# Contents

1. Introduction ................................................................................................................................. 3  
   1.1 Radio Data Terminals ........................................................................................................ 3  
   1.2 Where Does Voice Technology Fit Into The Warehouse Infrastructure? .................. 3  
   1.3 Recent Successes .............................................................................................................. 4  

2. An Overview Of The Voice Technologies ................................................................................. 5  
   2.1 What Is Voice Technology? ............................................................................................... 5  
   2.2 The Elements Of Voice Technology .................................................................................. 5  
      2.2.1 Speech Synthesis ........................................................................................................ 5  
      2.2.2 Speech Recognition .................................................................................................... 5  
      2.2.3 Speech Processing On The Mobile Device Vs Remote Speech Processing ............. 6  
      2.2.4 So What Is Voice Technology Good At? .................................................................... 7  
      2.2.5 ...And What Is It Not So Good At? ............................................................................. 7  

3. How Can Voice Technology Be Used Within The Supply Chain? ........................................... 8  

4. Simplified Architecture Of A Voice Enabled Supply Chain Solution ................................. 8  

5. Summary ................................................................................................................................... 10  

6. Appendix A – Summary of the features of speech recognition methods ............................... 11  

7. Appendix B – Specification of the Talkman2™ terminal .......................................................... 12
1 Introduction

This paper reviews the technologies that are available for implementing voice-enabled systems within the supply chain.

It addresses the critical issues concerning voice technologies:

- The benefits that voice technology can deliver within the warehouse environment
- Selecting a voice technology
- The components of a voice enabled system.

The document discusses Microlise’s selection of a voice technology partner and the implementation of a pilot system for a high-street retailer. This includes the methods that have been employed to develop the pilot system, and the components that have been developed to deliver a robust solution.

The document considers the voice technologies separately from the existing business systems to which they are connected. However, some of the document is concerned with the means by which voice technology can be integrated with these systems to produce a complete solution.

1.1 Radio Data Terminals

Radio Data Terminals (RDT) are currently the most popular method by which warehouse staff interact with Warehouse Management Systems (WMS). They provide a portable computing platform for the presentation, capture and validation of real-time data.

They typically have the following properties:

- A rugged casing and battery-based power supply
- A small screen and an alphanumeric keyboard
- Wireless communication with a central application server
- Ability to run local applications.

RDTs are well suited for use by semi-skilled operators who need to perform repetitive, well-defined tasks within a warehouse environment.

The addition of voice technologies into a warehouse operation can offer benefits over and above those delivered by traditional RDTs alone. Voice terminals are are specialised RDTs, typically rugged, small wearable devices equipped with an ear-piece and microphone integrated into a head-set, which require no physical intervention from the operator during use. They can be fitted with barcode scanners, printers or other peripherals.

1.2 Where Does Voice Technology Fit Into The Warehouse Infrastructure?

Over the past few years, a number of manufacturers have produced wearable computers which provide a voice operated interface to the backend WMS. The major benefits of such units are that they provide a truly hands-free and eyes-free operating environment.

- Hands free; the operator doesn’t have to keep picking up and putting down a hand-held device.
- Eyes free; the operator doesn’t need to look at a screen to receive instructions, which improves safety and enables total concentration on the task in hand.

Voice technology is practical for use where manual-handling forms a major part of a task. It is safe for use on fork lift trucks because it requires no visual attention whilst in use and no distracting hardware installations on the FLT itself.
However, the first commercial products in the 1990s proved to be expensive and unreliable. The anticipated take up by the logistics industry did not materialise. Some costly implementations failed to meet expectations, and the market has remained sceptical of voice-enabled systems for some years.

Recent voice technology offerings have become more cost effective and usable due to increased development effort by major vendors such as IBM, Microsoft and Phillips. The availability of relatively powerful mobile computers means that it is now realistic to utilise this technology in a mobile environment.

1.3 Recent Successes
A number of voice technologies have been adopted by solution vendors within the supply chain sector. This has resulted in some successful, large-scale deployments, primarily in the US, but with increasing penetration into the European market. One supplier claims that its equipment is processing over 10 million voice transactions per day.

