Vocollect®

#1 in the Business of Voice™

Speech Recognition Technology Choices for the Warehouse

August 2010

A Vocollect White Paper

Table of Contents

Executive Summary

This White Paper reviews some of the basics of computer-based speech recognizers, including the challenges of maximizing performance. We then discuss technology choices in light of the needs of warehouse workers using speech recognizers, and calculate the cost of recognizer errors in the warehouse environment. For one of the major design decisions – whether to use a trained ("speaker-dependent") or an untrained ("speaker-independent") recognizer, we demonstrate that time spent in training the recognizer is likely to be repaid in a few days by the benefit of improved performance.

What Do We Want from a Speech Recognizer in a Warehouse?

Before we delve into technology details, let's take a look from the user/customer's point of view at the goals for a warehouse speech recognizer. In an ideal world the primary goal would be, "Understand instantly and correctly everything the user wants you to hear, and nothing he/she doesn't." As we'll see below, however, this ideal is impossible to reach, so we have to talk about getting as close as we can. Understanding that constraint, a warehouse recognizer should:

- Work effectively in a wide range of noise environments, including very quiet, very noisy, and rapidly changing.
- Work effectively for the widest possible range of facility employees, regardless of gender, language spoken, accent, speech patterns, etc.
- Respond "instantly" to operator speech, to eliminate both cost and user frustration caused by delays. Minimize "total cost of use," including time lost to any pre-use preparation required and to recognizer delays and errors while the user is working.

How Speech Recognizers Work

Computer-based speech recognizers match patterns. The sequence of events (much simplified) is as follows:

- The recognizer loads a set of sound reference patterns, which represent either words or partial words ("phonemes") that the application expects the user to say.

Technologists use the term “active vocabulary” to mean the list of words the user can be expected to say at any instant, and “vocabulary” to mean the list of all the words the user may speak while working with the application.

- The application then passes to the recognizer an unknown sound, which represents either an utterance (word or series of words) spoken by the user, or an extraneous sound, or perhaps an utterance contaminated by extraneous sounds.
- The recognizer “classifies” the unknown sound, and reports the best possible match between the unknown sound and a series of one or more of the reference patterns. If, for example, the reference patterns represent digits, the recognizer may report that the unknown sound best matches the digit sequence 123. The recognizer also typically reports on the closeness of the match between the unknown sound and the reference patterns. If the match “score” is poor, the application may decide that the unknown sound was most likely an extraneous noise, rather than user speech. In that case the application will ignore the reported output.

Why is Speech Recognition So Difficult?

Speech recognition (by humans and computers) would be a relatively easy problem if humans spoke identically and consistently. But we do not. Speech utterances are like snowflakes – no two are exactly the same. Person A’s way of saying “one” may be very different from Person B’s. What’s worse, even if Person A repeats the word “one” several times in a row, each repetition will be subtly different. Speech recognizers are further challenged by other effects:

- When we speak multiple words without pausing between them, the way we pronounce each word is affected by the words before and after it. This is called co-articulation. This also affects sounds within words, so the same sound may be spoken differently depending on the sounds before and after it.
- Utterances can be corrupted by background noise.
- The application may pass to the recognizer an extraneous noise that doesn’t contain any user speech.
- The sound the user makes may not be accurately conveyed to the recognizer (e.g., when there’s a telephone connection between the two).

When recognizers make errors, which all – both human and computer – do, those errors come in three flavors – insertions, deletions and substitutions:

Speech Recognition Error Example		
Error Type	**Speaker Says**	**Recognizer Thinks Speaker Said**
Insertion	**<Nothing>** **One Five Three**	**One** **One Five Nine Three**
Deletion	**One** **One Five Three**	**<Nothing>** **One Three**
Substitution	**One Five Three** **Five**	**One Nine Three** **Nine**

As you might expect, optimizing a recognizer's performance across all three categories of error is a challenge. For example, "tuning" the recognizer to be less susceptible to insertion errors tends to increase susceptibility to deletion errors.

Making the Problem Easier

An important goal for anyone designing a system with a speech recognizer at its core is to minimize recognizer errors by making the recognizer's problem as simple as possible. There are numerous ways to do this:

- We can constrain the recognizer's vocabulary. In a dictation system, the constraints we can apply are very limited – the user may say almost anything at any instant. In an industrial application, if the system is asking the user to enter a quantity, we can make the recognizer's problem far easier by telling it to expect only strings of digits.
- We can insist on a working environment that limits background noise. That is reasonable for a dictation system, but impractical for an industrial system intended to support workers in factory or warehouse environments. In this kind of environment the best we can do is to minimize background noise through the use of special "noise canceling" microphones, and to use algorithms in the recognizer that attempt to minimize the effect of "noise contamination."

