



# Speech Recognition

---

## Giving a Voice to Business

**WHITE PAPER**

**Release 2  
February 2001**

**Copyright**

Copyright © 2000 Mitel Networks Corporation. This document is unpublished and the foregoing notice is affixed to protect Mitel Networks Corporation in the event of inadvertent publication.

All rights reserved. No part of this document may be reproduced in any form, including photocopying or transmission electronically to any computer, without prior written consent of Mitel Networks Corporation. The information contained in this document is confidential and proprietary to Mitel Networks Corporation and may not be used or disclosed except as expressly authorized in writing by Mitel Networks Corporation.

**Trademarks**

Product names mentioned in this document may be trademarks or registered trademarks of their respective companies and are hereby acknowledged.

**Table of Contents**

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>2</b>	<b>SPEECH RECOGNITION TECHNOLOGY – WHAT IS ALL THIS TALK ABOUT?</b>	<b>2</b>
<b>3</b>	<b>HOW DOES IT WORK?</b>	<b>3</b>
3.1	<b>Conversion of Speech to Action</b>	<b>3</b>
3.2	<b>Methodologies</b>	<b>3</b>
<b>4</b>	<b>THE COMPONENTS OF SPEECH RECOGNITION</b>	<b>5</b>
4.1	<b>Accuracy</b>	<b>5</b>
4.1.1	Input Devices	5
4.1.2	Speaker-Dependent Versus Speaker-Independent	5
4.2	<b>Usability and User-Friendliness</b>	<b>6</b>
4.2.1	Continuous vs. Discrete Input	6
4.2.2	Grammars and Vocabularies	6
4.2.3	Natural Language Understanding (NLU)	7
4.2.4	Barge-in Support	7
4.3	<b>User Interface Design</b>	<b>7</b>
4.3.3.1	<i>Multiple Language Support</i>	8
4.3.3.2	<i>Voice Verification</i>	8
4.3.3.3	<i>Text-to-Speech (TTS)</i>	9
<b>5</b>	<b>CUSTOMISED SPEECH APPLICATIONS</b>	<b>10</b>
5.1	<b>Grammar tools</b>	<b>10</b>
5.2	<b>Dialog Design Tools</b>	<b>10</b>
5.2.1	Reusable Dialog Components	10
5.3	<b>Testing Tools</b>	<b>11</b>
5.4	<b>Recording Options</b>	<b>11</b>
<b>6</b>	<b>THE MARKET DRIVERS FOR SPEECH</b>	<b>12</b>
6.1	<b>The Benefits of Speech as a Technology Interface</b>	<b>12</b>
6.1.1	The User Experience	13
<b>7</b>	<b>THE APPLICATION OF SPEECH RECOGNITION IN TELEPHONY MARKETS</b>	<b>14</b>
7.1	<b>Voice Processing</b>	<b>14</b>
7.1.1	Voice and Unified Messaging	15
7.1.2	Virtual Assistants	15
7.2	<b>Call Centers</b>	<b>15</b>
7.2.1	Interactive Voice Response (IVR)	15
7.2.2	CTI/Call Centers	15
7.3	<b>The Internet</b>	<b>16</b>
7.4	<b>Return on Investment</b>	<b>16</b>
<b>8</b>	<b>SUMMARY</b>	<b>17</b>



## **1 Introduction**

Hal is here. The computer that ran the space ship in 2001 is no longer a myth, although the voice of Hal may or may not have an accent, or be male, and definitely has more personality. The use of speech recognition, whereby a computer talks and/or interacts with a person, is proliferating into applications far beyond those seen in sci-fi movies. In fact, it is becoming commonplace to encounter silicon speech personalities during our interactions with businesses or even self-chosen interactions through the purchase of business and consumer products that contain the technology.

So just what is speech recognition? This paper provides an introduction to speech technologies and is intended to help with the understanding of these technologies and their practical applications in the consumer and business market.

## **2 Speech Recognition Technology – What is all this Talk About?**

Simply speaking, Automatic Speech Recognition (ASR) is the technology that allows a machine to understand human speech. Alternately referred to as either ASR or speech recognition, the technology takes human speech input, digitizes it, and converts it into a machine-readable string of text. A technology component called a recognizer then manipulates the text into a form that the recognizer uses to identify what the speaker said. This is no simple task, as it requires a thorough linguistic understanding of speech combined with statistical analysis, plus a healthy dose of electrical engineering, and digital signal processing. Fortunately, with more than three decades of public and private R&D behind us, we now have good quality, commercially viable speech recognition solutions available.

