ITSC

Information Technology
Support Center

State of Maryland ◆ Mitretek Systems ◆ Lockheed Martin Corporation ◆ University of Maryland

Winner of Case Study Award - 1996 International Summit on Service to the Citizen
Winner of Showcase Award - 1996 Joint Employment and Training Technology Conference

White Paper On Speech Recognition In The SESA Call Center

Final Report

April 2001

Ron Mains, Senior Analyst

Tim Meier, Project Leader

Scott Nainis, Chief Technologist

Henry M. James, Executive Director

Executive Summary

Speech recognition (SR) has developed very rapidly over the last ten years. This has largely been the result of advances in computer processing power combined with correspondingly rapid advances in speech recognition algorithms. As computer processing power increased and became more widely available and less costly, the speech recognition algorithms could be implemented as software programs that could process information quickly enough to be truly usable. The speech recognition systems that are on the market today are the embodiments of these new algorithms that were once the province of the advanced R&D laboratories.

What is often referred to as speech recognition is actually two separate forms of voice input processing, speech recognition and voice recognition. Speech recognition is the conversion of spoken input into stored text, or commands to an application or a computer operating system. Voice recognition, on the other hand, is the identification or verification of an individual's identity using speech as the identifying characteristic.

Speech recognition systems used today fall into two main types: speaker-dependent continuous-speech PC-based systems, and speaker-independent continuous-speech server-based systems. Speaker dependent speech recognition requires "voice-training" by the user(s) prior to their use. Speech-independent systems require no prior input or knowledge about the speaker(s) prior to use. Older systems typically were discrete-speech systems in which each word had to be spoken separately. These systems have since been largely supplanted by the newer continuous-speech recognition systems.

Speech recognition represents a very real option for user input, whether that is a single user system, or one fielded as an adjunct or replacement for an IVR system. Speaker independent systems are generally more costly to purchase than individual speaker-dependent software, but can offer value in saving SESA staff time. Recognition accuracies for both PC and server based systems are now in the 95 percent range and better, with many high-end systems consistently delivering 99 percent accuracy. As good as these systems are, the entire field of speech recognition is still rapidly evolving. As these systems continue to evolve they will function both faster and more accurately, and more languages will be added to their repertoires. Additionally, as with all technology products, the price-performance of these systems will continue to drop as they evolve and penetrate the marketplace further.

Just as speech recognition is bringing competitive and workplace advantages to many companies in the private sector, it can provide distinct workplace and service advantages for SESAs. Examples of some of these benefits are: providing job opportunities for physically handicapped individuals; eliminating or mitigating the risk of repetitive stress injuries such as carpal tunnel syndrome; and automating tasks

that weren't previously feasible to automate. Many of these advantages also carry with them performance and efficiency improvements, helping to reduce the cost of delivery of the required services, and speeding up the time it takes to process initial and continuing claims. Typical cost savings, through reduction in the need for staff telephone contact time, have been found from implemented speech recognition systems to be in the range of from 70 percent to 93 percent per call minute, as compared to the same call being handled by a customer service representative.

This paper addresses the foundation aspects of what speech recognition is, what are the various kinds of speech recognition systems that are available today, and how speech recognition can be used in a SESA call center environment as well as its applicability to other areas of SESAs.

Table of Contents

1 Introduction 1

1.1 Audience 1
1.2 Purpose of this paper 1
1.3 Overview 1

2 Background 3

2.1 What is speech recognition? 3
2.2 Types / Categories of Speech Recognition 5
2.2.1 The Verbal Interface: Discrete or Continuous 6
2.2.2 User Community: Speaker-Dependant or Speaker-Independent 6
2.2.3 Platform Type: Single-User PC or Server / IVR based 7
2.2.4 Summary of Speech Recognition Types / Categories 8

3 State of the Technology 9

3.1 Historical development 9
3.1.1 Advances in algorithms and processing power 9
3.2 Today's Technology 12
3.2.1 Convergence of algorithms and processing power 12
3.2.2 Many vendors in the marketplace 15