Whilst it is possible to develop voice enabled systems from first principles, using the voice capabilities of the Microsoft Speech API (MS SAPI) or the Java Speech API for example, a successful system requires a specialised, high quality hardware platform on which to run. The hardware platform is especially important for the mobile solutions which Microlise develop, and has to meet similar performance criteria to those demanded of other RDTs, such as:

- An industry standard operating environment (such as Windows CE).
- Ruggedised devices meeting IP standards.
- Connectivity with major RF networks.
- Full-shift battery life.
- Low training requirement for users.
- Easy availability of servicing/maintenance services

The Microlise Approach
Microlise have chosen to partner with Voiteq, the UK distributor of Vocollect products. The core of the Vocollect product range is the Talkman 2™ Voice terminal, which we believe is the most suitable device of its type for use in the supply chain.

Vocollect have many customers in the US, with thousands of Talkman units in use within the supply chain.

As of November 2002, we are developing a pilot picking solution for an existing Microlise customer.

All examples regarding the implementation of Voice Technology in this document assume the use of the Vocollect product family.
2 An Overview Of The Voice Technologies

2.1 What Is Voice Technology?
When we talk about Voice enabled systems, we think primarily of a user interface through which all interaction takes place via spoken phrases: the user ‘speaks’ to the system, and the system ‘speaks’ back to the user. It is important to understand that merely implementing a voice user interface does not bestow any additional intelligence upon the system. The user can no more engage the system in a normal conversation that (s)he can type random words and phrases into a traditional graphical user interface and expect to get meaningful results.

A voice interface at it’s best is a simple and intuitive way for a user to interact with a system without intruding in any way upon the manual task being performed.

The use of voice technology does not preclude other complementary human-computer-interfaces. Voice can successfully be combined with visual displays or barcode scanners to give a hybrid user interface. We shall henceforth consider voice only interfaces.

2.2 The Elements Of Voice Technology
A voice-only interface comprises two kinds of speech processing: Speech synthesis and speech recognition.

2.2.1 Speech Synthesis
Speech synthesis is the means by which data from the host system is converted into spoken words that the operator hears through a headset. There are broadly speaking two methods by which speech synthesis can be achieved. One involves the use of a human speaker to record every word that the unit needs to say. These word ‘samples’ are stored in a database and then strung together and replayed at runtime, making coherent sentences. This is expensive, in terms of the overhead required to record the words, and the maintenance effort to re-record the samples as applications change and new words are required. In addition, it cannot be assumed that simply chaining together a collection of recorded words will sound anything close to natural speech.

A simpler and cheaper option is to use a real-time text-to-speech synthesizer, which takes an ASCII input string and creates an audio output stream suitable for feeding into a headset. The advantages of text-to-speech are that the synthesizer can say any word sequence without requiring any development effort, and that the synthesized speaker can be customized at will: for example by changing the pitch or speed of the spoken output. There are numerous text-to-speech engines available, and most produce an acceptably intelligible output. Operators quickly become accustomed to the synthetic ‘accent’. Where specific words cannot be pronounced accurately by the engine, alternative phonetic spellings can be used to give an acceptable rendition (The word ‘Microlise’ is a good example of this. Text to speech engines pronounce it “Microleeze”, but this can be corrected by using the alternative spelling “microl eyes”).

The nature of most of the tasks which are implemented with speech technology means that users will be hearing the same phrases over and over during a shift, and so are using the prompts as a confirmation of an action rather than as detailed instructions.

2.2.2 Speech Recognition
Speech recognition is the means by which the user can interact with a system through the use of voice commands only. It is the reverse of speech synthesis in that it converts speech into text. It is a critical part of a voice enabled system, and the quality of speech recognition has the greatest effect upon the success of the overall system.
Speech recognition is a complex problem for software designers to solve, and although commercial applications have been around since the 1970s, few would suggest that a perfect solution yet exists. Speech recognition requires complex speech analysis which uses a large amount of processing power. Larger vocabularies increase the analysis needed to differentiate between similar words since the larger the vocabulary, the more difficult it is to differentiate any word from all the other words in it. Systems which are speaker independent also have to be able to recognize the same word pronounced by different people. Of course, as more processing power is needed to process the speech, response times increase, and so system designers are therefore forced to make compromises between quality of recognition and performance. Speech recognition engines vary in the way in which they separate one spoken word from the next. When we speak, we do not put explicit gaps between words i.e.