The goals of early speech researchers, while altruistic, also had business components, as well as novelty and entertainment value. Researchers found that there were numerous needs requiring command and control of devices and applications in a hands-free environment, either out of safety concerns or because of physical limitations of the user. Potential applications include using speech recognition to control a computer, dictate words and paragraphs, or control an action such as dropping fire retardant from a helicopter, when your hands are busy with flight control. Both novelty and convenience were motivators to create the ability to control a home's appliances, HVAC and security systems from a distance using the telephone. Between these two extremes was research on the practical business applications that could be enhanced by speech. The resulting development of speech technologies has evolved into three main areas: PC software, such as dictation or command and control applications; over the phone applications; and embedded systems applications.

### **3 How does it Work?**

It takes many components to move from human interaction with a computer to a desired outcome. This interaction can take place in a number of ways however, for the purpose of clarity, we will use the example of a person interacting with a speech application over the phone for the majority of this paper. Therefore, in telephony applications, the user will be referred to as the caller.

#### **3.1 Conversion of Speech to Action**

Without any refinements to the technology, a basic speech recognition interaction with a caller develops as follows:

##### **Step 1 – Caller Input**

A caller states a phrase or sentence using an input device such as a telephone. The system captures the speech in the form of an acoustic signal.

##### **Step 2 - Digitization**

The system converts the words from an analog to a digital signal it can understand. This step converts the input signal into something closely approximating the acoustic properties of the human ear.

##### **Step 3 – Phonetic Breakdown**

The speech recognition software breaks the digitally converted words down into the basic components of speech. These components are the basic building blocks of speech and correspond to consonant or vowel sounds.

##### **Step 4 – Statistical Modelling**

The system then tries to match these sounds to their phonetic representations.

##### **Step 5 - Matching**

The speech recognition application tries to map the possible phonetic representations to words or phrases defined in the grammar of that application. For example, in a banking application a caller might say, “transfer” a specific amount of money and the application would find a match. But if they said, “send” an amount of money, the application might not find a match if the word “send” was not previously defined as an acceptable synonym for “transfer.”

#### **3.2 Methodologies**

To make speech recognition work, the software has to take into account the acoustics of the words being spoken, the vocabulary being used, and the language model. The language model contains all the statistical information about usage of the vocabulary so that the recognizer can make a reasonable guess at what is being said.

Recognizers can be either hardware or software-based and various methodologies exist to transform the digitized text string into words. Techniques vary from recognizer to recognizer with advantages and disadvantages to each.

There are different methodologies used in recognizer models. One of the most common is the Hidden Markov Model (HMM), on which the majority of phonetic, grammar-based recognizers on the market are founded. Other models include segmental modeling, host and/or DSP-based models, or telephony versus microphone input models. Many small vocabulary applications still make use of whole-word templates rather than phonetic based algorithms. The methodology greatly depends upon the environment in which the recognizer will run and the goal of the application. For example, a limited vocabulary application for command and control of a PC software package initiated by a single user using a microphone has different requirements from a commercial over-the-phone stock brokerage application with a large vocabulary, vast number of speakers, and diverse variances in input environments.

## **4 The Components of Speech Recognition**

The above application flow is very simplistic in that it focuses on a single utterance. In a lab environment or over the phone in ideal conditions it is easy to make a simple call flow work. However, under normal conditions there are many factors that can make it difficult for the recognizer to be accurate. The type of input device, ambient noise, differences in caller accent, tone, and gender, as well as regional differences in terminology are all examples of challenges that the speech developer faces. Further, the ultimate goal for speech recognition is to bring a human touch to the interaction by creating a natural dialog with the caller. “How may I help you?” applications, in which there is unconstrained caller input by way of an interactive dialog with the computer, is the ultimate marker of speech recognition success. This conversational speech technology, in which a user has a dialog with the application, increases the speed of the interaction as well as comfort and satisfaction by the user. Therefore, to really make speech recognition dramatic as well as accurate, the speech recognition industry has had to do a lot of development on different aspects of recognition.

### **4.1 Accuracy**

Although we are long past judging the viability of an application on accuracy alone, the original true measure of a speech recognition application was in the accuracy rate of the recognizer. The accuracy rate refers to the percentage of time that a recognizer will accurately recognize what a caller said.

