Hey Siri, Tell Me About Advanced Speech Recognition

By: Al Balasco

Within a very short time, consumers have turned speed into a preferred way of interacting with their devices. Today, voice-enabled applications—such as Siri, Alexa, or Google Assistant—are increasingly woven into the fabric of everyday life. In the last year alone, the market for smart speakers has grown 128 percent—just in the U.S. This growth will increase, too: advanced speech recognition, text-to-speech, and speaker verification constitute a market that will be valued at $18.3 billion by 2023, according to Markets and Markets' *Speech and Voice Recognition Forecast Report.*

This dynamic market presents tantalizing opportunities for communications service providers (CSPs) to capitalize on such potential. And they're well-positioned: CSPs already have millions of subscribers—both consumer and business—using their networks every day for voice and video services. Devices with speech interfaces are now nearly ubiquitous, with *nearly* being the operative word. As *nearly* is eclipsed, CSPs can bridge the gap for speech enablement anywhere, anytime by offering speech recognition services in the context of a live voice or video call. Doing so will create the opportunity to offer value-added applications that generate new revenues in the process.

As the adoption of speech recognition capabilities to continues to grow at a rapid pace, however, key factors will prove critical to success. These include implementation costs, quality of experience, responsiveness, and streamlined user interfaces.

Market Overview for Speech Recognition

While speech recognition technology has actually been around for decades, recent advances have driven dramatic evolution (and the corresponding adoption) within the last five years. Machine learning has finely tuned recognition accuracy, and cost-effective deployment of mobile and other small form factor devices has driven adoption. As a result, speech is now the de facto—and, in some cases, the only—user interface for a number of devices. At scale, however, these solutions can be expensive, increasing the need for a new solution—or, more aptly, new solutions.

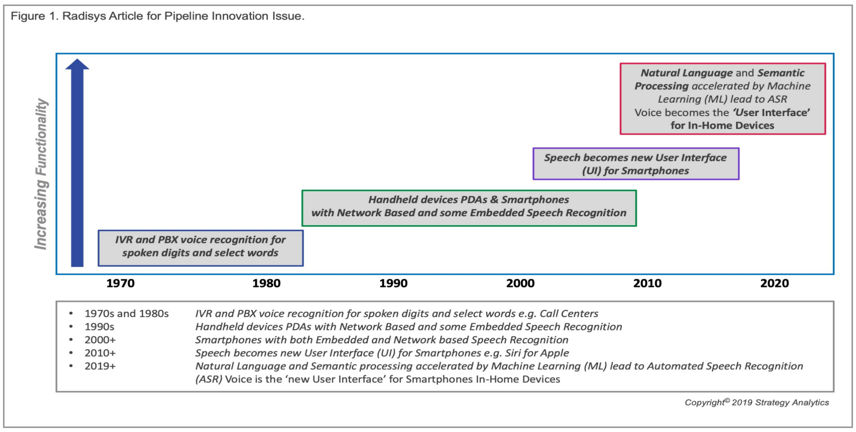

Figure 1. Radisys Article for Pipeline Innovation Issue.

Increasing Functionality
IVR and PBX voice recognition for spoken digits and select words
Handheld devices PDAs & Smartphones with Network Based and some Embedded Speech Recognition
Speech becomes new User Interface (UI) for Smartphones
Natural Language and Semantic Processing accelerated by Machine Learning (ML) lead to ASR Voice becomes the 'User Interface' for In-Home Devices
1970
1980
1990
2000
2010
2020

• 1970s and 1980s	*IVR and PBX voice recognition for spoken digits and select words e.g. Call Centers*
• 1990s	*Handheld devices PDAs with Network Based and some Embedded Speech Recognition*
• 2000+	*Smartphones with both Embedded and Network based Speech Recognition*
• 2010+	*Speech becomes new User Interface (UI) for Smartphones e.g. Siri for Apple*
• 2019+	*Natural Language and Semantic processing accelerated by Machine Learning (ML) lead to Automated Speech Recognition (ASR)* Voice is the 'new User Interface' for Smartphones In-Home Devices

Copyright© 2019 Strategy Analytics

Figure 1. History of Speech Recognition Technologies

According to Strategy Analytics, the hot markets today for advanced speech recognition include:

- In-home smart home assistants, such as Amazon Echo or Google Home Control
- On-the-road smartphone assistants, such as Siri on Apple devices and
- Vocabulary- and context-specific industry verticals, for example for medical surgery heads-up displays or for in-vehicle speech interfaces for hands-free navigation

Another growing market and major opportunity for service providers is “in-call” speech recognition. These in-call capabilities can support person-to-person interactions, person-to-machine interactions, person-to-bot interactions and more. They are able to serve the requests of callers who are already having a conversation or are on a conference call. During the call, callers can easily invoke the application to dial someone to join a phone call or to record the conference call. To add these capabilities to their networks, however, service providers need to overcome inherent challenges.