4 Applicability to Call Centers 17

4.1 Call Centers in General 17
4.2 Cost Savings of Speech Recognition 18
4.3 SESA Specific Factors 19

5 Conclusion and Summary 22

Appendix A - Biometrics 24

Definition 24
Types of Biometrics 24
Technical Aspects of Voice Biometrics 25

Glossary 28

References 30

List of Figures

Figure 1. Typical Speech Recognition Processes 4
Figure 2. Typical Voice Recognition Processes 5
Figure 3. Moore's Law Example 10
Figure 4. Speech Recognition Complexity 11
Figure 5. Speech Recognition Timeline 11
Figure 6. CSR vs. Speech Recognition Call Cost Per Minute 19
Figure 7. "Good Morning" Audio Waveform 26
Figure 8. "Good Morning" Audio Spectral Distribution 27

List of Tables

Table 1. Speech Recognition Categories 8
Table 2. Speaker-Dependent Speech Recognition Vendors and Products 16
Table 3. Speaker-Independent Speech Recognition Vendors and Products 16

1 INTRODUCTION

1.1 Audience

This paper is intended to provide State Employment Security Agency (SESA) IT Directors, and Unemployment Insurance (UI) Directors, Management, and Staff with an overview, as well as a moderately in-depth understanding of speech recognition and how it is important to SESAs.

1.2 Purpose of this paper

This paper is intended to familiarize the SESA and SESA-IT personnel with the basic principles and different types of speech recognition that exist today. From this understanding the reader will be able to make a knowledgeable decision as to whether or not speech recognition can play a role in their SESA environment, and what that role will be. In addition, examples of how speech recognition can be of use in the SESA and SESA call center environments will be given.

While some vendors and their products are listed as offered at the time this paper was written, this is done only as a rough guide to what is available in the speech recognition marketplace. The speech recognition market is very dynamic at this point and as such is evolving and changing very rapidly. Readers are encouraged to perform their own market research to determine what vendors, products, and costing information is pertinent to their needs.

1.3 Overview

Section 2, Background, presents a definition of speech recognition and covers the different types of speech recognition that exist today. It also discusses the differences between speech recognition and voice recognition.

Section 3, State of The Technology, covers the history and development of speech recognition from both a software and hardware perspective. This view of the evolution, and possible future developments, of speech recognition is far ranging and not limited to any one arena of usage.

Section 4**,** Applicability to Call Centers, looks at speech and voice recognition from a perspective of applicability to call centers in general. The view is then refined to look more specifically at SESA and UI call centers.

Section 5**,** Conclusion and Summary, presents the results of this investigation into speech recognition, and its potential for use in SESA call centers.

Appendix A**,** Biometrics, presents a brief description of biometrics as it relates to voice identification and voice verification.

Glossary, describes the terms and acronyms used in this paper.

References, details the materials referenced in this paper.

2 BACKGROUND

2.1 What is speech recognition?

The term "speech recognition" has had a number of definitions, from both academic and business perspectives, over the course of its evolution. Currently, two different terms, speech recognition and voice recognition, are often used interchangeably. While often used interchangeably, these two terms have very different, yet allied, meanings.

Speech recognition, in its simplest definition, is the automated process of recognizing spoken words, i.e. speech, and then converting that speech to text that is used by a word processor or some other application, or passed to the command interpreter of the operating system. This recognition process consists of digitizing the sound-stream and then parsing that digitized data into meaningful segments. The segments are then mapped against a database of known phonemes and the phonetic sequences are mapped against a known vocabulary or dictionary of words. What is done with the words once they are identified depends upon the speech recognition application. In its simplest form on a PC, the words are entered into a word processor, as if they had been typed at the keyboard. A diagram of a typical speech recognition process is shown in Figure 1, Typical Speech Recognition Processes.