Critical elements used to bolster accuracy rates are the ways in which an application is designed to handle error conditions. There are a number of ways that an application can be designed to handle errors. For example, with “N-Best” recognition, when the system doesn’t understand the utterance, it searches for alternatives and presents them back to the caller. In response to ambiguous input a caller might hear “I did not quite understand what you said. Did you say ‘Smith?’” Alternatively, the application might reply with “I did not understand what you just said. Could you please repeat?” or simply repeat the question a second time. Other tactics, such as rephrasing the question, prompting the caller with possible responses, or asking several simpler questions, can also be employed.

#### **4.1.1 Input Devices**

The type of input device employed by the user greatly affects the accuracy of the recognition. Contrast the difference between a person using a microphone to talk to a dictation package on the PC and a business traveler calling into a travel application while using a cell phone in an airport. Even the difference in transmission channel or telephone handsets can affect the input to the recognizer. Therefore, the conditions under which the application is being used must be considered so that the application is optimized for different conditions.

#### **4.1.2 Speaker-Dependent Versus Speaker-Independent**

Another factor that can seriously affect recognition rates is whether the system is tuned to individual users. Speaker-dependent applications are those in which the application “self-tunes” to individual speakers. Employing what is termed speaker verification or voice verification algorithms, the system is trained to recognise individual users of the application. In most dictation software, for example, a user spends time training the system to recognise his/her voice by reading a series of passages that cover a wide range of phonetic representations. In telephony or other applications, the application uses a smaller set of training criteria and makes a voice print of the caller to verify when the caller enters the system. For example, when using a telephone credit card application, the user would be voice-printed by the system upon enrolment. The system would then ask for a name or password at the beginning of subsequent calls to identify the user and give them access to the application.

Speaker-independent applications are those that are open to a larger, variable pool of subscribers. In this case, the system needs to be able to recognise input that takes into account large variances in the way that words are spoken. Speakers may be male or female, young or old, with high or low-pitched voices. They may have foreign accents, regional dialectical differences, or speech impairments such as a cold, stuttering or other speech impediment. In order to counteract these challenges, a grammar is defined for the application, which limits the number of variables that can be spoken in order to elicit a positive response from the application.

## **4.2 Usability and User-Friendliness**

As important as accuracy is the usability and user-friendliness of the application. Obviously, the system to be accurate or users won't use it; however, there are many other things that developers have created to enhance user-friendliness of these applications. These include the following:

### **4.2.1 Continuous vs. Discrete Input**

Most early speech recognition applications were capable of recognizing discrete words or digits, so that a user had to say one word at a time for the recognizer to work. In a telephone credit card application, for example, a user would pause and say one number, then wait for a beep, then state another number, progressing until the entire credit card number was entered. The current standard goes beyond this with continuous speech recognition, allowing the user to speak input in an unbroken string or sentence. Therefore, a credit card number could be spoken as an uninterrupted utterance. While this may seem simple, developers had to take into account the various ways that a user might enunciate input, such as saying "zero" versus "oh" in a credit card number, or the parsing of numbers in a digit string, taking into account natural language pauses. The algorithms had to factor in every possible word in every possible position in the sentence, as well as alignments in the word boundaries of the speech signal.

### **4.2.2 Grammars and Vocabularies**

A crucial aspect of the application is the vocabulary and grammar it employs.

A vocabulary is the list of words used in the application. It is often a reusable component that can be modified to fit a number of customer applications. As such, much development work has been done in defining, building and testing specialized vocabularies for various vertical markets, such as health care, finance, or hospitality. One of the most highly touted vocabularies is that of stock brokerage applications. These vocabularies, refined by a number of vendors, contain all the words and phrases that a brokerage customer would normally speak to a customer service representative (CSR) when placing a trade, getting a quote, or making some other type of transaction. They include all the names and common abbreviations (IBM for International Business Machines, for example) for the commodities on different stock exchanges.

A grammar is the compilation of words and phrases (from the vocabulary) comprising the expected range of input and output for the application. It defines what the application is listening for when interacting with the caller, and uses different linguistic and statistical models to put boundaries around the application. It might be a simple grammar of 50 words or one that contains thousands of words, names or addresses. A grammar must also take into account any specialized abbreviations, acronyms, etc. Since a grammar defines the boundaries of the application, the recognizer listens for the components of the grammar and handles extraneous input from the caller with exception rules or error handling conditions. In a brokerage application if the recognizer heard the word "filibuster," for example, it wouldn't recognize the word and would re-prompt the caller.

#### 4.2.2.1 *Dynamic Vocabularies*

Dynamic vocabularies are exceptions to developer-owned and -run vocabularies. A dynamic vocabulary allows the recognizer to accept input from the caller not included in the original vocabulary. This is particularly useful in an auto-attendant application or directory service program in which a caller speaks an unknown name or address.

