflows were executed, and fallback mechanisms were triggered appropriately when needed.

- **Was the voice experience high quality?** Consider how natural and clear the voice sounded (TTS), how accurately user speech was transcribed (STT), and how well the bot handled real-time voice behaviors like pauses, interruptions, or corrections.
- **Was the system responsive?** Evaluate latency and whether the interaction felt smooth or had noticeable lags.

Recommended Success Scorecard

This scorecard reflects **ServiceNow's recommended baseline for evaluating voice assistant performance**. You can evolve it over time and add your own metrics that align with your specific goals and use cases. As you test the voice assistant, fill out the scorecard for each call and capture detailed notes on the quality of the interaction, including any issues, anomalies, or notable strengths.

We recommend evaluating **30–50 test calls per voice assistant**, covering both expected and edge-case scenarios. Distribute calls across different testers, speaker profiles, and use cases to capture a representative view of performance across diverse real-world conditions.

Metric	Score Options	Guidance	Recommended Target
Call Success (North Star)	3 = Fully successful 2 = Partially handled 1 = Failed	Did the agent complete the task or hand off correctly when needed?	70%
Agent Detection	Yes / No	Did the correct voice agent get triggered based on the user's initial request?	80%
TTS Quality	Yes / No	Was the bot voice clear, pleasant, and natural-sounding?	80%
Conversation Handling	Yes / No	Did the bot handle pauses, corrections, or user interruption naturally?	80%
Response Speed	3 = Fast/Great 2 = Acceptable 1 = Too Slow	Was the system's response time acceptable for a smooth conversation? Note any noticeable lags.	80%