

Speech Enabled Auto Attendant Design: Using Application Intelligence to Deliver Superior Voice Recognition

Abstract

Speech recognition is an important way to interact with computers and communications devices, freeing your hands and expediting collaboration. A speech-enabled Automated Attendant (SEAA) solution can have many designs, but a typical one has three main components: the speech-enhanced user interface, the directory (or grammar), and the speech engine. The user interface provides the dialogue you hear and the manner in which the SEAA interacts with you to determine the person you want to reach. The grammar is the directory of allowed responses for the SEAA application. The grammar for most SEAs is an employee directory, with phone numbers and other information. A speech engine is a background calculator that uses computational algorithms to recognize speech and deliver information to the SEAA application. It is important to note that a properly designed speech solution is much more than a speech engine; it must have an intelligent application on top of the speech engine in order to deliver accurate results and an effective user experience.

Introduction

Business owners and employees today conduct business worldwide 24 hours a day using seemingly infinite combinations of phones, voice messaging, e-mail messaging, fax, mobile clients, and rich-media conferencing. Without unified communications, however, these tools are often not used as effectively as they could be. Information overload and misdirected communications delay decisions, slow down processes, and reduce productivity. Unified communications solutions can save time and help control costs while improving productivity and competitiveness.

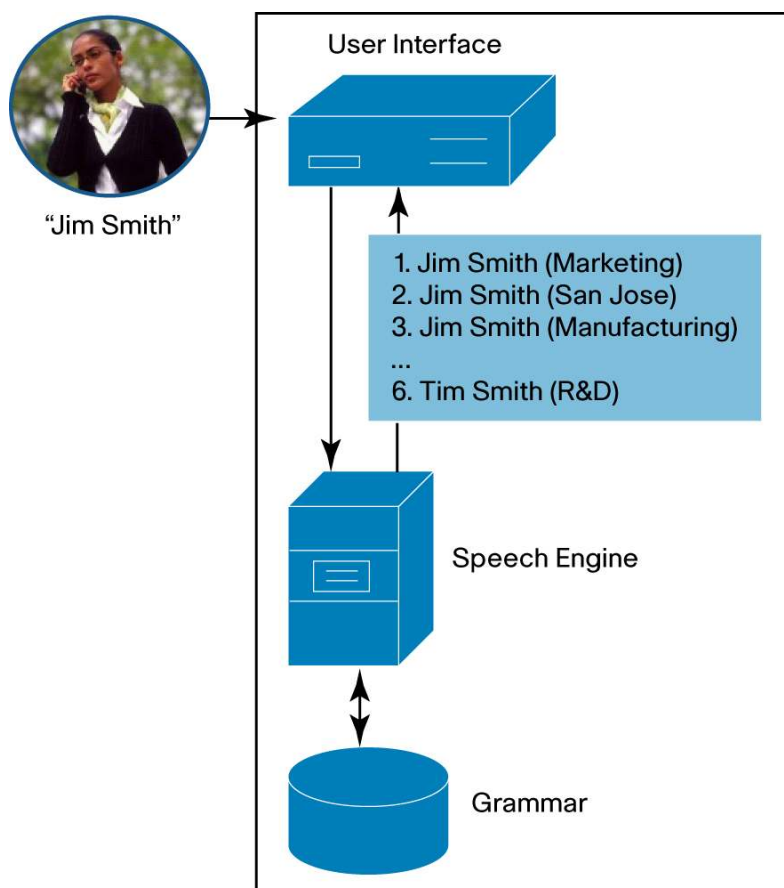
As unified communications solutions integrate applications, phones, and computers, speech recognition plays an increasingly important role in the way we interact with these devices and applications. Speech recognition frees our hands and lets us control our unified communications experience with spoken commands instead of memorized, menu-controlled clicks, keystrokes, and button pushing.

Yet speech recognition solutions, for a variety of reasons, have failed to evolve to maximize the effectiveness of unified communications solutions. In particular, many Automated Attendant products have added speech recognition to improve the user experience and increase customer satisfaction by allowing customers to use natural language to command the attendant to direct their calls. But whether from a lack of a complete and unified solution, underdeveloped application intelligence, or limited integration to other places in the data and communications network, speech-enabled Automated Attendants have not delivered a time-saving, customer-satisfying experience.

Some of the shortcomings in SEAA solutions can be attributed to the fact that there are numerous approaches to designing the solution. A typical SEAA solution is made up of three key components: the speech-enhanced user interface, the directory (or grammar) and the speech

engine (Figure 1) and the features and interactions among the three components of an SEAA differ, depending on the designer.