#### 4.2.2.2 *Vocabulary Size*

Vocabulary size is related to the accuracy, speed, and usability of the application. In the early days of speech recognition, callers were limited by the capabilities of a grammar or vocabulary due to technical limitations and the cost of processing power. However, with the advent of faster, cheaper and more powerful systems this has been alleviated. Today, applications such as the huge operator service vocabularies support thousands of phrases.

#### 4.2.3 Natural Language Understanding (NLU)

Natural language understanding (NLU) is related to both accuracy and usability. Alternately called natural language support or conversational speech technology, NLU adds a human component to speech recognition by trying to understand the meaning of what the person said, rather than just the actual words spoken. It is the driver behind the “Holy Grail” of the speech industry, that of unconstrained speech in an application. The benefits of employing NLU are dramatic in that it can increase both the speed and the accuracy of the application. More importantly, however, it gives the caller a better user experience because the application more closely mimics human interaction. Instead of the caller being forced to follow prompts such as, “Please speak the name of the airline for which you want the flight number.” A speech recognition system might offer the less constraining, “How can I help you?” The caller could respond in various ways, such as, “Please give me the flight number for United.... “ or “Can I have the flight number for United?”

NLU presents a number of linguistic challenges for developers of speech recognition. They not only have to deal with the content of what the caller says, but the context as well. As an example, if a caller mentions books, he could be talking about reading books, doing accounting books, or “booking” a commodities trade.

#### 4.2.4 Barge-in Support

Barge-in support allows the user to interrupt or override a spoken prompt. The recognizer keeps a channel open, “listening” for the caller while speaking the next prompt or providing feedback. This function greatly enhances the usability of the application by allowing power users of the system to get to what they want efficiently through interrupting the system to correct either their input or the system’s output, or by bypassing redundant prompts.

### **4.3 User Interface Design**

Apart from an appropriate vocabulary and the technology itself, the most critical component in the success of the application is the user interface, known as call flow or dialog. The beauty of applying speech recognition to an application is to transform the user experience from automated input to something more closely resembling that of interaction with a human. Therefore, if the application is poorly designed, the system has failed. Studies have shown that a properly designed user interface can reduce hang-ups and the occurrence of a caller “zeroing out” to a CSR, while increasing transaction rates. However, if the system doesn’t handle error conditions properly, doesn’t provide the caller with the correct options, or is too difficult to navigate, the reverse happens. A good user interface will simulate a personality and will encourage both caller interaction and repeat calls.

Setting a voice and style for the application is critical for its success. Whether it is a replacement for a simple touch-tone or a more complex, multi-layered application, the design must encourage interaction with the caller to obtain the proper response while being socially appropriate for the type of caller using it. It is also critical that the application's "personality" has characteristics that represent the company style and brand. The types of dialogs that are used in designing an application drive the "feel" of the call flow. For example, an application designed for dealing with damaged merchandise would have a very different "feel" to it than one that is processing inquiries about Broadway show tickets.

#### 4.3.1 Directed Dialogs

In a directed dialog, also known as system initiated dialog, the caller is led through a fixed set of prompts requiring specific input. The dialog is in a step-by-step fashion. The most basic examples are similar to simple touch-tone IVR applications such as greeting the caller, asking for a social security number or other identifier, and leading the caller through questions requiring yes/no answers or numerical input. The application limits and directs caller input, occasionally needing to re-prompt the caller to refine the input before moving on to the next stage of the dialog.

#### 4.3.2 User Driven Dialogs

User-driven dialogs provide more flexibility in the user interface as they permit variations in input, so that the caller feels as though they are driving the application. After greeting the caller, the system might say, "How may I direct your call?" or "How may I help you?" leaving an open ended question for the caller to answer as they would in a live interaction with a CSR. User-driven dialogs employ natural language understanding software.

#### 4.3.3 Mixed-Initiative Dialogs

In a mixed-initiative dialog, the system uses a combination of directed and user-driven options to achieve balance between user-friendliness and speed. For example, in a travel reservation system the application might begin the call with "How may I help you?" and let the caller answer as they wish, while switching to directed input when asking for credit card information.

##### 4.3.3.1 *Multiple Language Support*

It is increasingly popular to provide multilingual support, just as some call center and IVR applications do. For example, in Canada, where there is a high percentage of French speakers, or California, where Spanish is prevalent, it makes practical and business sense to set up the application with bilingual prompts, similar to the expectations of a live agent. For example, the name Richard St. Jean could be pronounced as "Ritchart Saint Jeeen" or "Reeshard Sain Jon." This requires support for two grammars on the same system, or two systems networked together. In some cases, applications are set up to allow real-time switching between two grammars or two languages. Applications can also be set up to recognize the caller's language and automatically switch to it for the duration of the call.