Voice recognition is in some ways a simpler process than speech recognition. Voice recognition is defined as the automated process of identifying a specific individual's voice. There is no attempt in voice recognition to necessarily identify the content, or speech, of a sound-stream, merely to identify its auditory and vocal characteristics. In this case the sound signal is digitized and then the digitized signal is compared to previously recorded samples held in a database. The result is a simple yes/no decision as to whether the speaker has been identified. Again, what is done with this information is dependent on the application(s) associated with the basic voice recognition application. Voice recognition has also been known as speaker verification or voice identification. A diagram of a typical voice recognition process is shown in Figure 2, Typical Voice Recognition Processes.

While this paper discusses some aspects of voice recognition in Appendix A, Biometrics, it is more concerned with and primarily addresses speech recognition.

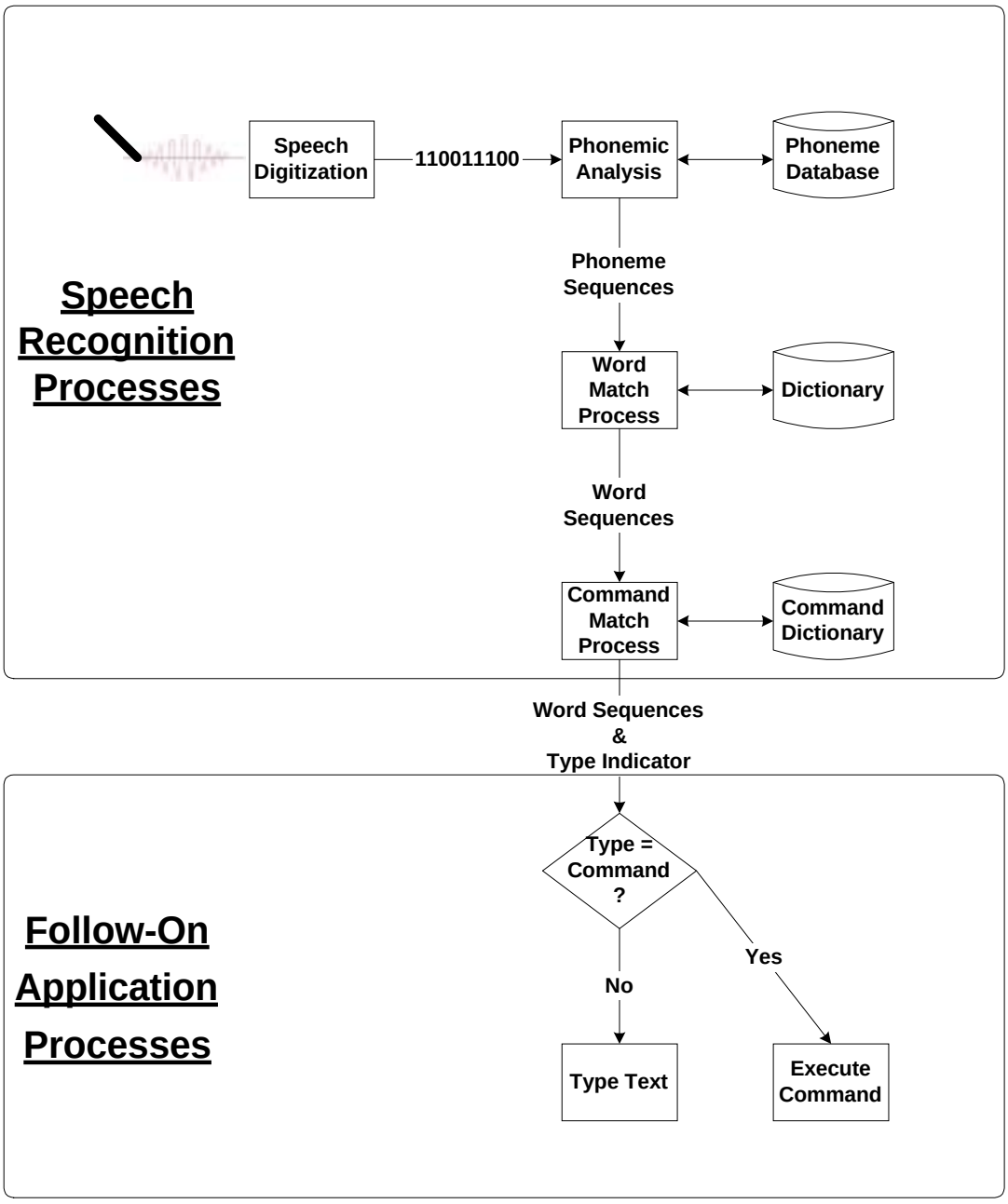