Figure 1. SEAA Architecture



The user interface provides the dialogue you hear (“Who would you like to reach?”) and the manner in which the SEAA interacts with you to determine who is really intended (“Did you mean Jim Smythe?”). Just about every SEAA will interact differently with you, all with varying levels of success. The user interface is a critical element for delivering a satisfying experience. Some solutions offer a “natural-language” interface where you are allowed to speak any phrase. However, you are better served by a well-structured dialogue where you know what you are supposed to say and which commands to give by the style of the questioning. For example, “Who would you like to reach?” will result in less confusion among callers than “Hello, what would you like to do?”

After the user interface prompts you, it accesses the grammar, the directory of allowed responses for the application. For most SEAs, the grammar is an employee directory, with phone numbers and other information, but it may also include product names, locations, conference rooms, or any other person or place a caller may be directed to. Features of an effective grammar are discussed later in this paper.

When the user interface collects a spoken name, it passes the name to the speech engine in the background. Some commonly used speech engines are provided by Nuance, IBM, Microsoft,

Loquendo, and Sphinx. All speech engines are essentially based on the same computational algorithms and deliver similar results.

The Speech Engine: A Powerful Core

A speech engine is a background calculator — it needs a “brain” to deliver a complete and useful user experience. The SEAA application and its user interface sit above the speech engine, acting as a layer of intelligence. A speech engine powers the SEAA solution much like an automobile engine powers the automobile. They share similarities in that all engines perform essentially the same function, just at varying speeds and levels of power. Much like automobile engines are tuned differently for a race car versus a family sedan, speech engines can be configured in a variety of ways to compensate for background noise, pronunciation error thresholds, and other characteristics that may affect the overall accuracy of the SEAA application.

Advanced Disambiguation: Intelligence Above the Speech Engine

Whereas the speech engine provides the computational power, the SEAA application controls the user experience. Understanding the difference between the engine and the application is critical. Without the application intelligence, the speech engine cannot deliver a robust user experience, cannot improve over time, and will not add value to your organization. This scenario is similar to the way a powerful automobile engine without an effective cooling system or an adaptive steering system will ultimately fail the driver through changing conditions.

In general, when a speech engine receives your utterance from the user interface, it analyzes the audio content and attempts to match the audio to the stored pronunciations in its grammar. It develops a list of possible matches and passes that list back to the user interface for presentation back to you. At that point, the user interface proceeds through a critical process — disambiguation — the dialogue through which your final choice of calling destination is made. When multiple employees have the same name, the speech engine begins the disambiguation process and asks you to provide additional information. The disambiguation process can be handled in a variety of ways, so it is critical that an intelligent, user-centric dialogue be part of the disambiguation process or the SEAA application will not quickly deliver accurate results.

To illustrate, following is an example where you speak “Jim Smith” to a SEAA user interface with a grammar of a typical large company’s employee directory. In this example, the speech engine produces a results list of six names:

Jim Smith (Marketing, Chicago, Ill.)

Jim Smith (Marketing, San Jose, Calif.)

Jim Smith (Manufacturing, location unknown)

Jim Smith (Product Management, San Jose, Calif.)

Jim Smyth (Sales, London, United Kingdom)