- We can allow the system to make use of knowledge about the user. Again different systems have differing levels of ability to apply such knowledge.
- An over-the-telephone system, such as an airline information provider, cannot realistically require users to identify themselves, and transactions are so short that the system can learn, during the transaction, very little about the user's speech patterns.
- A dictation system can require new users to speak to it, typically by reading one or more fixed scripts totaling perhaps five to fifteen minutes in length, before they use the system. This allows the recognizer to gain information about the user's "voice type" (e.g., high-pitched vs. low-pitched) and accent.
- A "small vocabulary" system, such as used in a warehouse, can require new users to speak to it the specific words he or she will speak while working. The system can then form user-specific "voice templates" for each of the words in the vocabulary. A system that makes use of this kind of knowledge about the user is known as "fully trained," or "speaker-dependent."
- Some recognizers require users to speak "anchor words" before, or before and after, each utterance (e.g., "start 1 2 3 stop"). While anchor words can improve some aspects of performance of a recognizer with weaknesses, they also substantially increase the number of words the user must speak, which has a negative effect on productivity. At Vocollect, we've chosen to ensure that our recognizer provides optimal performance without putting on the user the additional burden of speaking anchor words.
- The most advanced recognizers learn about the user while he/she is working. They use that knowledge to further improve performance. Vocollect refers to this technology, which we incorporated into our products in 2006 and continue to refine, as "adaptive recognition".

Choices for a Warehouse Speech Recognizer

When designing a speech recognizer for warehouse use, some decisions are easy to make, while others require more thought. It is clear that one should:

- Constrain the vocabulary to match the task
- Use equipment (e.g., microphones) and algorithms to minimize the impact of background noise
- Avoid the use of anchor words
- Adapt to the user

The primary remaining design decision is whether to use an untrained ("speaker-independent") or trained ("speaker-dependent") recognizer. Let's examine the characteristics of warehouse applications that affect the choice of technology:

- Small, fixed vocabulary – this characteristic, not present in many other speech applications, allows a fully trained recognizer to be used.
- Large number of transactions per user and high transaction rate– as explored below, these traits make recognition accuracy and response speed important, because errors and delays can add up quickly.
- Multi-lingual, non-native workforces – requires broad language, dialect, and speech pattern coverage.
- Must work for every user – there are no practical alternative means of entering the data.
- Short phrases and short words, spoken in noisy environments – short phrases and short words can cause insertion errors, so imperviousness to background noises is important.
- Changing speech patterns and background noise – Users' speech patterns change for many reasons; for example, they may change as they tire over the course of a shift.

The following table revisits the major design goals for speech recognizers in the warehouse, indicating whether an untrained or a trained recognizer has an innate advantage for each.

Goal	Trained 'Speaker-Dependent'	Untrained 'Speaker-Independent'
Minimize pre-use training time		✓
Maximize accuracy & worker productivity	✓	
Work in any language	✓	
Be impervious to accents, voice type, gender, etc.	✓	
Maximize background noise rejection	✓	
Maximize user satisfaction	✓	
Maximize benefits of adaptive recognition	✓	

Cost Impact Calculations

Clearly any calculation about which type of recognizer to use should consider the cost of training against the benefits of improved performance. While we cannot readily measure the benefits or cost of user satisfaction, we can fairly readily estimate the costs of pre-use training, and of in-use errors.

We use the following assumptions:

Operator payroll (salary and benefits) cost:	$20 / hour
Voice device usage:	8 hours per day, 360 days per year
Transaction rate (e.g., picks per hour):	200
Spoken words per transaction:	4
Time cost of recognition error:	3.5 seconds
Pre-use training time:	20 minutes

Notes on the assumptions:

- Voice device usage could be lower if the warehouse operates 5 days per week rather than 7, but could be substantially higher in a multi-shift operation.
- The transaction rate corresponds to a typical “case pick” operation. Other tasks could have substantially higher or lower transaction rates.
- Four words is typically the minimum for any warehouse operation – a simple “no exceptions” pick transaction in which the operator speaks 3 check digits to confirm the pick location and a single digit to confirm pick quantity.
- Vocollect has measured through observation the time cost of a recognition error. The recovery time from an error may be longer than the figure used here, but experienced operators can sometimes continue to work while recovering from an error.
- The pre-use training time figure is typical for a Vocollect warehouse voice system.

