WHITE PAPER

EVOLUTION OF SPEECH RECOGNITION TECHNOLOGY IN THE WAREHOUSE

BY STEVE YURICK, LUCAS SYSTEMS, INC.

Voice-directed warehouse applications are a proven solution for improving distribution efficiency in industries ranging from grocery and foodservice distribution to apparel and industrial supply. A typical voice application combines a voice-directed workflow—the system provides audio prompts directing users what to do—with speech recognition technology that understands a user's spoken responses. Voice applications seamlessly integrate with other warehouse systems to enable hand-and eyes-free operations that drive new levels of associate productivity and accuracy across picking and other warehouse tasks.

Underlying every voice application is a sophisticated speech recognition platform designed to ensure users are understood quickly and accurately, every time they speak. High recognition accuracy means users don't have to repeat their responses, which maximizes efficiency and user acceptance of voice as a tool that helps them do their jobs better.

Over the past decade the underlying speech recognition technology used in warehouse voice applications has undergone a significant evolution due in large part to developments in consumer and non-industrial markets, everything from GPS and smartphones to medical transcription and call centers. Although warehouse voice applications have different needs than these other applications, the advances in the wider speech recognition space are having a direct impact on advances in the warehouse market. But rather than adopting consumer speech recognition technologies in whole, the ideal recognition solution for the DC incorporates a combination of technologies that provide the best-possible speech recognition accuracy and ease of use in a noisy environment with a widely diverse user population.

This paper explores the evolution of speech recognition concepts and technologies, including the pros and cons of alternative technology approaches for a warehouse environment. The goal is to give non-experts an overview of the state-of-the-art so they can better understand the differences among their technology choices. We also explain why recognition technology that may work well in a consumer application—or in a conference-room product demonstration—may not be adequate for the warehouse.

The final section of the paper describes how Lucas Systems has incorporated the best available speech technologies in the Serenade speech platform that is a core component of our Jennifer™ VoicePlus system. Serenade provides superior recognition accuracy in a warehouse, reduced training time, and greater user acceptance than any previous warehouse speech technology. In fact, warehouses that have upgraded from legacy recognition technologies to Serenade have seen measurable improvements in recognition accuracy, along with the additional benefits of reduced user training.

I. Speech Recognition Challenges in the DC

Before describing the speech technology landscape, it's useful to describe the requirements for effective speech recognition in a warehouse or distribution center. Most importantly, warehouse applications require far better recognition accuracy than a call center or other consumer application. Acceptable recognition for a game-player would not be acceptable for a voice picker. If a warehouse worker using a voice picking system has to repeat him- or herself, or if the speech recognizer mis-recognizes commands, the user's productivity suffers and acceptance and adoption of the voice application is jeopardized.

Every five percent difference in recognition accuracy—99% accuracy versus 94%, for example—translates into 2-6 minutes of productivity benefit per user per day (depending on pick rates and other factors). While the individual time benefit of better accuracy is small, the cumulative benefit for operations with large numbers of users is significant. On the other side, the effect of superior recognition accuracy on individual user satisfaction is large but far harder to quantify or monetize.

Two additional factors compound the recognition challenge in the warehouse: non-standard accents and variable background noise profiles. In any country, pronunciations may differ widely from region to region, and individual user accents, speech patterns and speech impediments add another level of complexity. Unlike a call-center application that may eventually default to a live person if a user cannot be understood by the recognizer, a warehouse system has to work for all users all the time—there is no fall-back.

Similarly, background noise in any given warehouse can be extreme due to blowers and fans, conveyors, fork-lifts, and other factors. What's more, the level of background noise can vary from area to area and may also change quickly (for example, when a pallet is dropped or conveyors are turned on). Any recognition technology used in a warehouse has to provide outstanding recognition accuracy across diverse user speech patterns amid loud and highly-variable background noise patterns.

Two additional factors compound the recognition challenge in the warehouse: non-standard accents and variable background noise profiles.