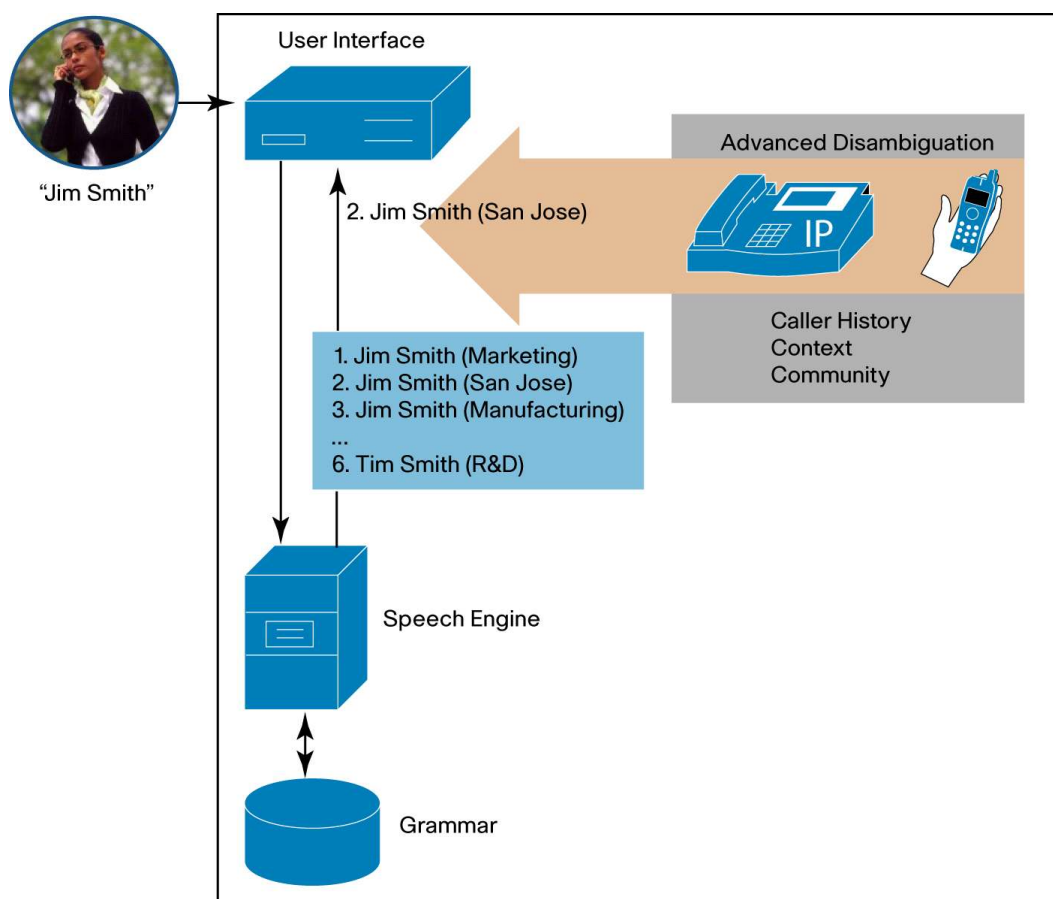
Tim Smith (R&D, San Jose, Calif.)

The typical SEAA product then presents these results to you through a dialogue such as “Press 1 for Jim Smith in Marketing, Press 2 for...” or “Did you mean Jim Smith in Chicago? Press 1...” This approach fails in most organizations, not because you have to work your way through all six results the first time you participate in this dialogue, but because there is no change in the procedure after the 100th time you are presented with this dialogue. For as long as you speak “Jim Smith,” you will

always have to tolerate the same interaction. Before long you will fall back to dialing numbers, meaning the SEAA product failed to deliver any value to you and your organization.

To combat this problem, you need a layer of intelligence on top of the typical SEAA architecture (Figure 2). Advanced disambiguation adds intelligence to the user interface, learning from past disambiguations and applying reason to reduce your time spent (and level of frustration) connecting with the people you are trying to reach. After all, at the very heart of unified communications is the efficiency and timeliness of collaboration between people and among teams. The Gartner Group states “The largest single value of UC is its ability to reduce ‘human latency’ in business processes.” (Magic Quadrant for Unified Communications, Bern Elliot, Gartner RAS Core Research Note G00150273, 20 August 2007) In fact, many organizations have identified “unintelligent” repetition of results lists as a reason to avoid speech-based user interfaces.

Figure 2. SEAA Architecture with Advanced Disambiguation



An advanced disambiguation layer of logic and knowledge should become even more fine-tuned over time. Advanced disambiguation can be compared to the cabin of an automobile, where the driver’s experience is heightened by adjustable seats, climate control, navigation system, and a stereo system — all of which can be automated and fine-tuned to the driver’s preferences. Even better, cars today can store and restore a driver’s preference for seat adjustment, mirror placement, and cabin temperature at the touch of a button. As you use a SEAA, it should be able to better choose the person you are looking for over time, remembering your preferences and customizing your experience. The goal of advanced disambiguation is to direct the results list to

one person every time, thereby reducing the latency of human collaboration across the network and the enterprise.

The ideal advanced disambiguation process should therefore fine-tune your experience over time, building a history of your prior interactions with the SEAA application and using this knowledge to weed out unlikely choices in the results list that the speech engine provides. In Figure 2, the advanced disambiguation logic has concluded that you intend to reach Jim Smith in San Jose. An intelligent SEAA should use numerous sources of information to understand your preferences -- personal address books, buddy lists, phone distribution lists, departments, locations, and even call records. Over time, an intelligent SEAA with advanced disambiguation will construct an accurate grammar of those you are most likely to call. Only with proper disambiguation logic will your speech interfaces be highly accurate, reducing your frustration and enhancing your unified communications experience.

