



# SPEECH STRATEGY NEWS

SPEECH TECHNOLOGY IN BUSINESS AND COMMUNICATIONS

ISSN 1932-8214

March 2009

---

## VUI Visions

### The Evolution of Speech Technologies in Warehouse Voice Picking

*Doug Brown, VP, Product Management & Marketing, Datria*

*In this guest column, we ask designers skilled in creating Voice User Interfaces to highlight a particular aspect of VUI design inspired by actual deployments. In this issue, Doug Brown, vice president of product management & marketing at **Datria** discusses the breadth of speech technologies used in the past dozen years to voice-enable warehouse applications. The in-depth article presents an interesting evolution of the technologies used in a difficult problem that many don't realize has been one of speech recognition's earliest successes, both for vendors and the companies that buy it.*

*All three generations of technology offer a viable approach to today's supply chain automation, and Doug examines the reasons behind the diversity in approaches, pros/cons and recent trends, basing his article on discussions with colleagues as well as his own extensive experience. Doug became involved with speech recognition technologies at the Conversant Systems startup venture at AT&T in the mid-80s. His current work at Datria focuses on an ever-increasing set of packaged speech applications automating mobile employee processes (including warehouse workers). Datria was formed in 1997 as a spin-out of Lockheed Martin, and has been delivering multimodal data collection and field service solutions over the past 10 years. Significant customers include Johnson Controls, Bell Canada, Coca-Cola Enterprises, Energy South, TELUS, and Cardinal Health. Datria has partnerships with SAP, Cisco (e.g., SSN, May 2008, p. 27), and Nuance (among others).*

Over the last 10 years, the most widely known speech recognition application for many has been self-service in the contact center. Lately this has been changing, as speech recognition (voice user interfaces) has started becoming more prevalent in consumers' lives. People now talk to their navigation systems, to 411 information services, to their PCs, to the music systems in their cars, their cell phones (voice dialing and text-message creation), to corporate systems for password reset, for Google and Yahoo voice search, and so on. The unique hands-free, eyes-up attribute of a speech interface is becoming increasingly appreciated, especially as people are "always-connected" while mobile and away from a desktop interface.

Voice picking in the warehouse – once a niche market filled with specialty point solutions and proprietary technologies – is another example of companies using speech technology outside of the contact center. While not as well known as some speech applications, it is a burgeoning market with greater than \$400M end-user spend in 2008 (*"The Guide to Voice Solutions in Warehouse Environments,"* Daniel Hong, Datamonitor, February 2009).

An unusual characteristic of the voice picking market is the range of varying speech technologies at play – embedded capabilities versus network speech, speaker-dependent versus independent, adaptive technologies versus fixed voice templates, etc. This article looks at why different speech recognition approaches exist, and their relative pros and cons. It is important for the reader to know three things:

- A 2009 voice picking deployment can be successful using any of the approaches discussed below.
- With today's technologies, most misrecognition issues are rooted in user behaviors and are not caused by technological shortfalls.
- Regardless of vendor marketing, no speech recognition technology works 100% of the time for 100% of the users.

## Voice Picking

Voice picking is the generalized name for speech recognition applications that automate order fulfillment and other supply chain processes occurring in a warehouse. This catch-all includes a range of process automation addressing goods receipt and put-away, order selection (picking), replenishment, inventory management (cycle counting), storage location moves, cross-docking, returns, value-added services (such as engraving your name on an iPod before it ships), yard management, inspection, loading/packing, and shipping/delivery. Voice picking is easily complemented with other data collection technologies, such as scanning and RFID.

A voice picking transaction is relatively simple when compared to other speech-enabled enterprise mobility applications. Workers log into the system and begin receiving instructions on where to go in the warehouse to pick items for order fulfillment. When they get to the right location they speak a “check digit” from a signage, to confirm that they are indeed in the right place. The system then tells them how many items to pick, with the worker confirming quantities as they pick items. Voice picking systems are flexible and deal with deviations such as stock shorts and damaged goods. From a speech recognition perspective, there is a very small vocabulary (grammar) that is typically 60-200 words.