Advanced Speech Recognition: Challenges and Opportunities

Voice continues to be a key application for businesses and consumers, but balancing the cost of voice interactions, customer experience, and business productivity remains a key challenge for service providers. The industry faces two main challenges:

1. Costly and Complex Solutions Limit Service Innovation and Mass Market Outreach

The traditional approach to the deployment of speech recognition applications requires external ASR (Automatic Speech Recognition) servers—in-network or cloud—using high performance speech engines to process the entire gamut of speech interactions, from simple keyword or small vocabulary recognition to natural language and long-form transcription. The cost of using these external ASR servers to solve all interaction requirements, even for wake-word detection, has until recently made many applications prohibitive, limiting innovation and mass market penetration.

2. Performance and Latency Issues Impact User Experience

Sending all media files for processing by an external recognizer results in unnecessary latency, particularly if the speech inputs have to be transmitted over the Internet to public cloud servers. Although machines are getting smarter and can learn better, inaccurate keyword recognition, media processing, learning, or analysis and calls to action can initiate wrong processes that deliver either a poor user experience or cost a substantial amount of time and money for business-critical applications. Processing all speech media in the cloud also creates multiple privacy issues.

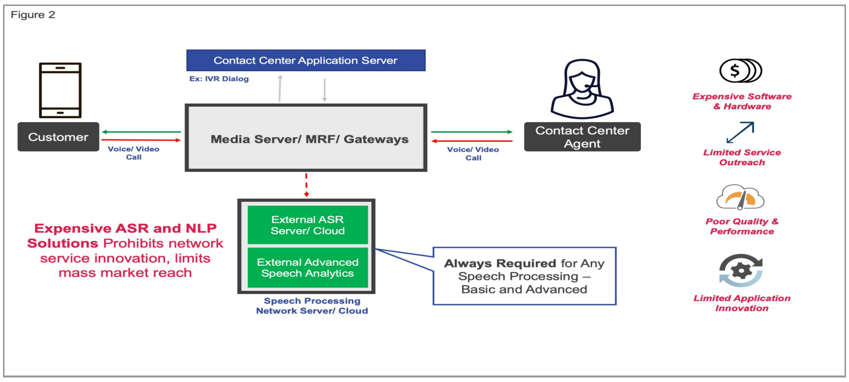
Figure 2
Contact Center Application Server
Ex: IVR Dialog
Customer
Voice/ Video Call
Media Server/ MRF/ Gateways
Voice/ Video Call
Contact Center Agent
Expensive Software & Hardware
Limited Service Outreach
Poor Quality & Performance
Limited Application Innovation
External ASR Server/ Cloud
External Advanced Speech Analytics
Speech Processing Network Server/ Cloud
Expensive ASR and NLP Solutions Prohibits network service innovation, limits mass market reach
Always Required for Any Speech Processing – Basic and Advanced

Figure 2. Challenges with Existing In-Call External Advanced Speech Recognition Solutions

To exploit new market opportunities and extend speech to many more applications, CSPs need to adopt new approaches to overcome these challenges. The traditional approach of leveraging natural language processing technology for everything may be too costly in terms of license and hardware requirements, and too complex for many interactions.

CSPs should consider a new approach that:

- Is device independent and agnostic, not reliant on a specific phone or smart speaker platform from a single manufacturer
- Offers lower total cost of ownership by cost effectively enabling multiple services
- Delivers fast processing
- Is able to process speech in long phone conversations without visual cues
- Overcomes call quality issues that are unique to each link of a telephony interaction
- Enables service providers to maintain their brand value
- Provides data for analytics to help CSPs better serve customers and enhance the value of services
- Allows dynamic selection of speech technologies to match specific use cases

The Solution: Embedded Advanced Speech Processing

Service providers can now embed advanced speech processing into software-based media server technology using platforms that are already deployed in their networks. This approach eliminates the need to purchase one-size-fits-all speech recognition engines for every scenario or to support costly integration with other elements in the network.

The software-based media server runs on off-the-shelf servers and exposes a full set of media server capabilities as well as speech recognition through a consistent set of open APIs. It decouples speech recognition processing from any specific device, thereby extending the application reach. By leveraging open APIs, service providers and application developers can simplify integration with any application server and support a variety of speech engines. Each speech engine has different strengths, which provides service providers with choice. By using a combination of wake-word and natural language technology, CapEx and OpEx can be reduced by up to 90 percent. Furthermore, by using an in-call wake word to determine the action required, the solution limits the use of expensive full-natural-language processing platform to an as-needed basis, thereby significantly reducing the overall cost.

The embedded solution with voice quality enhancements also overcomes the challenges associated with various call quality conditions that previously required additional network elements to improve the user experience.