Background noise from fans and conveyors add to the recognition challenge in a DC.

II. Core Speech Recognition Technology—Phonetic- and Word-Based Engines

Underlying every speech recognition system are mathematical algorithms (typically referred to as "engines") that translate what a user said into data by matching the characteristics of digitized audio against a pre-defined model. There are two basic methods for doing this: word-based recognition, where the voice database is based on whole words, and phonetic-based recognition, where the voice database is based on phonemes, the sound components that make up words.

99% vs. 94%

Translates into 2-6 minutes of productivity benefit per user per day.

(depending on pick rates and other factors)

Many early warehouse voice systems utilized the word-based recognition approach. At the time, speech recognition technology was still developing and the techniques for the more rigorous statistical modeling framework required for phonetic-based recognition had not yet matured. But over the past decade the bulk of speech recognition R&D has been focused on phonetic-based systems. As a result, phonetic-based engines have matured to the point where they are suitable for industrial applications. (For more information about the history of speech recognition in general, see Automatic Speech Recognition: A Brief History of the Technology Development, by B.H. Juang and Lawrence R. Rabiner.)

The word-based method has some distinct advantages and disadvantages. A major disadvantage is that it requires every word that is to be recognized to be "trained" by each user before they can use the system. Since a typical warehouse application consists of 100-200 words that the user can speak, this can be time-consuming. Training times with typical word-based recognizers range from 20-40 minutes per user. Additionally, it is important in a word-based system for users to speak their words in a consistent manner while they work as there is generally less room for pronunciation variations than with the phonetic-based method.

On the plus side, the word-based method handles heavy accents, non-standard pronunciations, and speech impediments especially well because there is no pre-built dictionary in the system that defines one or more pronunciations for each word. In a word-based system, the pronunciations are exclusively defined by how the user pronounces the word during training. In fact, since a user is in complete control of the pronunciations he records, he or she may elect to choose a completely different pronunciation of a word. Like the word-based method, the phonetic-based method has pros and cons. By mathematically modeling sub-word sounds, one needs to train only those distinct sounds in a language (usually around 45) in order to recognize large vocabularies of words and phrases. Because of this, the phonetic-based method is used for systems that require reduced or no voice training time (such as consumer-facing call center systems) or applications that require large vocabularies. In addition, the phonetic-based method is also strong at allowing for continuous speech recognition, which allows users to talk naturally without the need to pause between words or commands.

It's worth noting that phonetic recognizers may still require user training to achieve acceptable recognition rates for voice picking or other warehouse applications. In fact, first-generation phonetic recognizers used in the warehouse often required as much user training as word-based systems. On the other hand, because of the phonetic-based model underlying the recognition approach, any adaptations in an individual user model—to improve accuracy in recognizing a single word—may degrade recognition of other words. Similarly, a phonetic recognizer does not give users the same degree of flexibility to completely alter how a given word is pronounced and recognized, which is necessary for individuals with non-standard pronunciations.

Going a step further, what if you could do away with training altogether so that anyone could use the voice system without creating a voice model? That's the intent of speaker-independent systems used in automated customer service applications and other consumer-oriented products.

First-generation speech recognition engines were typically integrated with a voice hardware appliance, such as these devices.

III. Recognition Approaches—Speaker-Dependent and Speaker-Independent

Until recently, warehouse voice systems all used speaker-dependent technology in which the voice engine is trained to recognize each user's speech patterns. Many of these first-generation speech recognizers were originally built, tuned and optimized for a specific voice-only hardware platform in a closely-coupled hardware/software system. Although these proprietary recognizers may now be used on general-purpose hardware devices, they may not perform as well as they did on the special-purpose hardware for which they were originally designed.

In addition to the inherent limitations of earlier voice-only hardware (which is the subject of a separate paper on mobile computing platforms for warehouse voice applications), these first-generation warehouse voice recognition systems typically required a 20-40 minute voice training process for each user, as described above. This time investment in user training was seen as a small price to pay to get high recognition accuracy.