Companies are increasingly investing in voice picking for very tangible cost reductions, improved staff productivity, increased safety, faster employee ramp-up, regulatory compliance, reduced staff churn, and enhanced operations agility. Accurate supply chains also please end customers who value the right items arriving on time without damage and with the proper paperwork.

Vendors of Voice Picking Solutions  
(See article for terminology in columns)

Vendor	Speech Recognizer	Speaker-Dependent (SD)	Speaker-Independent (SI)	Adaptive
Cadre Technologies	embedded		commercial	manual
CTG	embedded		commercial	manual
Data Systems Int'l	embedded		commercial	manual
Datria	network-based		commercial	automatic
Genesta (SyVox)	embedded		commercial	manual
Itworks	embedded		commercial	manual
KBS Industrielektronik	embedded		commercial	manual
Lucas Systems	embedded	proprietary		automatic
Naurtech	embedded		commercial	manual
SAE Systems	embedded		commercial	manual
topsystem Systemhaus	embedded	proprietary	commercial	
Vanguard Voice Sys.	embedded		commercial	manual
Vocollect	embedded	proprietary		
Voice-Insight	embedded		commercial	manual
Voxware	embedded	proprietary		
Wavelink	embedded		commercial	manual
<b>Total --</b>		<b>4</b>	<b>13</b>	

## Pioneers (the 1990s)

The earliest challenge for successful use of speech recognition in warehouses was the level of noise – both steady state (conveyor systems, production lines) and dynamic spikes (forklift honks, pallets being abruptly dropped onto concrete, etc.). In the early 1990s, commercial speech recognition technologies were not robust enough to provide accurate performance in such noisy environments.

Pioneering vendors addressed this issue by simplifying the speech recognition effort. By taking a fat-client speaker-dependent (SD) approach, spoken input (utterances) would only need to be matched against one-person's established voice templates. This simplified the computing requirements and narrowed the

possible outcomes. The small footprint of the recognition software and compact computing requirements were an excellent fit to the more rudimentary mobile devices available in the 1990s.

Given the lack of commercial market alternatives in the 1990s, voice-picking vendors were forced to become speech R&D operations, creating proprietary speech engines. While market options would change over the years and commercial alternatives would emerge, these proprietary speech engines can still be found in today's markets, and are still valid alternatives for customers to consider.

In the 1990s, the strength of using SD technology was that quality speech recognition became possible in noisy warehouse environments, creating the voice picking market. SD algorithms were also an excellent fit to limited CPU and memory resources of the proprietary rugged mobile devices of that era, units that preceded the commercial handhelds available today.

Yet SD technology was not a panacea. Recognition against a user's voice templates was very specific, which is to say narrow. If users spoke differently as they tired later in their shifts or had colds, they would not sound like their original recorded voice templates and recognition problems could exist. One vendor, MCL Technologies, created a tool to identify SD users having recognition problems (MCL-Voice Manager). This tool helps solution administrators identify workers having recognition performance difficulties, and allows taking corrective actions. Misrecognition issues were often behavioral (such as yelling instead of speaking in the same voice in which the templates were recorded), but when problems were technology-based, the user had to re-record their voice templates. A recent customer case study published about Carib Sales noted that all their workers had to re-record their voice templates. The original voice samples were said to be "robotic" whereas actual field use had a more relaxed voice sound. This highlights the need in the SD approach for workers to speak in a manner consistent with their stored voice templates.

According to industry pundit Judith Markowitz (J. Markowitz, Consultants) Lucas Systems became the first of the pioneering SD vendors to use an adaptive speech model approach to avoid re-recording of voice templates. Markowitz notes that, "An adaptive approach modifies the user's voice template over time, adding new acoustic model information through actual use in the warehouse." Adaptive speech recognition, as discussed below, is an excellent technique for improving speech recognition performance in an automated, unsupervised manner.

Larger customers perceive one aspect of SD technology to have hidden costs: the creation of voice templates for each worker. The requirement to have a worker spend 30-40 minutes to record their voice template initially sounds benign. Yet customers with larger user populations and multiple sites have become sensitive to the management resources involved, the lost hours of productivity, and the lengthening of rollouts that delay reaping cost reductions.