##### 4.3.3.2 *Voice Verification*

Voice verification is speech recognition software designed to recognize a specific user of the system. To do this, the system is trained to recognize a user by making a voiceprint of a set of utterances representing that person's unique speech patterns. In order to ensure accuracy of the system, and to reduce instances of false acceptances or false rejections, the "training" process for capturing a voiceprint can be as thorough as the developer deems necessary.

#### 4.3.3.3 *Text-to-Speech (TTS)*

A precursor to speech recognition and a sister technology, text-to-speech (TTS) is the technology whereby a text string is converted into synthesized speech. In a mixed or complex application, TTS might be used to read a large quantity of text back to the caller. As a supplement to speech recognition, it can be used, for example, to avoid the cost prohibitiveness and inconvenience of pre-recording large chunks of text with human speech (either because of time or PC memory constraints). It is also used when there is a variable quantity of text that might be only intermittently required or constantly changing. Certainly this is true in unified messaging applications, where voice navigation is relatively static but the content of e-mails is constantly and unpredictably changing.

## **5 Customised Speech Applications**

Although many turnkey speech recognition applications are being sold, others require varying degrees of development and customization. This customization ranges from adding names to a list to designing a telephone or even to web-based applications using speech. In addition, the industry as a whole is creating standards to facilitate such things as how speech should be incorporated into an application or how to speech-enable web development so that applications can access information from intranets and the Internet. At the same time, just as the technology itself has been refined, so have the tools to develop an application. From the developers' perspective, there have been great improvements with the ease with which a speech recognition application can be developed and deployed. This is particularly true for speech-enabled call center and IVR applications because of the parallel development of GUI development tools being used in those industries. On the voice processing side, improvements have been made in simplifying the process of creating and updating directory entries as well.

### **5.1 Grammar tools**

Grammar tools are those that a developer uses to build the grammars for different speech applications. These tools have evolved in response to the difficulties in writing separate grammars for different applications. Similar to advances made in complementary technologies such as call centers and IVR, these tools are typically intuitive, GUI-based tools that hide the details of grammar syntax from the user. Now a developer can add the words and phrases that will be used in the grammar by typing phrases into a template (for example, into an Excel spreadsheet). This acts as a visual aid to the developer and greatly speeds up the building of the grammar. The developer inputs words and phrases to build the grammar and the tool generates the proper output, including any error codes.

### **5.2 Dialog Design Tools**

Many aspects of a design can make or break an application. The conveyance of tone and attitude in prompts, as well as their length and verbosity, can affect a user's willingness to proceed. For example, just as a lengthy touch-tone menu puts off callers, so do long-winded voice prompts. It is imperative that the dialog be succinct, yet instructional enough to elicit the proper input from the caller.

Beyond the user friendliness and speech ergonomics of the dialog are the tools that enable a developer to quickly design an application. These tools allow the developer to input the prompts and create the application flow of the dialog. Current state-of-the-art technology for complex applications uses graphical interfaces with drag and drop capability for building the call flow. Additionally, most applications will allow the developer to write specialized subroutines using an application-programming interface (API) to popular programming languages such as C++, Visual Basic or JAVA.

#### **5.2.1 Reusable Dialog Components**

Speech technology vendors now create libraries of reusable dialog components known as dialog modules or speech objects, which are application building blocks or intelligent objects that encapsulate frequently used subsets of applications. In an IVR application, the call flow for greeting the caller and accepting a user ID number would be one example. Other modules commonly developed include collecting credit card numbers or addresses, voice verification of the caller, or confirming and correcting input.

### **5.3 Testing Tools**

Testing and tuning tools allow the developer to test the application in the lab before going “live” with real customers. Many of these tools simulate the application and allow for rapid prototyping to test it before the majority of application details are written, or they ensure that the application will solicit the proper response from the caller. There have also been many advances in the creation of easier to use, testing tools that help guide the developer to successfully testing and completing an application.

### **5.4 Recording Options**

With many speech applications, the client has the option of recording their own prompts or having them professionally recorded in a studio. In most cases, recording can be done with a microphone or over the phone. The key to the success of any speech application is to give users the ability to record information that would otherwise have to be expressed using TTS. For example, even though auto-attendant vocabularies typically contain thousands of names, there is always the unusual last name or nickname that is new, so providing the administrator with the tools to properly record database exceptions is a must.

