THE TALKING WAREHOUSE

An eBook by Roger Byford and David Maloney

Understanding How Voice Unleashes Higher Performance in Product Distribution

Chapter 2: Demystifying Voice in the Warehouse



In Chapter 1, we provided a brief historical perspective on how voice technology in distribution circles came to be, and the early experiences of Vocollect as told through the voices of its cofounders and one early adopter (Kroger).

We see Chapter 2, "Demystifying Voice in the Warehouse," as a technical springboard to all the future chapters of *The Talking Warehouse*. In this chapter, we'll look at many types of voice recognition systems and explore why they are not created equally. We'll give special emphasis to voice recognition in the warehouse world, explain the optimal ways to maximize your voice recognition quality, and finally, examine the business value you can gain by using recognized best practices in speech recognition accuracy. Vocollect Adaptive Speech Recognition expertise has evolved over the course of more than two decades of research and testing and real-world application, and this chapter shares much of what has been learned through trial and error in creating and delivering voice-centric solutions to distribution organizations around the world.

While it isn't exactly "light reading," understanding the basics of speech recognition will help you fully appreciate the content of our future chapters.

Sincerely,

Roger Byford
David Maloney

CHAPTER 2

Demystifying Voice in the Warehouse

by Roger Byford and David Maloney

At one time or another, we've all had to talk to a computer via a speech recognition system.

Let's be honest, this "human speaking to computer" type of experience has not always been positive. For instance, we may have phoned a customer service center and been asked by the electronic "receptionist" to state our reason for calling from a list of options. We spoke a response, and often the system failed to fully recognize what we were saying. We have to repeat the process, practically shouting at the phone and over-emphasizing our diction until the system finally understands. It leaves many of us frustrated and shaking our heads at such technology.

It is little wonder, then, that many supply chain managers fear that voice-directed workflow systems will offer similar results. If their experience with voice has been tainted by their encounters with consumer speech recognition systems, then they probably have low expectations for how well voice can succeed for their distribution operations in a "human talking to computer" type of arrangement. Little do they know that not all speech recognition systems are created equal.

In this chapter we will demystify voice and highlight its unique design to generate productivity and accuracy improvements in a distribution environment. We will pull back the curtain to reveal how there are different types of speech recognition systems that vary widely in their abilities to process information and provide desired accuracy. We will examine how some speech recognition systems have very high recognition rates in comparison to others, and how each is built to recognize various sizes of vocabularies, numbers of speakers, and types of environment (e.g., noisy versus quiet). We will look at how some speech recognition systems are built on sounds within words, while others use words themselves as the building blocks.

We will also explore why the speech recognition technology used in warehouse-based voice systems is a completely different animal from what we have experienced in the consumer arena. In particular, we will investigate the techniques the builders of warehouse voice systems use to create recognizers with very high accuracy.

The environment in which voice systems operate can greatly affect their performance. We will see how companies have engineered their products to match the demands of industrial warehouses. In this chapter we will also discuss how hardware, software and computer platforms can affect how well a speech recognition program functions.

Finally, since this book is directed to supply chain professionals, we will look at how the various pieces of the voice puzzle work together within the distribution environment to provide a system that is highly accurate in its ability to recognize a user's speech, using the attributes of speech interaction to build a productive workflow – thus driving increased business results.

This chapter focuses on how a computer understands speech from a human. To carry on a conversation, the computer obviously has to be able to speak to the human as well. We'll review computer speech output in Chapter 6 of *The Talking Warehouse*.

THE CHALLENGE

Voice recognition systems within computers have a challenge. When human beings work with a computer via screen and keyboard, we know that we are dealing with a dumb machine, so we are at least somewhat tolerant of flaws in the interaction. We also have rapid visual feedback, which in many cases helps out, e.g., "Oh. I misspelled that word – but the spelling engine in my word processor fixed it for me."

On the other hand, people are much less tolerant when it comes to spoken communications. Everything we've learned about live conversations has involved a second human being, so we fully expect the other party to respond intelligently, and often to know what we mean, even if we misspeak. And we expect this intelligence to apply in a wide variety of environments, whether it is a quiet room, a noisy coffee shop, or a loud concert.