Given the assumptions we can calculate:

Words spoken per day = 200 * 8 * 4 = 6,400

Vocollect has devoted substantial time and effort, over a period of years, to creating a database of many hours of speech examples from many users working with our voice systems in warehouses. We use this database to give us the best possible estimate of error rates users are likely to see "in the real world." Whenever we make an enhancement to our recognition algorithms we first test the change against this database, then we verify performance with field tests. The use of a data set specifically recorded for applications in the warehouse provides us with very good correlation between improvements in the lab and results reported from the field.

Demonstrably real in Vocollect's experience, but even more difficult to measure, is the impact of poor recognizer performance on employee satisfaction, overall employee performance, and equipment abuse. Our conviction that such "soft" issues are real and important drives us not only to continually seek recognizer performance enhancements but also to focus on product and service attributes that go far beyond recognizer algorithms, including, for example, a wide range of headset design and user training issues.

Test Results

Vocollect recently ran tests, using our in-warehouse database, to provide a performance comparison of our own trained speech recognizer against several of the untrained recognizers available from others (including the one most commonly deployed in warehouse voice systems from other suppliers).

The results of our tests suggested that, for warehouse 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 increase ranged from 6% to more than 20%. As the graph below shows, the resulting cost increase per voice unit in use, given the assumptions above, could easily exceed $1,000 per year, even in a single-shift operation, and could be several thousand dollars per year in a multi-shift facility.

It's important to note that the analysis in this document applies only to voice in the warehouse. Vocollect, for example, has developed and deployed an untrained recognizer for use in our healthcare business, where the application has very different attributes. But we continue to believe very strongly that our trained recognizer is the product of choice for use in the warehouse.

So that for every 1% increase in "word error rate" (e.g., the number of times the recognizer makes an error goes from 1 per hundred words spoken to 2 per hundred), we have:

Increased errors per day = 6,400 * 1% = 64

Time lost per day = 64 * 3.5 = 224 seconds = 3.7 minutes

Time lost per year = 3.7 * 360 / 60 = 22.4 hours

Cost per year = 22.4 * 20 = $450

Note that errors can be any of the three types described above. Untrained recognizers are particularly susceptible to insertion errors.

Therefore, if using a trained recognizer decreases the word error rate by only 1%:

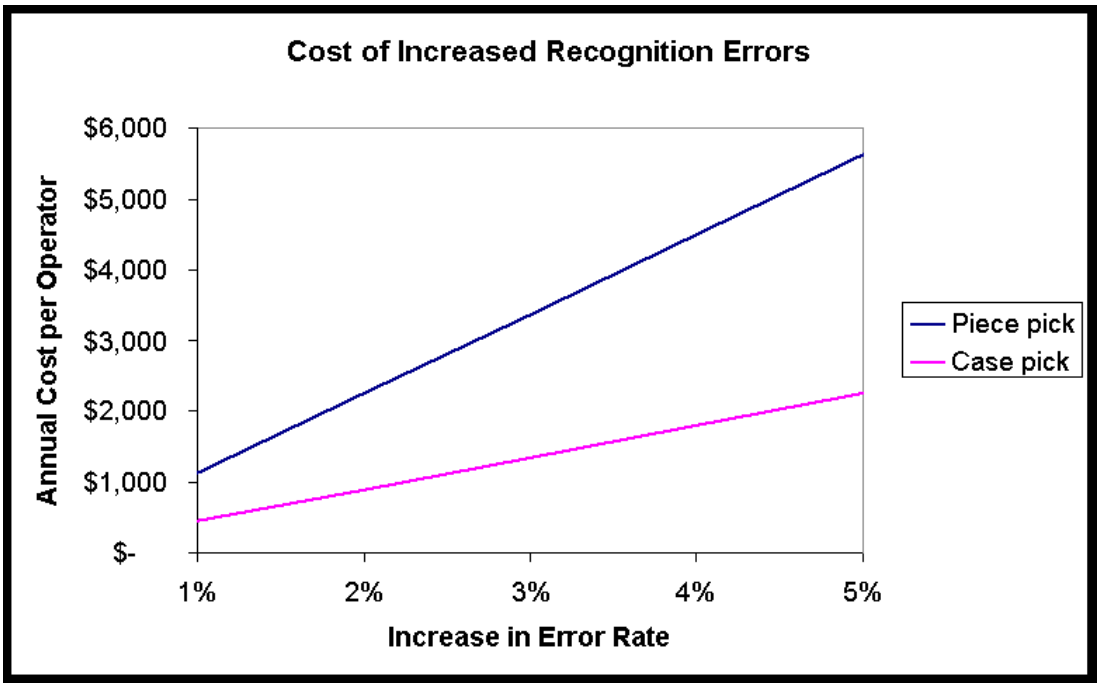
The payback period for the investment in pre-use training is less than 6 workdays.

