Effective Enterprise Voice Technology Solutions: A Primer

A successful enterprise voice journey starts with clearly understanding the range of technology components and options, and often includes selecting a suitable solutions integrator.

Executive Summary

As voice automation goes increasingly mainstream, enterprises are looking at avenues to drive additional efficiency and save money. As with most technological pursuits, that’s easier said than done.

The emerging enterprise voice technology solutions space is comprised of a wide range of applications, including natural language self-service call center applications, sophisticated voice biometric applications and automatic speech transcription solutions. It’s also a space with numerous standards and evolving products. This complex landscape makes it challenging for the enterprises to take the “first right step” in selecting the most appropriate voice technologies.

In this white paper, we present a product-agnostic view of the voice applications landscape. We introduce the reader to the gamut of solutions and describe the best ways for navigating and embracing them.

Types of Voice Applications

Voice applications can be classified based on the nature of interactions users have with them. Figure 1, next page, breaks this down by broad classifications and typical applications with each category.

Interactive Voice Response (IVR) Applications

These are voice applications that are typically used to reduce a call center executive’s involvement in servicing calling customers. IVR solutions range from ones that respond to user inputs from the dial pad to ones that can handle natural language like speech inputs.

- **Non-voice input/voice output:** These are systems where the users interact with voice applications through a PSTN telephony system. Touch tones generated by the dial pad are the means of providing input. The system responds to a user’s input with appropriate prerecorded voice responses, generated using a voice synthesizer. Such IVR systems are used...
in call centers to identify and segment callers before they are routed to an appropriate call center executive.

• **Voice input/voice output:** These solutions allow users to provide input with spoken words instead of dial-pad-generated DTMF tones. The user can interact with these applications hands-free throughout the transaction. The level of sophistication of these applications ranges from supporting a predefined set of voice commands to supporting natural speech such as sentences as inputs. Sophisticated call center IVR solutions that can “steer” the caller to the right support personnel based on spoken input are examples of this type of application.

**Dictation Applications**

These applications create transcripts from speech inputs. Such solutions are used by transcription services to build in automation into the transcript creation workflow.

These differ significantly from IVR applications as they don’t “speak back” or “ask” for inputs. They are instead designed to interpret everything that is spoken and generate a text equivalent.

Voice-automated transcription solutions are an example of this kind. These solutions are mostly industry-specific to address the complexity of interpreting jargon in the spoken input.

**Voice Biometrics**

This class of voice applications uses voice as a substitute for traditional authentication mechanisms such as a PIN or a password.

Voice biometric systems convert the caller’s voice into voiceprints, or unique algorithms based on the specific characteristics of the voice, which are even more unique than fingerprints.

This set of solutions comprises two broad subcategories:

• **Voice password solutions:** These solutions require the user to enroll with the voice biometric system through using a predefined spoken phrase that has to be repeated for identity verification during subsequent access.

• **Conversation-based voice authentication:** These make the authentication process transparent to the user as the user’s identity is established in the background when he is in conversation with the call center executive.

**Speech Analytics**

These solutions are used to extract valuable information from voice recordings. They are typically used by contact centers in analyzing recorded calls to discover avenues for increasing operational effectiveness.

**Voice Technology Application Landscape**

![Diagram of Voice Technology Application Landscape]

*Figure 1*
Speech analytics solutions can help in gaining insight into the following:

- Customer satisfaction levels.
- Customer intent insights.
- Maximizing opportunities for making contextual sales.
- Developing effective training for improving live agent performance.

**Automatic Speech Recognizer: The Heart of Voice Applications**

Since approximately 80% of the voice applications discussed above depend on the ability to generate the text equivalent of the spoken input, converting speech to text is clearly job one. This function is typically handled by the automatic speech recognizer (ASR) component of any voice solution. The ability to tune the ASR accurately determines the success of the voice solution.

To understand the various subcomponents that make up the ASR, let’s examine how a sample phrase is processed by the ASR (see Figure 2). The flow starts with a user speaking a phrase into the computer’s microphone and ends with the ASR detecting the text equivalent of the spoken phrase.

### Building Blocks of a Speech Recognizer

![Diagram of Speech Recognition Engine]

- **Acoustic Model**
  The acoustic model is used to break down a digitized speech input to its pronunciation equivalent. This pronunciation equivalent is represented using “phonemes.” Every language has a finite set of phonemes. These phonemes can be used to represent any spoken word in the language.

  While the set of phonemes that constitute a language is predefined, the digital representation for a given phoneme may differ based on the usage context. This possibility of multiple representations for a phoneme results from colloquial variations, differences in dialect and nuances of tone.

  The same phoneme can therefore have different representation values based on the acoustic model applied. Hence, to ensure accurate speech-to-text results, it’s critical to have an acoustic model that fits the business’s needs.

- **Dictionary**
  Dictionary is a component that stores a collection of words mapped to their phoneme equivalents.
All the words in the output of the speech-to-text conversion should have entries in the dictionary.

The following are a few examples of the phoneme representation of words.