It's little wonder, then, that users sometimes get frustrated even when a computer-based recognition system is near-perfect in understanding their speech. This becomes even more the case in a distribution environment, where every misunderstanding or slow response has an impact on productivity. Thus, it behooves the designers of warehouse voice systems to not only strive to maximize speech recognition accuracy, but also to be sensitive about managing the dialogue with the user as naturally and "intelligently" as possible.

I'LL KNOW IT WHEN I HEAR IT

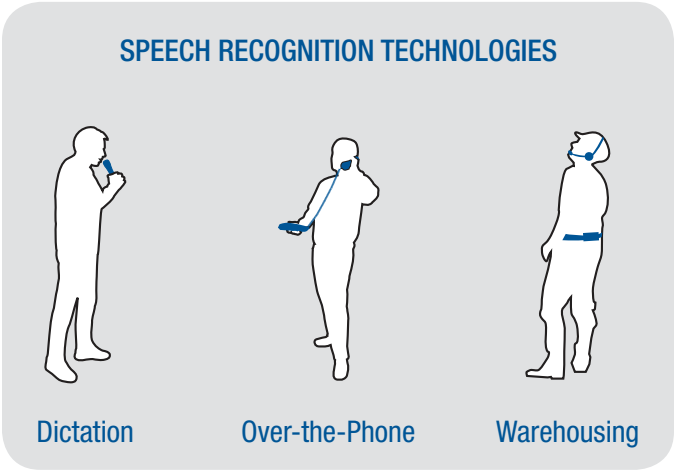
The goal of any type of speech recognition system is to interpret speech accurately and in a timely manner. Ideally, the system will understand everything the user wants it to hear, but will ignore everything she does not want it to hear, such as extraneous environmental noise that could be misinterpreted. Recognition systems also should be able to work in any language and within a variety of sound environments, whether loud, soft or shifting rapidly between the two extremes.

Meeting these goals can be a problem for humans, even though we typically make use of non-speech hints, such as visual information and knowledge of the subject being discussed, to help us out. For computers, which don't have this benefit, the problem is *very* challenging. How well a speech recognizer meets its objectives depends on the choice of the speech technology, how it is applied and how stringent are the user's requirements.

Let's review briefly the three most common applications of speech recognition today. Then we'll delve further into the challenges of the technology and how those challenges can be met.

DICTATION

Dictation recognizers are designed to, "Take a letter, Computer," and to immediately convert the speech to text. Some are tailored to specific domains, such as doctors or lawyers dictating specialist material. Others are simply intended to replace the keyboard for anyone writing a document on a computer.



By its very nature, dictation software requires a huge vocabulary – literally tens of thousands of words per language. With so many words in a dictation software program's vocabulary, it is easy to see how complexity can lead to confusion. Think of English alone. How many words can be created in English simply by changing a vowel sound or the way a particular syllable is accented?

How many words are homonyms – pronounced the same, but spelled differently, with very different meanings, such as *to*, *two* and *too*? And once these various words are put into sentence structures, even more complexity is created. Furthermore, unlike systems that ask users for specific items of information, a dictation system has (almost) no idea what the user might discuss.

Fortunately, several attributes of the application make life a little easier for dictation systems. They typically function in quiet environments. Users can be expected to work with fairly high-quality microphones and to be cooperative with the recognizer, and it's acceptable to ask users to invest a little time "up-front" to familiarize the recognizer with their speech patterns (more about this process in a moment). Finally, dictation systems are usually powered from wall outlets, so they can use algorithms requiring substantial processing power and memory.

OVER-THE-PHONE

Over-the-phone recognizers are the ones most of us have encountered. They're designed to take the place, in whole or in part, of human customer service agents. Over-the-phone recognizers are typically expected to recognize a few hundred to perhaps a thousand different words – e.g., the city names in an airline information system – which is significantly fewer than in a dictation system. However, they must work for thousands, or sometimes millions, of different people with widely varying accents. Furthermore, since users do not know the limited vocabulary that these systems have or other rules designed into them, they may say unexpected things. Finally, the speech signal the system receives may be of very poor quality, coming from a noisy environment such as an airport tarmac or train station via a poorly positioned microphone and a poor phone connection. As we'll see when we compare the overall level of difficulty for the various categories of recognizer, it's not surprising that over-the-phone recognizers don't always live up to expectations.