I…would…like…to…buy…

Words run into each other with breaks often occurring at punctuation marks only. This can make life difficult for a speech recognition engine. These two sentences illustrate the potential problem:

I would like twenty for myself.
I would like twenty-four myself.

However, any usable application must be able to support recognition of continuous, unbroken speech, since it is unnatural and inefficient to make users speak discrete words. Methods discussed here all support continuous speech recognition.

There are two principles of speech recognition in widespread use.

In **Speaker Independent Recognition**, the vendor supplies a database of many everyday words which have been analysed from samples spoken by a number of regional speakers. When the user speaks a sentence, the voice pattern is broken into ‘words’ and the database is searched for a close match. Speaker independent speech processing requires a lot of calculations to make the match, and so in a mobile environment, speech samples from the user are usually sent across the radio link to a central speech processor. The matched words are then sent back to the mobile device, and the application takes appropriate action.

Speaker Independent engines are typically used in applications where the identities of users are not known in advance, such as a call centre system. They have the disadvantage that due to wide variations in accents and speech patterns, they only have an 80% average recognition rate although they are more suitable for applications involving a large vocabulary.

The other common form of speech recognition is **Speaker Dependent**, and uses a **Voice Template** for each user. Before the application can be used, each user ‘trains’ the system by reading all of the words that the application needs to recognize. The voice patterns for each word are then stored in a unique template for that user. When the application runs and the user speaks, there is only a small set of voice patterns to search for a match. Since the voice patterns in the database are based upon the user’s own speech, the matching process is greatly simplified, and can be performed on the mobile device rather than on a central server. The recognition accuracy of systems that use a trained voice template can approach 100%, and will always be better than speaker independent voice recognition.

### 2.2.3 Speech Processing On The Mobile Device Vs Remote Speech Processing

As already mentioned, speech recognition can require a great deal of computing power. A large vocabulary combined with top quality recognition often mean that a mobile device is not powerful enough for the task. If this is the case, captured speech must be sent to a remote computer, where it is processed into text and the result sent back to the device. This is expensive in terms of the volume of data transmitted to the host, which introduces added latency to the transaction. It
also limits the scalability of the solution, since remote speech processing computer must be able to support the requirements of many clients. A third feature of remote speech processing which is not related to the speech processing itself, but which is still significant, is that terminals must be within radio coverage virtually all of the time, and certainly at all points where speech input is required.

By keeping the vocabulary small and adopting a speech processing methodology that can be run on client devices, all of the disadvantages of remote speech processing can be avoided. Data transmissions are limited to small items of processed data rather than large speech samples. Response times do not include the latency involved with network connections and interfacing with remote systems, and scalability is no longer an issue because each new terminal connected to the network performs its own speech processing. Some degree of network independence can be achieved if applications are designed as thick clients possessing a high degree of local logic, and communicate with the host system infrequently.

All currently available commercial speech systems process text-to-speech on the client device.

<table>
<thead>
<tr>
<th>The Microlise approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>The types of supply chain activities which lend themselves to voice technology are simple, repetitive tasks involving the collection of largely numeric data. Most user input consists of either numeric data (product codes, check digits, URNs etc) or simple commands to control the application. This means that most tasks can be accomplished with a vocabulary of less than 40 words, which makes a trained system a viable option. We also believe that the additional benefits a trained system bestows, such as the high recognition rate even in noisy environments, and the built-in scalability, make trained systems best technology for supply chain tasks. We have selected a Voice technology which supports continuous speech for a small, speaker dependent vocabulary, and provides a customisable text-to-speech output component.</td>
</tr>
</tbody>
</table>

A summary of the features of speech processing methods can be found in Appendix A.

2.2.4 So What Is Voice Technology Good At?

A key to using any technology effectively is to determine where it can be deployed effectively. Voice technology is suited to a range of supply chain activities, where manual handling tasks and difficult environments are the norm. Below we list the characteristics of tasks with are well suited to being controlled by voice enabled systems.

- Repetitive tasks where users are directed by a host system.
- Data collection tasks that follow a regular process flow.
- Operation in environments where keyboard operation is not easily accomplished (eg coldstore).
- Tasks with intensive manual handling (no need to keep picking-up/ putting-down terminal).
- Tasks requiring user interaction whilst on-board moving vehicles, or where controlling a vehicle forms part of the task.