<table>
<thead>
<tr>
<th>Word</th>
<th>Phoneme Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>HELLO</td>
<td>HEHLOW</td>
</tr>
<tr>
<td>FAR</td>
<td>F AA R</td>
</tr>
<tr>
<td>FOOT</td>
<td>F UH T</td>
</tr>
</tbody>
</table>

The accuracy of the conversion is a function of the number of words configured in the dictionary. As shown in the above, a logical collection of phonemes is searched against a dictionary to detect the equivalent word.

Creating a dictionary that best fits a speech recognition requirement may require extending a dictionary with entries for new words. These new words should address the jargon or any business-specific words that may be used in the speech input. There may also be a need to edit the phoneme representation of an existing word to address colloquialism.

**Language Model**

A language model, as the name indicates, is a representation of the usage of words that make up a language. The goal of the language model is to aid in the detection of meaningful phrases rather than just individual words.

The premise for this kind of modeling is that words are not used in a random order when spoken.

The goal of the language model is to predict the spoken phrase based on a detected word. This prediction is represented using a probability value assigned to words. This reliance on probability ensures that meaningful phrases, and not just words, are detected by the speech engine.

There are two categories of language models that can be used by a voice application:

- **Grammar-based language models**: These models are less intensive to create and train.

But their capability is limited, too. They are used in applications that are expected to handle only a finite set of command phrases. An example of a command phrase is the user speaking out the menu option he wants to choose, in a voice-enabled IVR system.

- **Statistical language model**: SLM is more sophisticated and powerful than grammar-based language models. These models can handle conversational-style natural language inputs. An example of the application of SLMs is in “dictation” type voice applications, where a user can dictate any sentence to the transcription engine.

**Significance of Training**

Most ASRs come with a default set of language, acoustic models and dictionary. However, these models may not readily suit the requirements of the business given variations in use, tone and presence of jargon. It’s therefore critical to train these models until a satisfactory level of speech recognition accuracy is achieved before they are deployed in real time.

This training needs the creation of the right corpus of representative samples and the application of the right tools. You should be especially mindful of such considerations – along with a few other critical ones – at the early stages of the enterprise voice journey.

**Three Initial Steps for Embracing Voice in the Enterprise**

The preceding sections cover the full gamut of available voice solutions and complexities in speech recognition. As all the complexities in the previous sections are product agnostic, they are applicable to any voice solution under evaluation. Given these complexities, the following three steps must be undertaken by any enterprise looking to deploy its first voice solution.

- **Choose the right product partner**: Picking a partner that can scale to your enterprise voice needs is the critical first step. Here are a few considerations to be mindful of when making this decision:
  - **Richness of speech recognition models**: As discussed, effectiveness of speech recognition depends heavily on the availability of language, acoustic models and dictionaries that fit the business need.
Partnering with a product vendor that has a rich repository of “off-the-shelf” models will save costs and time.

» **Variety of deployment architectures:** There are multiple possibilities for deploying a voice product – from cloud-based to behind-the-fire-wall on-premise solutions. The choice of the right deployment architecture depends on nonfunctional requirements such as security and performance. It’s therefore vital to choose a product partner that can support multiple models of deployment.

» **Breadth of tools and standard development kits:** The voice channel for your enterprise can have multiple types of end points. These end points include touch-tone phones, smart phones and desktops. SDKs are therefore needed to integrate the end points with the voice solution. The training of models used by the ASR requires product-specific tools and verification mechanisms. A product partner that offers a variety of SDKs and tools would reduce customization costs.

- **Choose the right SI partner.** It is also critical to team with the right solution integration partner that can help navigate this complex landscape. Having an in-depth understanding of the voice products is a must for the SI partner, but there are other competencies to be considered:

  » **Product-agnostic standards and tools:** As indicated above, enterprise voice is an emerging space. While the availability of multiple standards and tools offers flexibility, it also brings challenges in making the right choice. It’s valuable to partner with a services provider that can objectively evaluate peer standards and tools to suggest the best fit.

- **Prepare for training and tuning as an iterative task.** It’s important to realize “setting up” your voice solution is just the start of your journey in enabling the voice channel for your enterprise. It’s recommended to set aside time and resources to periodically review voice analytics and training and retune the voice solution. This retuning helps the voice system adapt to the observed changes in usage patterns and feedback from customers.

Looking Forward

With the increasing interest in voice automation systems as yet another channel for customers to interface with enterprise systems, the time is right for organizations to begin evaluating and investing in those systems that best serve their business needs. As noted, there is a gamut of solution options.

While choosing the right solution is surely the first step, choosing the technology/product partner to enable voice solution is an even more complex task.

**References**

About the Author

Mohan Vamsi E.V. is a Principal Architect within Cognizant’s AVI Center of Excellence. In this role, he is responsible for incubating and developing voice-related innovations and service offerings, and he has been at the forefront of various voice-related initiatives executed by his group. Vamsi received a B.Tech. in electronics and communication and has 13-plus years of experience developing enterprise-scale IT solutions. He can be reached at MohanVamsi.Eswara@cognizant.com.

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 75 development and delivery centers worldwide and approximately 178,600 employees as of March 31, 2014, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world.

Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.