WAREHOUSE

A warehouse recognizer doesn't have to recognize many different words – usually less than 100. However, warehouse recognizers have to operate in extremely noisy environments. They must be able to recognize words very quickly, regardless of accent or dialect, and to deal with changes in voice patterns brought about by varying noise levels (otherwise known as the Lombard effect). They must also be able to operate dependably within temperature extremes, such as working in a freezer.

SO HOW TOUGH IS IT?

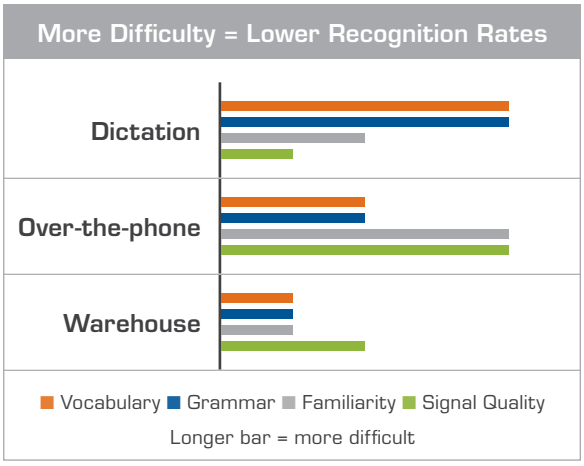
One can probably begin to see already why it's reasonable to expect, and demand, much higher performance from a warehouse recognizer than from one intended to work over the phone or for dictation. Let's summarize the most important "difficulty metrics" for each of the recognizers. Then we'll discuss some of the issues in greater detail. The measures we'll use are:

Vocabulary size – How many different words must the recognizer understand?
Grammar constraints – In how many different ways can the user enter information?
Familiarity – How much does the recognizer know about the user, and vice versa?
Signal quality – What is the quality of the sound signal the recognizer receives?

Here's one view of the relative challenges:

Challenges with Speech Recognizers			
	Over-the-Phone	**Dictation**	**Warehouse**
Vocabulary size	Medium	Large	Small
Grammar constraints	Medium	Loose	Tight
Familiarity	Low	Medium	High
Signal quality	Poor	High	Medium

And here's another:



While the graph to the right is not to scale, it clearly shows that if we design a warehouse recognizer to:

- take advantage of the limited vocabulary it is required to support,
- make good use of a constrained grammar,
- benefit from knowledge of its user, and
- minimize the impact of a high-noise environment,

then we can make the recognizer's problem much easier to solve than is the case for either a dictation or an over-the-phone recognizer. Let's look at how designers of warehouse recognizers work with these features of the application.

PHONEMES OR WORDS?

A recognizer works by taking stored "models" of sounds and comparing them with the sounds it hears from the user. For various reasons, it's not practical for a recognizer with a big vocabulary to keep models of all the words it may hear, so such recognizers (including over-the-phone and dictation systems) use *phoneme* models instead of word models. Phonemes are the smallest sound units of a language. The word *cat*, for example, is made up of the phonemes for the hard *c*, short *a* and *t* sounds. The advantage of using phonemes for speech recognition is that there aren't many of them. The hundreds of thousands of words in the English language, for example, are all formed by combining just 44 phonemes. Unfortunately, though, the pronunciation of phonemes varies, depending on the preceding and following sounds. Add the speaker's varying inflections, accents and speech patterns, and it is easy to see why speech recognition is a lot more complex than simply hearing and interpreting.

Another challenge for phoneme-based systems is their reliance on accurate phonetic transcriptions of words, which are difficult to create. This is evident by observing text-to-speech programs, which also rely on phonetics. These systems, as their name implies, turn textual information directly into computer-generated speech.

You may have used a GPS system in your car, for instance, that uses text-to-speech to provide spoken, turn-by-turn directions. If so, then you've probably encountered times when the system attempts to speak a street name and mangles its pronunciation to the point of being rather humorous. The exceptions to pronunciation rules are many and varied, so a phoneme-based recognizer must be built with a large dictionary of phonetic spellings.

In contrast to large-vocabulary, phoneme-based systems, recognizers for small vocabularies use *word-based* models. These are designed to model the sounds of an entire word, rather than piecing together phonetic segments. There's more information in a word sound than a phoneme sound, and word pronunciation doesn't vary as much as phoneme pronunciation does, so when a small vocabulary makes them practical, word-based recognizers have a performance advantage over phoneme-based ones. Therefore, almost all distribution center and warehouse systems use word-based recognizers.