For example, Mike Jacks, Senior Manager of Logistics and Transportation Systems at Coca-Cola Enterprises (CCE), recently noted the difference in using SD and SI technology approaches in their multi-location roll-out to more than 2,300 pickers at 100 sites. If SD technology were used, CCE would have had a warehouse supervisor or IT manager involved with each worker's creation of voice templates. Assuming 40 minutes for each picker and 40 minutes for a supervisor, the total investment would have equated to 383 days (more than a year) of lost staff time. By using SI technology instead, the solution worked out of the box for all workers and no time was lost to creating voice templates.

More importantly, deploying voice picking to 100 sites with SD would have taken far longer than the 16 months it took with SI technology. This is a critical point as a longer deployment means delays in CCE realizing the compelling cost savings. In comparison, a recent press release on Morrison's, a sizable UK grocery chain making a self-described "ruthless deployment" of SD voice picking (requiring each user to create voice templates) says that it will take two years to deploy only 30 sites. This is one reason why larger customers have become sensitive to the economic impact of choosing an SD technological approach to voice picking.

One other limitation of SD technologies – and again, smaller customers may be as less sensitive to this – is that SD is only a good fit to warehousing applications. The limitation to small vocabularies (up to 200 words due to the requirement for each user to record voice templates) means SD may not be suitable for cost savings applications elsewhere in the company. Optimizing other enterprise processes typically requires the ability to recognize vocabularies (grammars) with hundreds or thousands of words: a requirement best met with speaker-independent (SI) technology. Examples of optimizing processes with voice applications beyond the warehouse include: plant maintenance, enterprise asset management, field service management, transportation management, sales force automation, regulatory audit/inspection/reporting, crisis management, and human capital management.

Examples of early pioneering vendors taking the speaker-dependent (SD) approach are Lucas Systems, Topssystem Systemhaus, Vocollect, and Voxware (via the 1999 acquisition of Verbex Voice Systems). All still offer and deploy SD solutions today, with Topssystem becoming the first vendor to offer a choice of SD and speaker-independent (SI) technology in 2007. In the past year, Topssystem (also known as Top-VOX) said at the recent ProMat '09 event that almost all of its new customers have chosen their SI engine, but that it's nice to have the SD technology on hand in case there is a rare speaker better suited to that technology.

Another pioneering vendor, SyVox (originally Speech Systems Incorporated) took a different tack, attempting to move acoustic inputs from a device client application to a centralized server, where speaker-independent (SI) processing would be used to understand spoken inputs. This was a highly innovative approach – possibly years ahead of its time – and parallels can be seen with today's 3<sup>rd</sup>-generation VoIP approaches discussed below. SyVox also became one of the first vendors to switch from proprietary to commercial speech engines in 1998. Ultimately, the SyVox brand was acquired by Genesta where it is sold today as a second-generation, commercial speaker-independent (SI) solution.

### **Exploiters – 2<sup>nd</sup> Generation (2002-2007)**

Investment in the commercial speech recognition market increased, especially in the area of small footprint embedded automatic speech recognition (ASR) engines. New applications drove significant R&D and new mobility applications – catapulting speech recognition as the interface-of-choice for speaking to portable navigation systems, talking to your car's music system, dictating text messages, obtaining 411 information, dialing by name and most recently, doing multimodal Google and Yahoo searches. While unrelated to the voice picking market, many of these mobility applications faced similar challenges – how to accurately interpret spoken input in a noisy environment.

This resulted in a range of commercial speech recognition vendors – Nuance, IBM, Microsoft, Loquendo, SVOX and others – producing commercial off-the-shelf (COTS) products. These new engines made it far easier for application companies to introduce voice recognition solutions. Academia also made speech engines available via open source licensing. The impact of new speech engines was quickly apparent on the voice picking market, where a new software-only business model appeared and the number of competitors quadrupled.

In the 1990s, due to immature speech products, voice picking vendors had to invest R&D into proprietary engines (discussed in “Pioneers,” above). Once commercially viable engines became available, it was possible for software application-only vendors to enter the market. And many did, notably Cadre Technologies, CTG, Data Systems International, Itworks, KBS Industrielektronik, Naurtech, SAE Systems, Vanguard Voice Systems, Voice-Insight, and Wavelink.