Another drawback of many first generation speaker dependent systems was that they often required users to record a second voice template after starting to work with the system. Why? Because people usually speak very clearly and deliberately when they initially create their voice template, but as they get comfortable working with voice, they speed up and revert to their usual, natural speech patterns—words become combined, word endings are omitted, etc. When that happens, the recognizer starts having problems matching what users say against the templates they built, so the users have to perform a second 20-40 minute training process.

To eliminate the need to retrain, five years ago Lucas Systems introduced the concept of adaptive voice modeling, in which the speech software automatically adapts the user's voice template in the course of use—the recognizer is continually "training" as the user works. So as users start working faster, slurring words together and cutting off the ends of words, the recognizer keeps up. More importantly, rather than degrading, recognition accuracy actually improves with use, even when a user's pronunciation changes slightly due to fatigue (at the end of a shift, for example), or a head cold. The ability to provide consistently high recognition rates despite changes in the user's voice has a significant benefit in long-term user satisfaction.

Going a step further, what if you could do away with training altogether so that anyone could use the voice system without creating a voice model? That's the intent of speaker-independent systems used in automated customer service applications and other consumer-oriented products.

Another drawback of many first generation speaker dependent systems was that they often required users to record a second voice template after starting to work with the system.

Speaker independent systems have dominated mass user applications for the simple reason that you couldn't realistically ask every person who calls in to a voice-directed customer service phone line to train the system. While previous speaker-independent systems (all of which are based on phonetic recognition technology) provided acceptable recognition accuracy for consumer-facing applications (and those applications always had a fall-back to a live operator), they did not provide high enough accuracy rates required for warehouse applications, especially noisy environments. In a warehouse application, poor recognition degrades user productivity and, worse yet, frustrates users and impacts technology acceptance.

IV. Audio/Noise Adaptation

The third major component of a warehouse speech recognition platform is audio pre-processing to address background noise in the audio signal that is provided to the computer-based recognizer. First generation warehouse voice systems typically relied solely on 'noise-reducing' microphones. These microphones—which have improved significantly over the years and are still required for warehouse applications—include a dual microphone in the boom, one facing towards the user and one facing away from the speaker which captures and filters background noise.

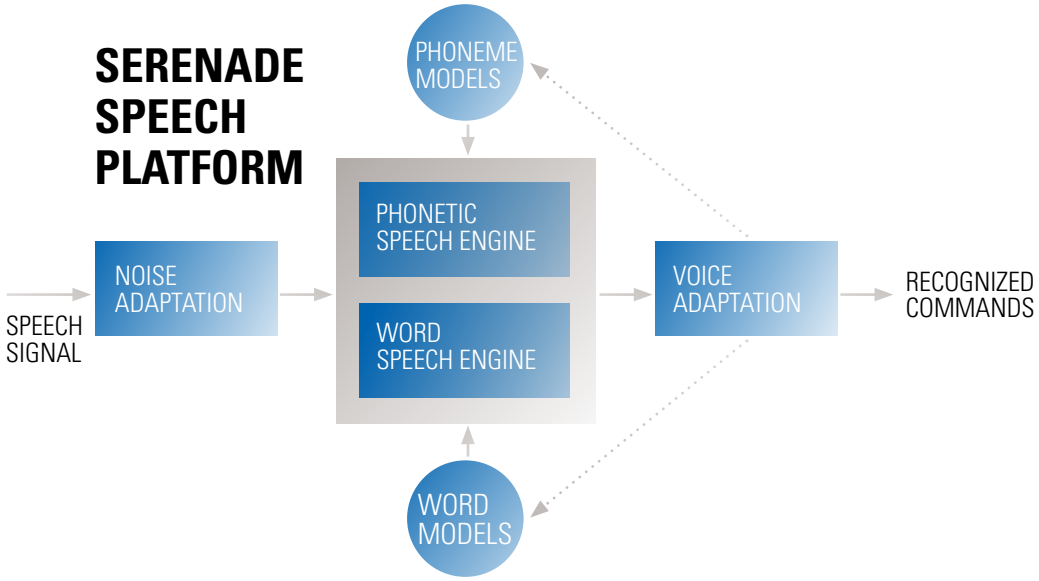
Noise-cancelling microphones filter out background noise at the source.