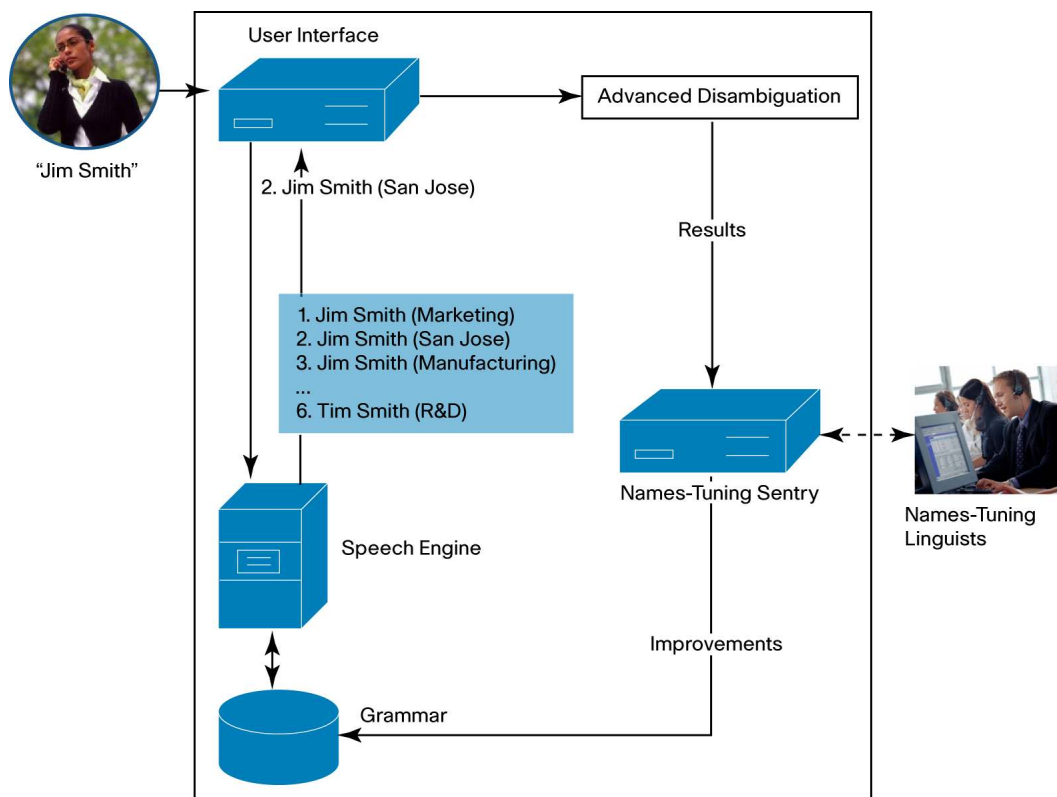
Returning to the example in Figure 2, consider what will happen if you want to call Jim Smith in Chicago. Speaking the name “Jim Smith” normally leads the SEAA to transfer you to Jim Smith in San Jose because of the compiled history of your SEAA interactions. An intelligent solution allows you to say “cancel” to the immediate connection to Jim Smith in San Jose and then prompts you with the original list of six choices so you can then select Jim Smith in Chicago.

Continuous Improvement Through Names Tuning

Advanced disambiguation is essential, but there is even more than the logic to ensuring a successful user experience. When a SEAA is installed, the grammar is compiled from master dictionaries of known names and pronunciations. But in today’s business world of many dialects, pronunciations, and nicknames, and varying audio quality conditions, mismatches can occur with a SEAA. You will become frustrated with a SEAA if you are misunderstood and the error remains in place for months. Even if the SEAA product offers an interface through which you can add pronunciations, you will not always have time to report every failure. (People rarely take the time to click a button to make an online report of a software failure, so you are unlikely to interrupt the process of calling a colleague to report a pronunciation error.) Furthermore, many corrections to the grammar require uncommon linguistic and speech skills. After all, if you knew the proper pronunciation of the person’s name, you would certainly use that instead of a mispronunciation.

A unique approach to solving and updating pronunciation errors is using a combination of automated software and a team of professional linguists to deliver timely error correction — a names-tuning service (Figure 3). An ideal solution uses a SEAA “sentry” program to collect the results of disambiguations from the SEAA along with your actions, sorting the information and routing the records to linguists. The service should provide data collection and reporting that includes actual spoken utterances, rather than data logs alone. A linguist can then accurately determine the source of the error — whether information might have been missing from the grammar, the name was mispronounced, or noise contributed to the problem. Timely corrections can then be made and transmitted back to the grammar, tuning the directory. Unfortunately, many SEAA solutions either leave error correction to the local administrator or provide a service where logs are reviewed on an infrequent monthly or annual basis. This scenario is like prepaying for a car tune-up service only to have mechanics change the oil, hoses, and filters months after their useful life expires -- and to have this service occur in long-duration intervals that only your mechanic can define.