Unlike their predecessors, these new companies exploited the advancements made in commercial speech engines designed to be embedded in mobile devices. In parallel, they also took advantage of the COTS mobility devices: ruggedized small form-factor computers from companies such as Intermecc, LXE, and Motorola (Symbol). While standardized mobile devices were still constrained in processing, memory and power, they were still highly capable of running more modern speech recognition algorithms.

Daniel Hong, lead analyst for Customer Interaction Technologies at Datamonitor, said, “These new commercial speech recognition offers bring a consistent attribute to the voice picking market: speaker-independent (SI) recognition technology, where the product works out-of-the-box without users having to train it to their specific voice. This eliminates many hours of lost staff time and speeds the time to ROI.” All voice picking market entrants since 2002 have chosen the SI approach. In fact, no new SD vendor has entered the voice picking market with SD in the past eight years.

Speaker-independent (SI) technology is based on a broad set of acoustic models, allowing it to be highly insensitive to the manner in which a person speaks. Yet, like all speech recognition technologies, it may not work for all speakers. As a result it is common to find adaptive speech approaches in SI technologies. At times the adaptation is manual in embedded technologies, where acoustic model sets can be appended from actual users.

Some of the pioneering vendors reacted to these new market dynamics. SyVox (acquired by Genesta in 2003) abandoned proprietary speech recognition and moved to commercial SI engines. Topssystem added an SI engine to its product line. Voxware moved away from being a hardware manufacturer to a software business model.

## SOA and the Emergence of Thin Clients – 3<sup>rd</sup> Generation (2007+)

In the past decade and a half, many companies have embraced speech as a powerful Customer Service technology for use in their contact centers (as readers of this newsletter know). Unlike the fat-client recognition technologies discussed above in the first two generations of voice picking, contact center solutions have always deployed speech as a shared network resource. This topology only required users to place a phone call to access the speech recognition technologies. With the advent of Voice over IP support via WiFi networks (sometimes known as “VoFi”), a network-centric approach has been enabled for voice picking and other “inside the four walls” enterprise applications.

The new thin-client approach moves the speech recognition effort from each worker’s mobile device to a centralized resource (server). This immediately reduces the processing and memory requirements on the mobile device; thus opening the door to support inexpensive devices, such as wireless IP telephones and smartphones. Handheld computers and PDAs can also be supported for multimodal transactions, as long as they support softphone capabilities (as many do).

Server-based speech recognition allows companies to tap into the most robust, open, and mature speech recognition technologies in today’s markets – those that they have already been leveraging in their Customer Service contact centers. These speech recognition technologies are proven in today’s market, handling billions of calls per day. They also excel at handling speech in noisy environments due to advancements made in the past 10 years to accommodate calls made over wireless networks and from noisy mobile environments.

Network speech solutions are speaker-independent (SI), bringing the “works right out of the box” and speedy deployment benefits discussed above. It is also common for network SI technology to be automatically adaptive, accepting new acoustic information into its database for improved performance based upon actual field use. Network speech recognition also supports the largest vocabularies (grammars) that are capable of understanding and acting on thousands of spoken words. This enables companies to use speech recognition resources beyond the warehouse.

W3C standards are also well-evolved for network-based SI technology, including a standard specification (MRCP) providing for the “plug-and-play” of speech recognition technologies from different vendors. This simplifies customer purchasing, as this new flexibility provides future-proofing, thereby eliminating being locked into a single speech recognition supplier. MRCP-compliant speech engines are available from companies making deep investments in speech technology, including Nuance, IBM, Microsoft, AT&T, and others.

“During any economic conditions, and particularly with the current global recession and credit crunch, Manufacturing and Supply Chain customers are intent to get more productivity and lifecycle services from their existing assets,” says Chet Namboodri, Global Director for Manufacturing Industry Solutions at Cisco. “Thin-client solutions like warehouse voice picking that utilize existing wireless and converged IP infrastructure fall into this category. With incremental ROI paybacks at less than 12 months, customers can leverage their enterprise investment throughout the business with voice automation for any manual work flow, instead of proprietary point solutions.”

The only vendor currently offering thin-client voice picking speech solutions is Datria, which brought it to market in 2007 in one of the industries largest voice picking deployments (2,325 concurrent users). That deployment enabled a large beverage company to use its Cisco wireless phones and speech recognition on Cisco routers to deploy 100 sites in a little over a year.