VOCABULARY SIZE AND GRAMMAR

The vocabulary, as we've discussed, is a comprehensive list of words the system can understand. It contains all the words a speaker might use in a normal conversation with the system. The vocabulary can be quite vast, so to narrow down that list and thus enable the computer to understand and respond more quickly, the system focuses on interpreting an *active vocabulary* - the possible words a speaker might use at any given moment. For instance, a telephone voice system may ask a question to which it expects only a *yes* or *no* answer. The system then has to look only at those two possible responses, which greatly increases its ability to understand what the speaker is saying, and to respond in a short time. At this point in the application, we say that the grammar is tightly constrained. Unfortunately, not all answers can be limited to "yes," "no," or other short answers from its active vocabulary. That requires the recognition system to dip deeper into its full vocabulary list – thus, the grammar becomes more loose, the chance of errors rises dramatically, and the time it takes for the computer to respond may increase.

Dictation recognizers can use a "statistical language model" to improve their accuracy. This technique makes use of grammar and context rules to help the recognizer choose between more, and less likely, sequences of words. This restricts the field of possible words, but rarely to a small and easily distinguished set. A warehouse recognizer, on the other hand, can use *tightly constrained* grammar all or most of the time. Quantity information contains only digits. Yes-no responses are fairly common. Score one point for the merits of the warehouse voice application!

FAMILIARITY – TO TRAIN OR NOT TO TRAIN?

A major difficulty with over-the-phone recognizers is that they must operate in an *untrained* mode, meaning that they don't have any information about the particular person using the system. They must be able to understand first-time callers without any history of that person to guide them. Such systems must function without knowing whether the speaker is male or female, has a high-pitched voice, or speaks in a gravelly manner.

Complicating things even more in untrained systems is that we do not all speak alike. English, for instance, sounds very different when spoken in Irish, Scottish, Canadian, South African, American and Australian accents. Even within a country, accents and dialects vary greatly. Someone from Liverpool will sound quite different than a Londoner.

To bring this idea home, think about how difficult it can be for us humans to understand someone with an unfamiliar accent. At first, we may pick up only a few words. After listening for a couple of minutes, our comprehension rate rises. If the speaker continues to discuss the same subject, reusing words with unusual pronunciations, we eventually figure out those words, and our understanding increases significantly. What if a recognizer were given the opportunity to learn from its users by hearing them say, in advance, the words they will use as they work with the system? We'd expect the computer to benefit in recognition rate and response time, just as we do. And it does.

Warehouse recognizers can be "fully trained" to better understand the user, which greatly increases recognition rates. A user simply has to invest 10 to 20 minutes teaching the system how she pronounces the words used in the warehouse environment. This is done prior to using the system on the facility floor. The ability for the system to learn from the user greatly affects overall recognition and is, as we will see, a smart investment of an employee's time.

The system prompts the user to speak the words in the vocabulary, which she repeats back into the headset's microphone. The software registers these words as typical responses from her, determining her unique tonal pitch, accent, inflections and pronunciations of each word. The recognizer uses the responses to build speaker-specific models of all the words in the vocabulary. It's easy to see that having the opportunity to record and work with each user's individual voice patterns makes the recognizer's task far easier. A trained recognizer is key to achieving the high levels of performance offered by today's industrial voice systems. Another benefit is that, because a trained recognizer knows exactly how its current user speaks, it can dramatically narrow down the range of sounds it will interpret as the word "one" (for example). In noisy environments, that makes such a recognizer much less prone to interpreting background noise or "off-topic" speech as an intended utterance.

Additionally, since a trained system bases its recognition on a person's individually-recorded voice patterns, it does not have to consider how the word *should* be pronounced, only how the worker *does* pronounce the word. It no longer relies on what the typical pronunciation should be, as dictation systems would. For instance, let's look at how the system determines the number "1" has been spoken to it. When the worker trained the recognizer, she was asked to say "1." The recognizer assumes that whatever sound she made at that time represents the number 1, and it stores that sound for future comparison with her speech. Trained voice recognizers, therefore, work equally well without modification for any language or dialect.