Figure 3. SEAA Architecture with Advanced Disambiguation and Names Tuning



The Directory: Maintaining Speech-Recognition Accuracy

The third element of a SEAA architecture that is important to maintaining superior accuracy is the directory. SEAA applications usually compile your directory, or grammar, by combining your employee directory list with the vendor's most current pronunciations dictionary, a master list of known pronunciations. The problem with this approach is that the vendor's master dictionary is static and may not be updated by the vendor on a timely basis, sometimes only annually. This scenario is similar to purchasing a new car and never changing the oil. Your car will work fine for a few thousand miles, but after time the old oil will gum up your engine and reduce the performance of your car — eventually leading to engine failure. Without an updated grammar, the speech engine becomes almost useless over time. As employees join your organization, move locations, and add new contact numbers, the grammar must reflect these changes or you will not be able to reach them through an outdated SEAA.

The Cisco Solution: Three Unique Features Provide Exceptional Accuracy

Cisco researched the essential components for an accurate SEAA solution and determined that advanced disambiguation, names tuning, and a dynamic dictionary are all critical for accuracy and must be used together in the ideal SEAA solution. Cisco recommends a speech solution with a layer of intelligence on top of the speech engine in order to intelligently monitor and improve your speech solution for the benefit of your organization and your customers. Cisco has designed such a solution and offers it to our customers today.

The Cisco solution, Speech Connect for Cisco Unity, is the SEAA for the modern global enterprise. The Speech Connect solution delivers superior voice-recognition performance through the use of advanced disambiguation intelligence and a names-tuning service that combines automated software and a team of professional linguists to deliver timely error correction. Additionally, the solution adds another layer of intelligence to your SEAA solution by compiling your grammar

against two dictionaries — the basic dictionary provided with the speech engine and the Speech Connect superset dictionary that refines pronunciations. The Speech Connect dictionary is derived from all prior results of advanced disambiguation and names-tuning service usages. The grammar becomes more and more intelligent over time, resulting in improved accuracy for you even before advanced disambiguation and names-tuning features achieve their full capability.

Conclusion

Advanced disambiguation, names tuning, and a dynamic dictionary combined in a single solution deliver exceptional performance in speech. As many people like to tinker with cars and engines, sometimes building a car from a collection of parts, you can also build your own ideal SEAA solution. However, both tasks require much time, expense, and expertise. In addition, many SEAA solutions offer speech-recognition performance that is accurate more than 80 percent of the time. But like a car that goes without maintenance, it will soon lose its value. Speech Connect for Cisco Unity, however, begins with a high level of performance and continues to improve over time. Advanced disambiguation that learns your preferences, names tuning that makes error corrections in the background, and a superior pronunciation dictionary make Speech Connect for Cisco Unity the SEAA that meets your needs from the beginning and over time.

For More Information

Learn more about Speech Connect for Cisco Unity by reading the [Solution Overview](#), viewing the [Video Data Sheet](#), or finding details in the written [Data Sheet](#). All these items are available on the [Cisco Unity](#) page at Cisco.com.



Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

CCDE, CCENT, Cisco Eos, Cisco Lumin, Cisco StadiumVision, the Cisco logo, DCE, and Welcome to the Human Network are trademarks; Changing the Way We Work, Live, Play, and Learn is a service mark; and Access Registrar, Aironet, AsyncOS, Bringing the Meeting To You, Catalyst, CCDA, CCDP, CCIE, CCIIP, CCNA, CCNP, CCSP, CCVP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Collaboration Without Limitation, EtherFast, EtherSwitch, Event Center, Fast Step, Follow Me Browsing, FormShare, GigaDrive, HomeLink, Internet Quotient, IOS, iPhone, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, iQuick Study, IronPort, the IronPort logo, LightStream, Linksys, MediaTone, MeetingPlace, MGX, Networkers, Networking Academy, Network Registrar, PCNow, PIX, PowerPanels, ProConnect, ScriptShare, SenderBase, SMARtNet, Spectrum Expert, StackWise, The Fastest Way to Increase Your Internet Quotient, TransPath, WebEx, and the WebEx logo are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0804R)