2.2.5 ...And What Is It Not So Good At?

Voice enabled applications can rapidly get unmanageably complex when applied to tasks to which they are ill-suited. This is largely due to the fact that the user is presented with information in a sequential manner, one after another, whereas applications with a visual interface can display many items of data concurrently, relying on the user’s cognitive skills to navigate to the required items. For this reason, large menus and complex list selection functionality is difficult to implement successfully.

Here are some other features of applications which make them unsuitable for voice interfaces.
Data collection tasks requiring a large specialised vocabulary (although in practice these can often be simplified enough to be implemented easily).

Tasks requiring complex or long-winded instructions (by the end of an instruction, the user may have forgotten the beginning of it)

Tasks with complex control flow options (users can lose their bearings such that they do not know where in the task they are!).

3 How Can Voice Technology Be Used Within The Supply Chain?

We have seen that whilst voice technology offers some exciting opportunities for creating innovative and efficient applications, but how can we apply the technology to real-world solutions?

It will be readily apparent that many warehouse and supply chain operations meet the criteria we have defined for being well suited to voice technology. Much of the work done is primarily concerned with the movement of stock, either by being handballed or moved using stock moving equipment under the control of an operative. In both of these situations, anything system interaction requiring a handheld device will be seen to slow down the task. Whilst in control of a vehicle, any use of a ‘screen and keyboard’ device is likely to prove hazardous.

In certain environments, such as a coldstore or outside, where gloves are worn, the practicality of voice operated equipment becomes obvious.

Bearing in mind the types of task that are most suited to voice technology, the biggest gains are likely to be seen in the following areas:

- Case & Item picking
- Bulk picking (by vehicle)
- Putaway and replenishment
- Perpetual inventory and Stock Checking
- Marshalling, assembly and vehicle loading
- Goods receipt
- Cross docking

In each of the above tasks, the operative can be given instructions and feedback collected data without interrupting the task in any way.

4 Simplified Architecture Of A Voice Enabled Supply Chain Solution

There will be few, if any, voice-enabled systems that do not require interfaces to external business systems. A typical application will include an ERP or WMS as ‘source’ and ‘sink’ for data captured on the voice terminal. Given the wide range of host systems that could be called upon to drive a voice terminal, a reliable and flexible method of data transmission is a key component of the overall solution. An appropriate way to achieve the required level of integration is by using a general purpose messaging (examples of which are JMS and IBM’s MQSeries). The benefits of messaging systems are discussed extensively elsewhere, but for this kind of application, they provide us with a standard mechanism to integrate our voice-enabled user interfaces to a variety of business applications, irrespective of the location or implementation specifics of either system.

The following simplified diagram shows how the major hardware components are integrated to support two-way communication between voice terminals and back-end business systems.
**Wireless Voice Terminals.** These devices run a thick-client application providing the user functionality and speech processing. Where interaction is required with a host system: for instance where a new piece of directed work is required or collected data is ready for transmission back to the host, the terminal communicates via its own proprietary protocol, over a standard radio network, to a listener on the Voice Terminal Server. The client application is capable of operating autonomously in case of temporary radio failure. All speech processing takes place on the terminal, and so network traffic comprises only that data which is required by the host system, making the bandwidth requirement of a voice terminal similar to that of a conventional RDT.

**Voice Terminal Server (VTS).** This provides the bridge between the Voice Terminals and the message routing network. A proprietary Terminal Listener application handles communications to and from the Voice Terminals, and a further piece of software forms a bridge between terminals and the messaging system. The bridge software developed by Microlise is called OPUS Voice Connect.

OPUS Voice Connect is Microlise’s own mechanism for interfacing between voice terminals and host systems. It is a flexible data manager for routing messages between voice terminals and host systems, and provides comprehensive statistical and monitoring information regarding terminal activity. Additionally, a separate terminal manager application may be resident on the VTS to monitor terminal usage and control software downloads etc.

**Message Routing Network.** The introduction of a messaging service in preference to piece of bespoke point-to-point interface code allows a high degree of decoupling between Voice terminal application and the back-end business system. It also provides a guaranteed once-only delivery mechanism and built-in scalability.