## Technology Enablers

It would be disingenuous to attribute all technological advancements to the speech recognition software. That is far from the truth. All three speech recognition generations have benefited from dramatic improvements in noise-cancelling circuitry and directional microphones in commercially available headsets and mobile devices. Many handhelds are optimized for high quality audio transmission as well.

WiFi networks (802.11 a/b/g/n) have also evolved tremendously, providing stable and uninterrupted coverage inside warehouses and standardized levels of quality of service (QoS). VoIP support is now common on WiFi, easily engineered for roaming, shift-long connection times and 10+ hours on standard batteries.

Affordability has also been enhanced, as customers can now assemble voice picking solutions using commercial off-the-shelf mobile devices, headset, WiFi, and speech technologies.

## Languages

At times, vendor marketing will focus on languages and dialects as another differentiator of SD and SI speech recognition technologies. This is a discussion that can mislead buyers. Multilingual capabilities imply two-way communications, both the ability to speak a language as well as hear/understand it. In most voice picking solutions – where text-to-speech synthesis can be a requirement to give voice to large catalogs of SKUs – it is the speaking to the worker that is the limiting factor. Most vendors use commercial text-to-speech (TTS) engines to synthesize speech in different languages, and those engines typically offer 10-25 dialect choices for embedded solutions and up to 35 for network-based solutions. This is far less than SI recognition (up to 58 dialects) or SD recognition (unlimited dialects). Notably, the number of language or dialect packs available tends to exceed the number of languages that supervisors and pickers actually speak in warehouses.

Language skills are exasperated in multimodal situations, where the ability to read a language (e.g., what is on a display) further complicates multilingual employee management.

One final anecdote about languages. Coca-Cola Enterprises deployed Spanish language support at some US locations due to the mix of employees speaking Spanish as their primary language. Unexpectedly, these employees chose to work in English, to build their second-language skills. To them, the voice picking solution acted as a language tutorial and provided a path to broadening their skill value to their employer.

## Misrecognition

While this article has spent a considerable effort distinguishing pros and cons to varying speech-enabled warehousing technologies, it would be naïve to ignore that most speech recognition errors occur due to behavioral issues. Recognition issues commonly arise from users neglecting to:

- Wear the headset microphone properly;
- Speak in the expected manner (e.g., yelling to overcome background noise instead of speaking in a normal voice);
- Speak the expected words (being at a different prompt where a different set of words is expected); or
- Use the right equipment (e.g., a headset with an omni-directional microphone).

It is incumbent on the voice picking vendor and its customer to jointly provide the necessary end-user training to avoid behaviors that result in misrecognition errors.

## Summary – Standards transform markets

The voice picking market is following a natural market maturation curve familiar to many in IT. Early pioneers delivered vertically integrated solutions to establish that the concept worked and the business case was valid. Second-generation solutions moved towards openness, increasing hardware choices while reducing capital expenses and Total Cost of Ownership. The emerging third-generation of voice picking solutions completes the migration to 100% commercial-off-the-shelf components, W3C standards compliance, and unsurpassed choice, flexibility, and affordability. In addition, corporations can now choose voice picking solutions that create an enterprise-wide speech automation resource within their SOA architecture.

Today's diversity of speech recognition approaches is a boon to companies considering speech-enabled solutions in their supply chain.

---

**Copyright TMA Associates 2009; All rights reserved.**

**Mail or fax orders to: TMA Associates, P.O. Box 570308, Tarzana, CA 91357-0308 USA. Tel: (818) 708-0962. Fax: (818) 345-2980, or register on web site: <http://www.tmaa.com/sru/subscribe.htm>.**

Speech Strategy News is published twelve times per year by TMA Associates, Editor: William S. Meisel. Trademarks mentioned in this publication are the property of the companies mentioned; they are used editorially. The material herein is based on data from sources believed to be reliable, but is not guaranteed as to accuracy and does not purport to be complete. From time to time, the author or TMA Associates may have consulting assignments, advisory positions, own stock or have other business relationships with organizations in speech recognition and associated areas, including companies discussed in this newsletter. Speech Strategy News is a trademark of TMA Associates.