Figure 3

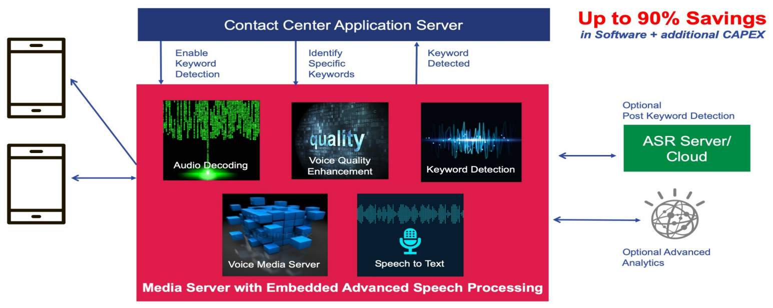
Contact Center Application Server
Up to 90% Savings
in Software + additional CAPEX
Enable Keyword Detection
Identify Specific Keywords
Keyword Detected
Audio Decoding
quality
Voice Quality Enhancement
Keyword Detection
Voice Media Server
Speech to Text
Media Server with Embedded Advanced Speech Processing
Optional Post Keyword Detection
ASR Server/ Cloud
Optional Advanced Analytics

Figure 3. Embedded advanced speech processing for in-call speech recognition capabilities

Real-World Use Cases for Advanced Speech Recognition

Communications service providers have the opportunity to support a wide range of new applications with in-call speech processing and media analytics embedded in their networks, thereby opening the ability to drive new revenues. During a phone call or conference call, the caller can invoke a wake word that launches the application. These applications include:

- Embedded “in-call” digital speech assistant (peer to peer, multi-party calls)
- Real-time speech transcription and speech analytics
- IoT-triggered conversational speech services and analytics

In the following examples, the wake word is “Hey Sophie.”

During a call with a contact center, John can interact with a voice-enabled bot by speaking to it with speech being processed by the media server. If a key word is detected such as “upgrade phone” or “technical issue,” the bot can send messages into the contact center and direct the call to the appropriate person. The agents themselves will have the wake word available to them, and they can say “Hey Sophie” to escalate the call by adding a manager or transferring a call to a tech expert, billing or other appropriate agent.

Figure 4

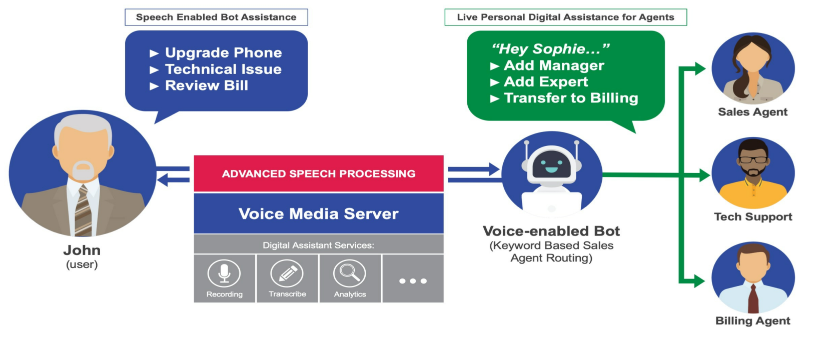

Figure 4. Call Center Example: Call Triggered Key Words and Speech Transcription

In an IoT-triggered example, the IoT device can detect a key word or sound and then be triggered to place an inbound call over a VoLTE network that enables its microphone or camera. Any audio in the surrounding area can be automatically transferred across the network in a live session. The voice media server in the network will receive the audio and be able to detect key phrases to, in turn, take specific actions. A typical use case could be a smoke detector that triggers the IoT device and then triggers a call to the fire department, enabling the firefighters to see and hear the live video stream.

Figure 5

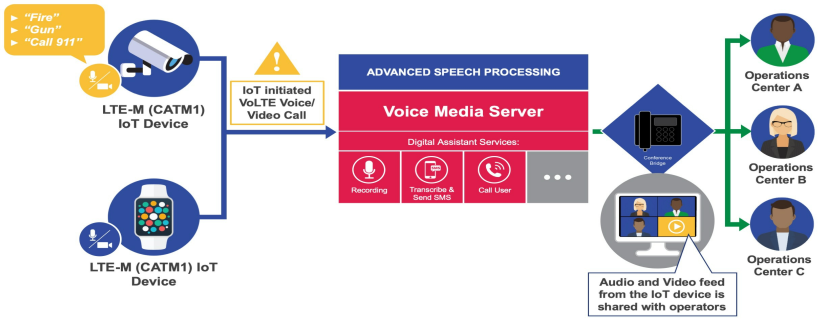

Figure 5. IoT Triggered Example: Emergency Response Speech Detection and Real Time Multi-party Assistance

Summary

As speech technology has evolved, an ever-expanding number of opportunities are emerging for CSPs to develop new network-based speech-enabled applications that will drive new revenues. By embedding advanced speech recognition capabilities into a media server that is already deployed in its network, CSPs can add significant value and extend applications to the mass market with in-call speech recognition services at significantly reduced CapEx and OpEx costs.

The opportunities for in-call speech recognition technology are endless and go far beyond the use cases described above. Integrated speech recognition solutions allow CSPs to offer their customers unique and extensible service capabilities that grow over time. As service providers work to reduce churn and grow ARPU for their enterprise and consumer subscribers, speech recognition applications and services can create a strong ROI and innovative edge for next generation value-based services.