All data flow requirements between the systems are identified up front, and appropriate message definitions are agreed to support these requirements. The choice of message service is then an arbitrary one provided that both the terminal listener and the host system can access it. Microlise use a Java Messaging Service (JMS) implementation called SwiftMQ.

**Host System.** This is a business system or combination of systems which are capable of supporting the Voice Terminal application. In order to integrate with the messaging system, a bridge component may be required between the system’s existing interfaces and the message service.
5 Summary

Voice technology is beginning to attract a lot of attention, both from system developers and potential users of the technology. The supply chain is seen as a prime target for voice enabled systems for several reasons.

- The nature of the sector is very competitive, requiring players to 'keep ahead of the crowd'.
- Many of the activities carried out in a supply chain environment are repetitive manual handling tasks which are ideally suited to being managed by voice enabled applications.
- Many businesses in the supply chain have already made significant investment in IT infrastructure, making the introduction of voice technology a simple step.

Several major suppliers have emerged and have demonstrated reliable and cost-effective voice technologies, and solution providers such as Microlise can now offer integrated supply chain systems with voice control.

Microlise believe that voice technology will rapidly gain an important role in supply chain solutions, bringing simpler ways for interacting with new and existing business applications, and greater operator efficiency.
## Appendix A – Summary of the features of speech recognition methods

<table>
<thead>
<tr>
<th></th>
<th>Microlise Approach</th>
<th>Alternative Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Where is speech processed?</strong></td>
<td><strong>On The Voice Terminal</strong>&lt;br&gt;• Speech recognition and speech synthesis performed on voice terminal.&lt;br&gt;• Only data is transmitted between voice terminal and central server.&lt;br&gt;• Terminal can buffer up instructions so that operation can continue when no RF coverage.&lt;br&gt;• No latency.</td>
<td><strong>At The Central Server</strong>&lt;br&gt;• Speech recognition and synthesis performed at central server.&lt;br&gt;• Compressed speech transmitted between voice terminal and central server (much higher volume than data only)&lt;br&gt;• Terminal cannot operate without RF coverage.&lt;br&gt;• Latency depends on load on RF network and central server.</td>
</tr>
<tr>
<td><strong>What method is used?</strong></td>
<td><strong>Speaker Dependent</strong>&lt;br&gt;• Individual profile for each user. Requires training beforehand (approx. 15 minutes). Profile downloaded to voice terminal when operator logs on (takes a few seconds).&lt;br&gt;• Vocabulary limited to only those words required for application.&lt;br&gt;• Independent of language, accent or dialect.</td>
<td><strong>Speaker Independent</strong>&lt;br&gt;• Uses natural speech recognition.&lt;br&gt;• Independent of user, no training needed, no user profiles.&lt;br&gt;• Practically unlimited vocabulary.&lt;br&gt;• Needs more computing power, hence deployment on central server.&lt;br&gt;• Language specific – currently available for English, French, Spanish, etc. Claimed to handle most regional accents.</td>
</tr>
</tbody>
</table>
### 6 Appendix B – Specification of the Talkman2™ terminal

<table>
<thead>
<tr>
<th></th>
<th>Talkman T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor</td>
<td>Voccollect</td>
</tr>
<tr>
<td>Dimensions</td>
<td>165 x 86 x 38mm</td>
</tr>
<tr>
<td>Weight (with std. battery)</td>
<td>440g</td>
</tr>
<tr>
<td>Battery (standard)</td>
<td>Lithium Ion 1.5Ah</td>
</tr>
<tr>
<td>Operating temperature</td>
<td>-29°C to 50°C</td>
</tr>
<tr>
<td>RF communication</td>
<td>IEEE 802.11b</td>
</tr>
<tr>
<td>Serial ports</td>
<td>1</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel Strong-Arm SA-1110, 200MHz</td>
</tr>
<tr>
<td>Memory/storage</td>
<td>16MB Flash, 32MB RAM</td>
</tr>
<tr>
<td>Operating system</td>
<td>Windows CE</td>
</tr>
<tr>
<td>Speech Recognition technology</td>
<td>Bluestreak</td>
</tr>
<tr>
<td>Warranty</td>
<td>1 year</td>
</tr>
</tbody>
</table>