An untrained recognizer, on the other hand, must be specifically programmed for each language it is required to recognize, and must be told ahead of time what language the speaker will be using. In the warehouse environment where multilingual workforces are becoming commonplace, workers using a fully trained recognizer can speak to the system in whatever language makes them most comfortable. Such flexibility greatly increases the ease with which workers adapt to a voice system.

We also have to consider how familiar the user is with the system he's talking to. If not familiar with it, he's likely to say things the recognizer doesn't expect. Even if the recognizer determines that the user said something that was outside the expected vocabulary, about all it can do is to either ignore the speech or apologize for not understanding it. Neither is very helpful to the user. This is an unavoidable problem with over-the-phone systems. It's less of a challenge for dictation systems, both because they are designed to respond to a very large vocabulary and loose grammar, and (for most users) visual feedback can help to clear up any confusion.

Industrial voice users work with their systems for many hours per day, conducting a relatively limited range of transactions. The associates quickly learn, either through classroom or on-the-job training, what responses are acceptable to the recognizer at each point in their dialogue with it. Unlike most users of over-the-phone systems, industrial recognizer users generally have a strong vested interest in being successful with the system and don't have the option of pressing zero to reach a human being.

The fact that the users become familiar with the system through formal training and extended use adds to the performance difference between these specialized systems and more general-purpose, over-the-phone and dictation recognizers.

Increased performance is good, but one has to ask if the cost of training a system pays for itself by improving recognition accuracy. Vocollect has evaluated the return on investment associated with trained systems in comparison to untrained systems. We first calculated the cost to train a warehouse recognizer, a process which takes 10-20 minutes. At US $30 per hour for salary and benefits, training the recognizer costs a facility about US $10 in employee time. We compared that expense with the costs the same employee would incur if he or she experienced the higher recognition error rates associated with untrained systems. The criteria for determining the additional costs were based on delays in work time caused by the need to repeat words or add extra words to allow the untrained system to process a response correctly.

According to the study's calculations, if training the recognizer reduces the error rate by one percentage point (from 2% to 1%, for example), the payback period for the investment in pre-use training (recording the reference word sounds) is less than six work days. And the ongoing cost savings per device each year is roughly US $450 for case-picking and US $2400 for piece-picking. In a two-shift operation, the annual savings per device doubles. Our head-to-head tests of several untrained warehouse recognizers suggest that the actual performance difference is likely to be several percentage points. Add up the total number of devices used in a facility, and it is easy to see that the savings a word-based, fully trained recognizer provides are substantial.

ICH SPRECHE DEUTSCH

A fully trained recognizer brings an additional benefit for multilingual workforces. In many parts of the world, a large percentage of warehouse workers are non-native. This is certainly the case in North America, where many workers speak English as their second language. In European Union countries, workers easily cross borders to work – a typical warehouse may have five or more native languages represented. Untrained systems, therefore, must be able to recognize responses laden with strong accents. But how much does an accent affect recognition accuracy?

Vocollect has conducted considerable research into this area. In a recent study, technicians ran tests comparing the performance of our own trained speech recognizer against several untrained recognizers available for the warehouse. The modeling for our study was based on non-native workers speaking English with moderate accents. The results showed that, for distribution usage, the increase in word error rate when moving from a trained recognizer to an untrained one is likely to be several percent or more. In fact, for speakers with moderate-to-strong accents, the untrained recognizers showed error rates *ranging from 6% to more than 20%,* while the trained recognizer's error rate on the same task was below 1%. This is because a trained recognizer is not affected by accent. The system simply learns how the worker pronounces the words – accent and all. Accent is no longer a factor.

In addition, the ability for workers to speak to the recognizer in the language of their choice, regardless of accent, improves job satisfaction and virtually guarantees that all workers in a facility will be able to use the voice system to its best effect. Enhancing the performance and satisfaction of a multilingual environment is, therefore, another benefit of deploying a trained recognizer in a distribution center or warehouse.

MANAGING NOISE: THE HEADSET MATTERS

We stated earlier that a challenge for warehouse recognizers is minimizing the impact on performance of the noisy environment in a distribution center. The background noise from lift trucks, fans, blaring music, public address systems and other equipment makes it challenging to have exacting speech recognition every time. Even more challenging for the voice equipment designer, warehouse systems must work in freezers, on sun-baked loading docks, and everywhere in between. While the design and implementation of the recognizer itself plays a big part in maximizing performance amid noise, it's also crucial to provide the recognizer with the best possible input, maximizing the so-called signal- (user speech) to-noise (other sounds) ratio. Vocollect designs or directs the design of all of our own headsets, and we do this for a reason. Off-the-shelf headsets are simply not optimized to work for rugged material-handling applications.