The ongoing cost savings are $450 per operator per year.

As we shall see below, 1% is a very conservative estimate of the difference between a trained and an untrained recognizer.

Measuring Performance

It is notoriously difficult to measure recognizer error rates in a meaningful way. A company might claim that its recognizer is "99.7% accurate." And it's certainly possible to devise a recognizer test that will show 99.7% accuracy (0.3% error rate) for almost any recognizer. But it's equally possible to devise a different test that would show error rates of 10% or more for the same recognizer (30x worse performance), even for a seemingly simple task such as recognizing "yes" and "no." Recognizer accuracy claims should, therefore, be taken with very large doses of salt. The best any company can do is first to measure, using a large data set, the results its users experience. This is very expensive and time-consuming. Then it's necessary to create a test environment that replicates, as closely as possible, those real-world results. Now one can hope that if one makes changes in the recognizer and sees an improvement in test results, then users in the field will experience a similar improvement. Periodically, it's necessary to re-calibrate by gathering more real-world data, using the latest version of the recognizer, and again comparing the results against those from the test environment. Even with such rigorous efforts, it's very difficult to make useful and credible quantitative accuracy claims. A substantial improvement, for example, in rejection of background noises, will have no impact in an environment that doesn't have such noises.



Note: **Piece pick rate assumed at 500 lines per hour, case pick at 200 lines per hour.**

Conclusions

In warehouse applications, an untrained recognizer has the advantage that it does not require an initial investment of user time to perform the training. But a trained recognizer will generate far better returns in the long run. The application characteristics not only allow a fully trained recognizer to be used in the warehouse, they make it the obvious optimal choice for anyone who designs a recognizer specifically for the warehouse. First and foremost, they offer higher accuracy because they are able to better differentiate and recognize how each individual speaks each word – they do not need to allow for all the pronunciation variations of a region or language. This specialization also better enables them to reject sounds that should not be recognized, preventing costly insertion errors.

Furthermore, the changes in speech patterns that occur over time make adaptive recognition an obvious choice for warehouse applications. While both fully trained and untrained recognizers can be adaptive, fully trained recognizers, which use models of complete words, can achieve higher accuracy using adaptation than can untrained recognizers, which use models based on phonemes (the individual sounds within words).

The universal acceptance of Vocollect Voice by more than 300,000 users in dozens of countries, speaking scores of languages and many more dialects and local accents, is a testament to the success of Vocollect's trained recognizer approach.

To summarize:

1. The attributes of warehouse applications of speech technology strongly favor the use of a fully trained speech recognizer over an untrained recognizer.
2. In a warehouse application the minimal cost of training a recognizer is far outweighed by the improved performance that training provides.
3. The operating cost savings provided by a trained recognizer, as compared to an untrained recognizer, range from hundreds of dollars per year per operator to thousands, depending on application attributes and relative recognizer performance.
4. Trained recognizers offer significant additional advantages in supporting multi-language workforces.

About Vocollect

Vocollect, Inc. is the number one provider of voice solutions for mobile workers worldwide, helping customers achieve a higher level of business performance through voice. Every day Vocollect enables over 300,000 workers worldwide to distribute more than $3 billion dollars' worth of goods from distribution centers and warehouses to customer locations.

A global team of over 2,000 supply chain reseller and channel partner experts supports Vocollect Voice offerings in 60 countries and in over 35 languages. Vocollect's VoiceWorld Suite integrates with all major WMS and ERP systems, including SAP, and supports the industry's leading mobile device solutions.

For more information, visit www.vocollect.com

Vocollect North America:
info@vocollect.com
+1.412.829.8145

Vocollect EMEA:
voc_emea@vocollect.com
+44 (0) 1628.55.2900

Vocollect APAC:
apac@vocollect.com
+852 3915 7000

Vocollect Latin America:
latin_america@vocollect.com
+1.412.349.2675

Vocollect Japan:
japan@vocollect.com
+813.3769.5601

Vocollect Singapore:
singapore@vocollect.com
+65 6248 4928

#1 in the Business of Voice™

Published by Vocollect, Inc.
703 Rodi Road, Pittsburgh, PA 15235
(412) 829-8145, Fax (412) 829-0972, http://www.vocollect.com