## **6 The Market Drivers for Speech**

The spoken word is the most natural user interface in the world. Most of us use it all the time. Speech technologies have grown beyond the lab into mainstream business and consumer applications. Speech recognition technology is being deployed in four main business areas:

- Dictation
- Telephony
- Consumer applications
- Automotive applications

The dictation market currently comprises consumer software packages such as Microsoft Word and e-mail applications, and business dictation products for vertical markets such as the health care and legal industries, for everything from filling out forms to commenting on patient charts.

In the telephony arena, the application of speech recognition has enhanced call and voice processing applications, call centers, service provider offerings and personal productivity software.

In the area of consumer applications, speech recognition has been used in personal dictation products, personal telephony applications such as name-dialing from personal lists, in automated residential applications, and in consumer games and toys. Another growing market segment is in the embedded systems market, with speech being used on Digital Signal Processing (DSP) chips for the automotive industry, for everything from hands-free dialing to mobile access to the GSP (Global Satellite Positioning) system applications and the Internet.

### **6.1 The Benefits of Speech as a Technology Interface**

In the past two decades telephony applications have been of benefit to both businesses and consumers alike. These technologies have resulted in expanded business hours, increased speed of information delivery, and enhanced transactions for both businesses and consumers. Interactive voice response systems, predictive dialers, computer telephony integration software, automatic call distribution and auto-attendants demonstrate some of the ways in which businesses can offload repetitive and menial tasks so that employees can be used for more complex and challenging projects. Similarly, they enable customers to initiate self-service applications 24-hours a day, providing them with access to information and allowing them to make transactions without the aid of a CSR. During normal business hours, these technologies deliver callers to the CSR at the same instant that the customer information being delivered to the CSR's terminal. The addition of speech to these applications significantly enhances the efficiency and the efficacy of the transaction for both the business and its client.

### 6.1.1 The User Experience

How can encounters with automated applications be enhanced with speech? Speech recognition provides access to many people who, for whatever reason, could not otherwise use automated applications. For example, despite the proliferation of touch-tone telephones, not everybody has access to one. With speech recognition, this barrier is lifted. Further, speech simplifies applications by employing voice verification to obsolete long or arcane passwords.

Although there are both tangible and intangible benefits to automated applications, the applications are not universally accepted because they are neither as intuitive nor as user-friendly as the familiar CSR. However, just as the last decade has brought great business advances in the telephony world, speech is being used to advance these applications by providing a natural and comfortable user interface. Through speech the user feels more in control of the flow of the application, and feels that it is more personalized. These factors have led many well-known enterprises to speech-enable access to their businesses. Using speech, consumers are now navigating through flight reservation systems, checking on stocks and making trades in brokerage applications, and making purchases and reviewing order status in retail applications. Previously long and complex touch-tone menus are condensed with speech, putting the user in control rather than forcing them to conform.

User-friendliness and personalization are the biggest factors driving growth in speech applications, but there are other important market drivers as well. For the mobile market, speech enables a service provider to present greater content faster than is available on the screen of a Wireless Application Protocol (WAP) phone. With mobile phones outnumbering computers in the US and worldwide, this opens up a tremendous opportunity to offer information access and revenue generating opportunities to a much wider audience. More importantly, it increases hands-free use of a mobile phone, thereby increasing safety. Finally, the use of speech recognition options such as voice verification increase both the security of applications and the speed with which a user can be authenticated, ultimately increasing user satisfaction.

## **7 The Application of Speech Recognition in Telephony Markets**

There are numerous applications of speech technologies in telephony, ranging from the pedestrian to the complex. Even before the development of large vocabularies, and before we pushed the envelope on recognition rates, early adopters started adding speech to their applications. Operator services aside, a huge market of opportunity was available in businesses already, primarily in the vast installed base of automated applications in call centers, standalone IVR applications and voice messaging systems. Other opportunities were available through applications that could be conceptualized but not realized due to the limitations of previous technologies.

### **7.1 Voice Processing**

Voice processing, which includes auto-attendants, voice messaging and unified messaging, is primed for the addition of speech recognition. Perhaps the earliest commercial applications of speech recognition were in the area of operator services, arising from the availability of resources in telecommunications and the intense need to drive down costs in telco networks. The automation of operator services resulted in millions of dollars being saved. However, it was not just the service providers' side of the market that benefited from the addition of speech, but the enterprise market as well. The creation of speech-recognition driven auto-attendants for the enterprise market brought both reduced costs and expanded coverage for businesses, as an auto-attendant works 24-hours-a-day without pay. By the mid 1990s, thousands of voice messaging systems were using the dial-by-name function of voice messaging systems to transfer internal and external callers to employees. While highly workable, these interfaces were cumbersome compared to the naturalness and ease provided by a voice interface. The change to a voice driven interface not only sped up calls, but reduced errors in reaching the desired party and enabled a greater range of transfer options for callers as well. Today, callers can navigate more easily through menus or barge in without listening to a long list of prompts.