A number of factors determine headset performance and ultimately affect the accuracy of speech recognition. For example, the microphone must have good noise cancellation characteristics, eliminating noises that originate far from the microphone, while capturing sound that comes from the user speaking directly into the microphone. Noise-canceling microphones require very tight manufacturing tolerances, making them expensive; however, a poor microphone will compromise the performance of the entire system. Unfortunately, the response of a noise-canceling microphone varies widely if it is moved, relative to the user's mouth, by even a small distance. This means that the recognizer must be designed to compensate, as best it can, for any such variations. The headset should be designed to make it easy and intuitive to position the microphone the same way, all day and every day.

The microphone must also perform in extreme conditions, even after being subjected to dirt, debris, and liquids common to distribution environments. The microphone and terminal circuitry have to work together and must be built with an adequate *dynamic range* to accommodate quiet and loud speakers in quiet and loud environments. This problem is made more challenging by that previously-mentioned Lombard effect – whether they need to or not, people naturally speak more loudly in a noisy environment. So a naturally loud speaker working in a noisy environment is likely to speak extremely loud! The microphone must be able to pass this very loud speech to the recognizer without any distortion from overloading.

WHAT ELSE?

Let's assume we've built the world's best warehouse recognizer, making all the previously-described right design decisions. There's a lot more to do before we have an effective voice system for use in warehouses. Here are some other important criteria.

Adaptive Recognition

Thus far, we've assumed that users will speak the same way each and every time they talk to the recognizer. But, in practice, that isn't so. After eight hours of work, a user may be tired, or perhaps he came in with a cold. Such variations can change how a person pronounces words and, in turn, how the system interprets his sound.

An *adaptive recognizer* has the ability to continue to learn more details of the user's speech patterns as he or she interacts with it over time. For example, whenever the person speaks a particular digit, the system registers the response and fine-tunes its model of how the user is saying that word. Adaptation algorithms must be designed very carefully. Not all user utterances are good examples – we all have times when words just don't come out right, and in an industrial environment, some utterances will be corrupted by extreme background noise. The adaptation algorithm must take into consideration these anomalies and ignore them.

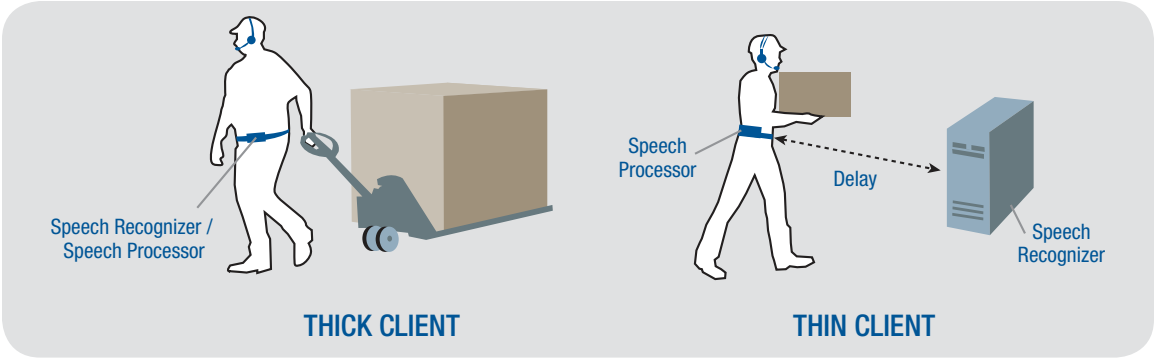
In comparison, an over-the-phone recognizer cannot generally take advantage of adaptive recognition, because it normally assumes it will hear a user only once. Since the next caller will have a completely different voice, there is no opportunity for the system to learn or adapt to a unique voice. Some dictation software programs do have a limited adaptive recognition capability that improves accuracy over time. However, adaptation is inherently less effective for a phoneme-based recognizer than for a word-based one, because in a word model, each component phoneme is adjusted to exactly how it is pronounced in that specific word, not a blend of how the phoneme is pronounced against all words. Adaptation at the phoneme level won't help if a user pronounces a word in an unusual way.