Figure 1. Typical Speech Recognition Processes

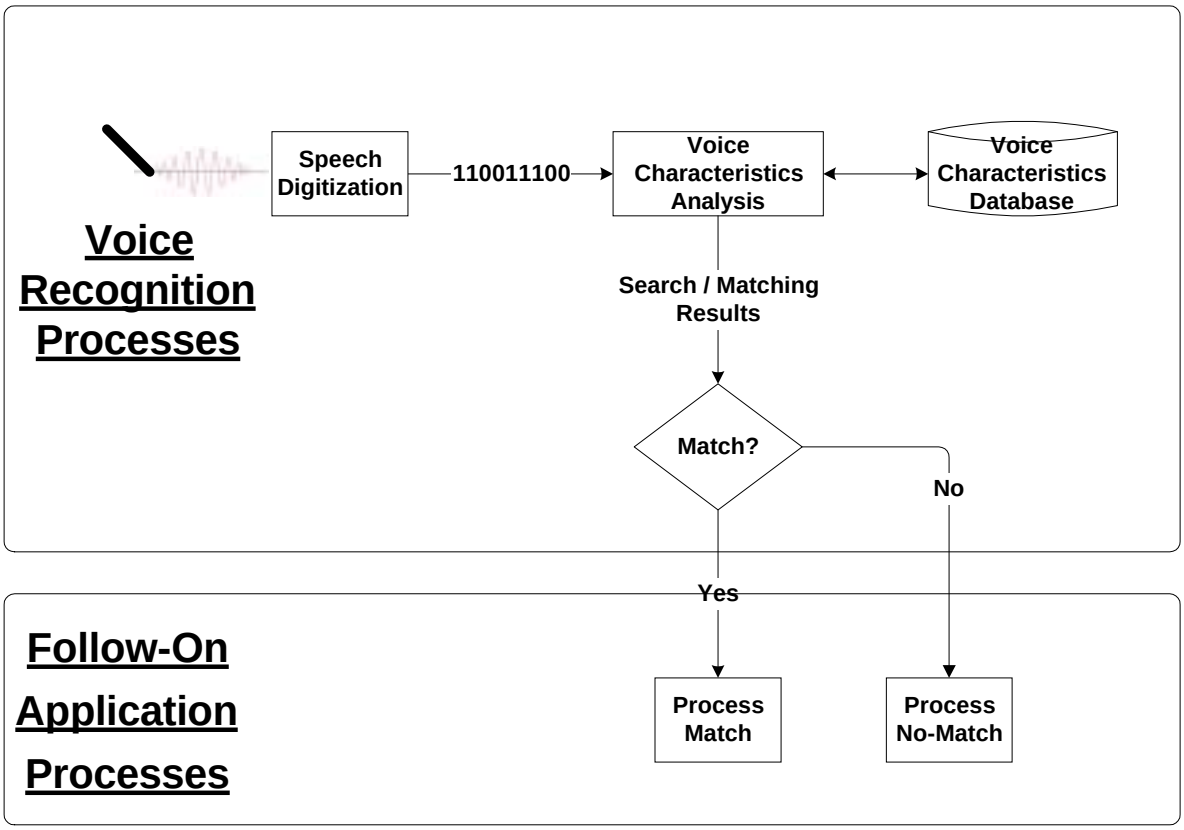

Figure 2. Typical Voice Recognition Processes

2.2 Types / Categories of Speech Recognition

Within the overall category of speech recognition there are several distinct types. These types are broken down by three main criteria. The first criterion is how words or phrases are spoken. There are two different types of systems determined by this criterion: discrete systems and continuous systems. The second main criterion is whether or not the speech recognition application is intended to function with just a single, previously identified speaker or whether it is intended for a broad population of speakers. These two different types of systems are known respectively as speaker-dependent and speaker-independent systems. The third and final main criterion is whether the application's operating platform is a single-user PC, or a server. Speaker-dependent systems are found primarily on single-user PCs, whereas multi-user speaker-independent systems are found almost exclusively on server-based systems.