Internally, companies can further reduce expenses by enabling employees to dial-by-name or extension without requiring an internal operator. Not only are human resource costs reduced, there is also no need to print directories on a periodic basis, which can cost into the thousands of dollars for a large enterprise.

Another unheralded benefit to speech recognition applications is fraud reduction. A common feature on many PBXs is DISA (Direct Inward System Access). DISA has long been used to give employees access to the features of the PBX remotely. A user dials into a switch, enters a pass code, and is passed to a dial tone, enabling an external telephone set to access all the network features of the switch including long distance overseas dialing. However, since DISA is an access code, this leaves companies open to fraud if the number gets out to unauthorized users. With voice verification, only the authorized employee can gain access to the switch, thereby eliminating the chance of fraud.

### 7.1.1 Voice and Unified Messaging

With the depth of research done in command and control applications, speech recognition is a natural choice for controlling telephony applications that require navigation. For this reason one of the most obvious targets for deployment is in voice and unified messaging applications. As with the auto-attendant function, the addition of speech recognition allows users to navigate through voice messages without remembering command codes or a series of keystrokes on the telephone or keyboard. It also levels the playing field for user interfaces in that any combination of words can be recognized as a command. For example, a system can be set up to respond to, delete or erase (not just one or the other) making it more flexible for the user. In addition, the Telephone User Interface (TUI) is now flattened as well, eliminating the need for a user to have to go up one tree and down another to access features or functionality. For example, a user could be listening to a message and then realize that they have to change their greeting. Rather than saving the message, working back to the main menu, and going through the user options to change the greeting, for example, they could be in the middle of listening to a message and simply say “change my greeting” to access the necessary function.

Since many voicemail products allow a caller to record their name for the recipient, in addition to dialing by name or speaking commands, a user could also search for a message in the queue by the sender's name. Speech will enable the full use of messaging, particularly as users typically only use 10 percent of the features of a messaging TUI because their inability or unwillingness to negotiate the TUI.

### 7.1.2 Virtual Assistants

An offshoot of the messaging market is that of virtual assistants. Also named personal telephony or unified communications, virtual assistants are speech-recognition driven personal call assistants that allow a user to access and control any number of communication applications from a telephone. Designed with personality in mind, these agents act on behalf of the caller to access and utilize applications such as unified messaging, voice messaging, voice activated dialing to internal and external numbers, integration with Personal Information Manager (PIM) applications, and calendar functions.

## 7.2 Call Centers

There are ample opportunities in the call center market for speech as well. In fact it is a natural enhancement to many applications in a call center including:

### 7.2.1 Interactive Voice Response (IVR)

IVR vendors were one of the earliest adopters of speech technologies, initially due to the limitations of touch-tone input. Touch-tone was a blessing in the automation of many applications, but had severe limitations when it came to some types of input, such as alphanumeric strings of data. Speech eliminated many of the constraints on these applications, and thousands of simple applications were upgraded to include speech recognition. The earliest applications were simple “one through ten” and “yes/no” applications, but with the advent of natural language processing and greater processing power, any number of IVR applications have now been speech enabled.

### 7.2.2 CTI/Call Centers

A high percentage of call centers employ IVR systems at the front end in order to collect caller data, offload CSRs, and reduce call hold times. The addition of speech recognition applications further enhances these operations by speeding up access to information and making the user experience better. This, in turn, reduces the instance of callers hanging up when they get stuck in queue, allows them more self-service options, and increases their willingness to use the application again. Speech recognition can be used to gather the same type of information from callers as would an IVR, but more quickly, and with greater flexibility as to the type of information being gathered, because callers aren't limited to entering input via touch-tone.

### **7.3 The Internet**

Finally, one of the most exciting emerging opportunities is the speech empowerment of the Internet. The World Wide Web (WWW) is a vast collection of business and personal sites linked together in a spider web of network connections. Recently, it was reported that all but one or two of the Fortune 1000 companies had a presence on the web. For a number of years the previously mentioned torchbearers of customer service applications, the IVR and call center vendors, have been web-enabling their applications to provide greater access to information and transactions for customers. Speech is a natural fit in this scenario. In addition to call center and IVR applications, voice portals (just like web portals) are appearing everywhere, allowing callers to “voice surf” the web over the phone, without benefit of a computer.