Later warehouse voice applications added a noise sampling function on the mobile computer. For noise sampling, the recognizer would take a sample of current background noise (while the user does not speak) in order to set appropriate noise-cancellation parameters and provide a higher quality audio sample for the speech recognition engine. Since this is not an automatic process, users might perform noise sampling several times in the course of a shift if they have recognition issues as the background noise levels in the warehouse change.

The final piece of the puzzle is to integrate the noise reduction approach with the speech recognition algorithms.

A better solution is to include audio processing technology that can automatically and continuously monitor and adjust to changing background noise levels. To do this efficiently requires far more computer processing power than was available in previous generation voice terminals. In addition, the algorithms for processing audio input have advanced dramatically, similar to the advances in speech recognition engines. Taken together, we now have powerful mobile computers that can efficiently support higher audio sampling rates to enable advanced audio pre-processing and sound adaptation. The final piece of the puzzle is to integrate the noise reduction approach with the speech recognition algorithms. With these techniques and technologies, the speech platform can effectively handle all types of warehouse noises, from steady state noises such as conveyors to noise spikes such as pallets being dropped onto concrete. (*For a more complete discussion of the approaches to audio pre-processing for speech recognition, see Multiple Approaches To Robust Speech Recognition, by Richard M. Stern, et al.*)

Serenade combines the power of phonetic-based recognition with the flexibility of the word-based recognition.



V. The Best of All Worlds For the Warehouse: Serenade Speech Platform

As described above, the warehouse environment presents unique complexities and requirements. The challenge for software companies delivering voice-directed warehouse applications is that most of the advances in basic speech recognition technology have been targeted to other markets and therefore do not address every aspect of what is needed in the DC.

Lucas Systems brings unique expertise and experience to this challenge. For the past 14 years we have delivered speech-based applications in the warehouse—Jennifer VoicePlus—using the broadest possible range of recognition technology. Early in our history we delivered production systems using each of the leading speech engines that were purpose-built for the warehouse. More recently, we have developed a speech platform—Serenade—that is designed to support warehouse-specific audio capabilities independent of the underlying speech engine.

Serenade's engine-independent design allows us to incorporate a variety of speech engines, so our customers can readily get the advantages of advancing speech recognition technology. In addition, Serenade incorporates continuous noise adaptation, multiple simultaneous recognition approaches, and adaptive voice modeling. This unique bundle of technologies provides the best of all worlds – optimal accuracy in noisy environments, maximum flexibility to adapt to different languages and atypical speech patterns, and minimal speech training.

Dual Speech Engine Technology

Serenade combines the power of phonetic-based recognition with the flexibility of word-based recognition. As described above, today's advanced phonetic based recognizers provide the best-possible recognition accuracy across the broadest spectrum of users without the need for individual speech training. Phoneme-based recognition also allows users to talk naturally without pausing between words or commands. On the other hand, word-based recognition offers the advantage of selective word-based retraining for out-of-norm speakers.

High Resolution Noise Adaptation

Serenade includes advanced high resolution audio processing that continuously adapts to changes in background noise levels. Leveraging the processing power of today's mobile computing platforms, Serenade's noise adaptation technology provides higher audio sampling rates designed to optimize the audio signal used in the recognition engine.