It should be noted that actual speech recognition systems exist as combinations of choices from the above three criteria. Each of these three major criteria and the options within each are described in the following sections.

2.2.1 The Verbal Interface: Discrete or Continuous

There are two types of speech recognition systems from the perspective of how words are spoken to be recognized by the system. These two types of verbal interfaces are categorized as discrete or continuous.

In a discrete speech recognition system, words are spoken individually and separately – not in a conversational manner. This is done so that the processing of speech can be simpler by not having to determine where each phonemic grouping, or word, starts and stops. This is a characteristic typical of older speech recognition systems and is seen less frequently today. Where it is encountered in current systems it is usually as an adjunct to an Interactive Voice Response (IVR) system, where the script will direct the caller to either press or say a specific number or word, such as " . . . Press or say 1" or " . . . Press 1 or say Yes."

With continuous speech recognition systems, words are spoken continuously, in a conversational manner, or as if dictating text into a tape recorder for later transcription. This form of speech recognition is much more natural for the user than the discrete type because the words being spoken do not have to be separated from each other by periods of silence. For this reason they are also much faster, from the perspective of the number of words processed in a given time, than discrete systems. It is because of these two reasons, naturalness of use for the speaker and speed of speech input, that continuous speech recognition systems are rapidly supplanting discrete systems. These reasons are also why speech recognition systems are being considered today for uses where they were not realistic solutions in the past, such as in telephone call centers.

2.2.2 User Community: Speaker-Dependant or Speaker-Independent

From the perspective of the intended user community for any particular speech recognition system there are two distinctively different types of systems: speaker-dependent and speaker-independent systems.

Speaker-dependent systems require training to a specific user's voice. They do not work well with people the system has not been trained for. More one than one person can use some speaker dependent systems, but each user must first identify him or herself to the system. Most often this is done through a keyboard and/or mouse process rather than through a verbal process. When it is done through a verbal process it requires a voice identification and verification component for the speech recognition system. Most desktop PC-based systems are of this type. Another commonly encountered speaker dependent system is the voice dialing systems used with wireless telephones, which are also beginning to show up on Private Branch Exchanges (PBXs). These systems are typically very robust and highly accurate, and have large user vocabularies, especially the PC-based single-user systems.

Systems that do not require training to a specific user's voice and which are designed to work well with many people are known as speaker-independent systems. These systems tend to have smaller vocabularies than speaker dependent systems. This tradeoff is made in order to have a system that works effectively with a larger and less defined user community. These systems are most often found in menu-based applications where the user is somewhat restricted in their responses to preset questions. An example of this type of system is package tracking for an overnight carrier service, a lost-luggage system for a major airline, or a flight status information system for a major airline. Companies have typically deployed these systems because they permit an easy user interface for all callers and handle non-revenue-generating calls that would otherwise have to go to a customer service agent. One of the advantages of using speech recognition for these applications, rather than IVR systems, is that the speech recognition applications are hands-free for the user, increasing user convenience and reducing the user's think-time and related data entry times. This reduces overall call length, as the caller does not have to switch back and forth from reading a document to find the information they need to enter and then looking at the telephone set to enter that information. Also, speech recognition systems are easier to use for the entry of alphabetic information, whether that information consists of just letters or entire words and/or phrases.