Responsiveness

Those old enough to remember making transatlantic phone calls via satellite will recall that the transmission delays, although quite short, made conversations surprisingly awkward. We humans are very sensitive to "conversational delays" when we don't have any visual cues to show us that

the other party heard us. Few things do more to reduce the productivity and overall satisfaction of a speech recognition user than a system with a slow or inconsistent response time. Increases of even fractions of a second to response times can make a pleasant conversation painful. If the user begins to think the recognizer hasn't heard her, she may begin to repeat herself. The recognizer, having heard her but been slow to respond, may interpret the new speech as another data item. Now the user and the recognizer are "out of sync." Delays and frustrations are inevitable.

For this reason, it is better for speech processing to be conducted directly on the wearable or portable computer that is with the user at all times. While it requires that these portable devices have significantly greater processing power, the gains in speed and performance provide this architecture, known as *thick client*, with a superior advantage. With thick client, the voice recognizer and the text-to-speech engine reside on the terminal. With each terminal's processing power fully dedicated to its one user and no requirement to move the speech data across a wireless network, a thick-client system can guarantee consistent and rapid response to its user's speech.



The opposite of thick client is known as *thin-client*. With thin-client, the voice recognizer does not reside on the device. Instead, data must be transmitted to a server for processing and speech recognition.

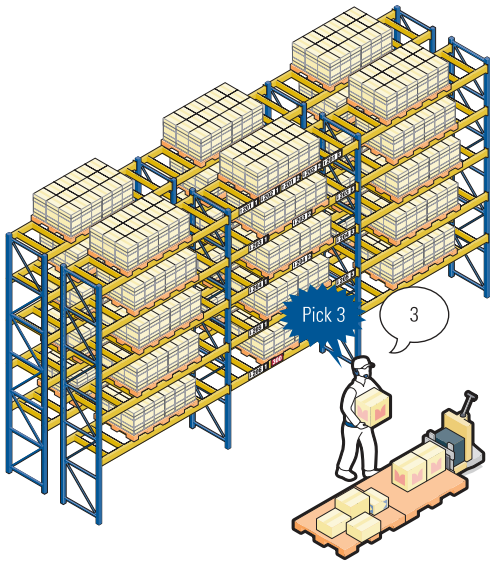
This can cause delays, depending on network design, server software implementation, and how many users are attempting to interface with the server at the same time. As we've noted, even very brief delays will seriously degrade the user's experience with the product.

Thin-client devices are also completely undone by connectivity problems. Areas of weak Wi-Fi coverage in a facility dramatically increase the likelihood that data cannot be exchanged properly. Since a thin client requires nearly continuous high-speed communication between the device and the server, workers may have to return to an area of the warehouse where they can receive a stronger signal before proceeding with their assignments. Thin client is also more taxing on an IT infrastructure. Each device with marginal connectivity to the network can affect all of the other devices, by increasing the number of data collisions and retries. This can lead to productivity-sapping delays.

PUTTING IT ALL TOGETHER

Now that we have detailed many aspects of how the various speech recognizers work, let's further examine how they are applied in a warehouse workflow. It's important to note that even a recognizer that can distinguish one word from another with unerring accuracy and blinding speed may not provide a satisfactory experience for the user, if not designed and applied properly. Let's see how a warehouse voice system leads its user through a typical transaction, then we'll see what we have to build "around the recognizer" to create a high-quality user experience.

The *application engine* manages the prompts delivered to the user and specifies the user responses the recognizer should expect. It has access to information about what the user is to do, and the sequence of commands and responses that should be exchanged with the user. The application engine communicates with the recognizer via the *dialog engine*.



Let's take the common voice task of picking a warehouse order as an example.

Selection is very much like grocery shopping. Items are stored in *slots* in the warehouse racking. The wearable computer being used by the order picker (or *selector*) has a list of slots and item quantities that make up an *assignment*. The only real difference from grocery shopping is that the items to be picked are specified by location, e.g., "*aisle 3, slot 124,*" not by description, e.g., "*banana*."