Minimal (or Zero) Training And Flexible Adaptation

Serenade supports a no-training option similar to a completely speaker-independent system. Unlike a true speaker independent system that does not have the capability to learn a person's unique speech patterns, Serenade's adaptive modeling is enabled for all users, whether or not they go through an initial training process. Most distribution centers elect to use a recommended five minute user enrollment process that provides the system with initial voice samples used to create a voice model or template for the user. This enrollment option allows the

system to get a jumpstart on learning a user's voice while at the same time giving the user a gentle introduction to using the voice system – including the headset and mobile computer.

Even with an initial enrollment process of about five minutes, DCs using Jennifer with Serenade save 20 minutes or more in initial training per user, get high accuracy from day one across all users (even challenging users), and don't have to re-train after getting comfortable with the system. For a DC with just 24 users, every 20 minute time savings equals one eight-hour work day. More importantly, with industry-best recognition rates, users get confidence in the system from day one and can concentrate on their jobs rather than the technology they are using to do it. At the end of the day, that's a critical requirement for any voice system.

VI. Final Thoughts

Like other enabling technologies, the use of speech recognition is highly dependent on the application environment in which it is used. Therefore, it's not surprising that the recent advances in phonetic-based recognition engines and speaker-independent recognition in consumer markets have not proven to be the "holy grail" for recognition methodologies in the warehouse. Because of the unique needs of the warehouse, the new methods, while bringing certain advantages, haven't completely superseded the previous methods that have dominated in the DC. For now, at least, the hybrid approach to speech recognition that is embodied in Serenade provides best-possible recognition accuracy, reduced training time, and greater user acceptance and flexibility than any previous warehouse speech recognition platform.

Voice-directed applications enable hands and eyes-free warehouse operations, driving new levels of associate productivity and accuracy across picking and other tasks. Lucas Systems takes a unique process-centric approach to voice system design that is combined with Jennifer VoicePlus, the most flexible, configurable software in the industry.

Tens of thousands of users at companies like Cardinal Health, C&S Wholesale Grocers, CVS/pharmacy, Do it Best Corp., Kraft Nabisco, and OfficeMax use Jennifer voice applications every day. For more information, visit www.lucasware.com.

About Jennifer VoicePlus and Lucas Systems, Inc.

Voice directed applications enable hands and eyes-free warehouse operations, driving new levels of associate productivity and accuracy across picking and other tasks. Lucas Systems takes a unique process-centric approach to voice system design that is combined with Jennifer VoicePlus, the most flexible, configurable software in the industry. The result is better, more comprehensive voice applications that are tailored to your operations and accelerate your return on investment.

Since 1998, Lucas Systems has been the leading innovator of voice-directed applications for the warehouse. We have consistently pushed the limits for what is possible with voice, developing creative applications that drive better results for our customers. We were the first company to provide end-to-end voice applications from receiving through shipping, and we delivered the industry's first voice picking systems on standard hardware terminals supporting voice and scanning, rather than special-purpose, voice-only hardware. Today we have more customers running voice applications on standard industrial mobile computers than all other voice vendors combined, and we are the fastest-growing voice provider in the industry.

Beyond our passion for process, Lucas has a unique track-record of customer success that translates into an intensely loyal, committed customer base. We work with single-site DCs and large multi-national distributors and retailers across a wide range of industries. Tens of thousands of users at companies like Cardinal Health, C&S Wholesale Grocers, CVS/pharmacy, Do it Best Corp., Kraft Nabisco, and OfficeMax use Jennifer voice applications every day. For more information, visit www.lucasware.com.

References

B.H. Juang and Lawrence R. Rabiner, Automatic Speech Recognition – A Brief History of the Technology Development, October 8, 2004 (available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.5614&rep=rep1&type=pdf)

Richard M. Stern, Fu-Hua Liu, Yoshiaki Ahshima, Thomas M. Sullivan, and Alejandro Acero, Multiple Approaches To Robust Speech Recognition. (available at: http://research.microsoft.com/pubs/78428/darpa92.pdf)

LUCAS SYSTEMS, INC.