2.2.3 Platform Type: Single-User PC or Server / IVR based

The final category of speech recognition systems is whether they are single-user PC-based or server-based. This is quite simply a determination of what type of processing platform the intended system is designed for. Most PC-based systems are single-user speaker-dependent continuous-recognition systems. They serve as either a dictation-type of system, or as a hands-free user interface. The hands-free requirement could be one that is unique to the specific nature of a job or task; one that enables a handicapped person to perform a job or task that they could not do otherwise; or to simplify a task or make it safer, such as stock-taking or wireless phone dialing. PCs currently have the necessary processing power to run these types of speaker-dependent, continuous-recognition applications.

When looking at speaker-independent, continuous-recognition systems, PC platforms typically start to become a constraining factor. These systems require significantly greater processing power and data storage requirements as compared to speaker-dependent systems. A large part of the processing power requirements of these systems is also driven by the fact that these are typically multi-user systems. Quite simply, more processing power is required to handle several or more users at the same time. Likewise, the phonemic analysis algorithms for a speaker-independent system are more complex than the algorithms used for speaker-dependent systems. Additionally, the phoneme database required by a speaker-independent system is larger than that required by a speaker-dependent system. This is necessitated by the fact that there is a greater variability to the pronunciation of a given phoneme when the user community consists of

more than one person. This is because it is necessary to allow for speakers with differing dialects and accents. Also, because different speakers will phrase their responses in different ways, it is necessary to allow for the multiplicity of sentence structures the speakers may use. This is referred to as defining the grammar of the responses. Because of the complexity of possible multiple grammar responses, the application development effort for multi-user speaker-independent systems is much greater and takes significantly longer than setting up a single-user speaker-dependent system.

2.2.4 Summary of Speech Recognition Types / Categories

Table 1, Speech Recognition Categories, shows the three areas of differentiation of speech recognition systems, and the choices that are available within each area. Choosing a specific type of speech recognition system is simply a matter of choosing one characteristic from each of the three columns, such as discrete, speaker-independent, and server-based. The reverse is also true: each and every general-purpose speech recognition product available today is characterized by one of each of the three categories shown in Table 1, Speech Recognition Categories.

The caveat in the preceding sentence of "general-purpose speech recognition products available today" is stated because speech recognition is rapidly showing up in many already existing devices in both the home and office. Speech recognition is being built into these devices through the use of dedicated digital signal processing (DSP) and speech recognition computing chips as a high-end option. These special-purpose speech recognition systems run the gamut from voice- and speech-recognition teddy bears and other stuffed animals, to voice-activated dialing services for wireless telephones. The day is not far off when a person will be able to walk into their kitchen at home and say, "Oven, set bake temperature to 350 degrees and tell me when the pre-heating is done." And the oven will do just that, and then inform you that it has reached the requested temperature.

Table 1. Speech Recognition Categories

Verbal Interface	Speaker Dependency	Platform Type
Discrete	**Speaker Dependent**	**PC**
Continuous	**Speaker Independent**	**Server**

3 STATE OF THE TECHNOLOGY

3.1 Historical development

As recently as ten years ago speech and voice recognition were still laboratory capabilities that had not made it to the marketplace in a realistic way. Ten years prior to that it was a computer science dream and a science fiction staple. While we have arrived at the year 2001, the human-like computer HAL-9000 of the movie *2001: A Space Odysse*y, first released in 1968, is still not a reality. But many of the speech recognition capabilities and speaking abilities of HAL are here today.

A number of advances have caused speech recognition, in particular, to move from the silver screen of science fiction to the reality of today's desktops and computer services interfaces. Chief among these are the progress in computing power at ever-smaller physical sizes and ever lower price-performance points, and the progress in software algorithms for speech processing.