After the selector completes a pick, the application engine looks in the assignment list to find the next slot to be visited. If it's in the same aisle as the current slot, the application builds a prompt for the selector that simply says, "*Slot 135*." (Warehouse voice applications are typically extremely concise – we don't want to waste the workers' time listening to unnecessary speech). The application engine sends the prompt to the dialog engine – an instruction for the recognizer to prepare to hear a sequence of (typically) three digits, and instruction that this prompt is not interruptible. The user must then listen to all of it before responding. A system's ability to offer both interruptible and non-interruptible prompts is essential. We must ensure that the user hears all essential information. However, at other times we may want the user to be able to interrupt the dialogue. For example, a user may ask for some kind of status report and then decides he doesn't need to hear all of it. It could be very frustrating if he can't interrupt and tell the system to move on.

Once the system issues the "*Slot 135*" instruction, the speech recognizer is started so that the system can listen to the response given by the user. In this example, the user would read off a *check digit sequence* to confirm the location. A check digit sequence is typically a two- or three-digit number physically attached to the facing of each storage position. Upon arrival, the worker must read off the correct number to confirm that he is at the correct location. The user reads the check digit, "*373*," and the recognizer interprets the spoken numbers and compares it to the models of the words in the active vocabulary that the system expects the user to say. If the recognizer finds a word or sequence of words it is confident in, then it passes the match to the application engine for processing. If the recognized speech does not match the correct check digit, then the prompt is repeated. If the recognizer hears a sound that it cannot find in the grammar, then it ignores the sound, assuming that it is extraneous noise that has entered the user's microphone (or perhaps the user talking to a friend about last night's game).

Warehouse voice systems must be "always on" and listening to the worker in case he needs to communicate with the system, perhaps asking for a prompt to be repeated, seeking a product description, or logging off of the system. Thus, the user does not have to push a talk button to speak (as is required on some voice systems). A button takes an extra step to push and does not keep both hands free. Because the software is continuously monitoring the user, the system must be carefully engineered to evaluate the noises it hears and to disregard extraneous noises or "off-topic" user speech, rather than interpreting them as intended responses.

For both maximum productivity and user satisfaction, the system should also minimize the number of words the user has to say. Some less effective voice systems require *anchor words*. These are words the user must provide to explain what will follow. For instance, before giving a check digit, the user may have to say "*Check*" to alert the system that the response following it will consist of numbers corresponding to check digits. Anchor words may also be added to the end of a sentence, such as "*Stop*" to tell the system that the response is completed. Additionally, a less effective voice system may have to add extra steps in its prompts to gain accuracy, such as confirming that what it thinks it heard it actually did hear, e.g. "*You said three, are you sure?*".

Let's continue with our processing example. Following confirmation that the user is at the correct location, the application engine gathers the next needed command, such as "*Pick*," and adds the quantity to be selected from information relayed by the warehouse management system. The dialog continues as the user hears: "*Pick three*." He confirms the quantity selected by saying "*Three*," and again the recognizer compares the response to the word models and finds the best match. All of this complex process happens as quickly as the user can speak and listen, with no delays.

Once the user confirms the correct location and the pick quantity, the process repeats itself for subsequent selections. With a high-quality recognizer using adaptive technology to maintain its performance over time, and with a well-designed system surrounding it, task performance can be extremely high. Workers never have to slow or interrupt their primary function in order to interact with the computer.

ENDLESS POSSIBILITIES

In this chapter we have attempted to draw a solid line of separation between consumer and industrial speech recognition systems. As conveyed, the recognition accuracy of speech recognition technology in warehouse settings is light years beyond that of commonly experienced consumer applications. We have looked at the components and variations of warehouse speech recognizers, and we have examined a host of factors that directly impact the quality of speech recognition. Finally, we have made a strong business case about the value that can be assigned to maximizing a voice system's speech recognition quality. In today's world where companies must eliminate as much waste from their supply chain as possible, getting the greatest payback from a voice system via good speech recognition is a bottom-line issue and opportunity.

In our next chapter, "Solving Warehouse Challenges Using Voice," we will explore how voice technology has helped many companies solve common strategic business challenges. We'll also look at how companies using voice have experienced significant boosts in productivity, accuracy and safety, reduced training time, and improved customer satisfaction.

Vocollect® and Vocollect Adaptive Speech Recognition™ are trademarks of Vocollect, Inc. All rights reserved.

"The Talking Warehouse: Understanding How Voice Unleashes Higher Performance in Product Distribution",
© 2010, Vocollect, Inc. All rights reserved.