This is occurring because of the development of a number of emerging standards including the most promising one to date: Voice eXtensible Mark-up Language (VoiceXML), which is an XML-based mark-up language that defines a spoken dialog (voice page) just as HTML defines a graphical web page. This and other developing standards, such as VXML, will eventually allow the interconnection of various “speech sites” to form a speech web, the way that individual graphic sites have made up the WWW.

Once these emerging technologies are implemented, anyone will be able to “surf” a web site without a computer, bringing the benefits of the Internet to anyone with a telephone. And, since so many people have mobile telephones, this only adds power to applications because people can get information from the Internet quickly without the limitations of a small screen of a WAP telephone, and without the hazards of with looking at a screen while driving.

### **7.4 Return on Investment**

As a complementary technology, there is clear business value to implementing speech recognition technology. By changing the caller experience, a business can:

- Provide extended business hours, which increases customer satisfaction and loyalty while opening the doors for new customers.
- Provide self-service applications during and after business hours, thereby increasing revenues.
- Reduce call times and telephone costs by employing voice verification to identify users before they get to a CSR.
- Reduce hold times in call centers by gathering more information from a caller before they get to a CSR.
- Automate applications and provide new services that previously couldn't be done with simple touch-tone.
- Increase e-commerce opportunities by providing access from mobile devices, not just from PCs.
- Increase the number of calls a call center can handle.
- Increase security and decrease fraud costs through voice verification.
- Reduce queue time by allowing users to get their own information.
- Reduce “port usage” of ports dedicated to an IVR system by speeding up caller interaction.

## **8 Summary**

You now have a voice in business that can help you. With readily available Pentium and RISC processors we now have the computing power to support large scale, complex, natural language speech applications at an affordable price, and we can upgrade simpler and smaller applications currently powered using DTMF input.

Speech recognition, once seen as the technology of the future, is today's reality. And it works.

## About Mitel Networks Corporation

Mitel Networks is a leading-edge provider of next-generation IP telephony solutions. The company creates advanced communication solutions and applications in the areas of speech recognition, wireless mobility, unified messaging, and customer interaction solutions. Through direct channels and strategic technology partnerships, Mitel Networks currently serves the education, hospitality, healthcare, and government markets, providing advanced communications solutions.

Mitel Networks is a market-leader in voice and data convergence. The foundation for Mitel Networks' IP-based platforms delivers the power of the Internet to the voice desktop. The company's successful integration of voice and data infrastructures, with patented dual-bus architecture in its Ipera IP platforms, lets communication systems easily accommodate IP and digital phones. Ipera platforms allows enterprise customers to seamlessly implement and/or upgrade to a Voice-over-IP structure without sacrificing any of the features or functionality of the traditional legacy PBX. In addition, clients also benefit from a broad range of IP-enabled applications.

At Mitel we are proving to enterprise that the migration to the IP world means improved features and services compounded with the scalability to grow with their changing needs.

Mitel Networks is headquartered in Ottawa, Canada, home to the company's product development, marketing, finance and administration functions. Regional operations are located in Herndon, Virginia (US Sales), Caldicot, Wales (European headquarters); and Hong Kong (Far East operations). Mitel Networks operates 71 regional facilities in the U.S., Canada, the UK, Europe, and the Far East. Manufacturing facilities are located in Canada, the UK, and Sweden.



**www.mitel.com 1-800-MITEL-SX**

**Mitel Corporation**  
350 Legget Drive  
Kanata, Ontario  
K2K 2W7 Canada  
(613) 592-2122

**Mitel Inc.**  
205 Van Buren Street  
Suite 400  
Herndon, VA  
20170-5336 USA  
(703) 318-7020

**Mitel Telecom**  
Mitel Business Park  
Portskewett,  
Monmouthshire  
NP6 4YR UK  
+44(0)1291430000

THIS DOCUMENT IS PROVIDED TO YOU FOR INFORMATIONAL PURPOSES ONLY. The information furnished in this document is believed by Mitel Networks to be accurate as of the date of its publication, and is subject to change without notice. Mitel Networks assumes no responsibility for any errors or omissions in this document and shall have no obligation to you as a result of having made this document available to you or based upon the information it contains.

© Copyright 2000, Mitel Networks Corporation. All rights reserved. Printed in Canada. PN 50000815 Rev B