3.1.1 Advances in algorithms and processing power

Webster's New Collegiate Dictionary, 1977 Edition, defines an algorithm as "a step-by-step procedure for solving a problem or accomplishing some end." In the world of computer software, particularly the type of software that performs speech recognition, programs are the embodiment of the algorithms that define how to go about the recognition process. In the early days of speech algorithm development (the 1960s) computer power was far more limited than it is today. One of the major factors that had to be dealt with in developing speech recognition software was the need for real-time processing. Speech recognition is not something that can be done in a batch-processing mode. This meant that the algorithms and their related programs that could be executed were also limited. In addition, the processing power needed to run these programs was not widely available, therefore it took longer to actually get an opportunity to test the programs and get results back from the tests. When the necessary processing power was available, it was very expensive, limiting the amount of time researchers could afford to use in their investigations of speech recognition algorithms and processes. All of this determined the pace of development, which was at times excruciatingly slow. The correlation between the advances in processing power and the resultant developments in speech recognition are shown in the graphs in Figures 3 and 4.

Figure 3, Moore's Law Example, shows the growth in integrated circuit density and complexity as expressed by Moore's Law. Moore's Law was originally stated by Intel Corporation's Gordon Moore. It states that the density and complexity of integrated circuits can be expected to double approximately every two years. Moore's Law has held

up over the last thirty years – chip transistor density has increased from 2,300 transistors on the original 4004-microprocessor chip in 1971 to over seven-and-a-half million transistors on the Pentium II microprocessor chip in 1993. (Moore's Law)

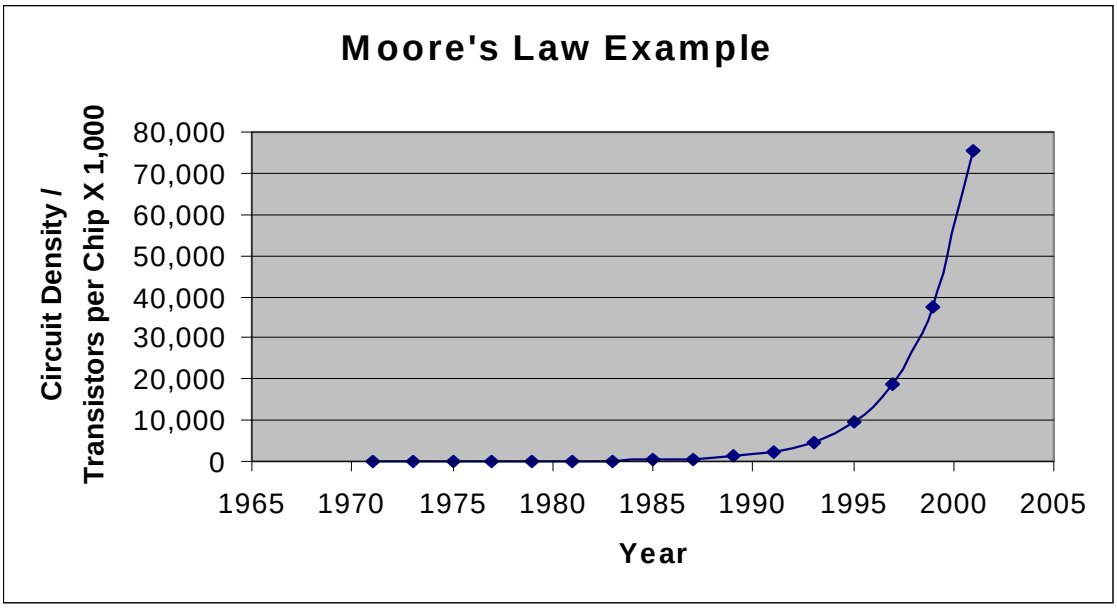

Figure 3. Moore's Law Example

Figure 4, Speech Recognition Complexity, shows the growth capabilities and complexity of speech recognition systems from the beginnings of speech recognition research in 1958 to the deployment of OnStar's latest service, the Virtual Advisor, in the spring of 2001. (Virtual Advisor is an extension of the OnStar system that will use a speech recognition system in place of live agents to provide motorist assistance.) The vertical scale on this graph is based on the premise that each succeeding generation of speech recognition technology has approximately ten times the capability of the previous generation.

A comparison of these two graphs clearly shows the correlation between the growth of the capabilities of the microprocessor and the development of the capabilities of speech recognition systems.

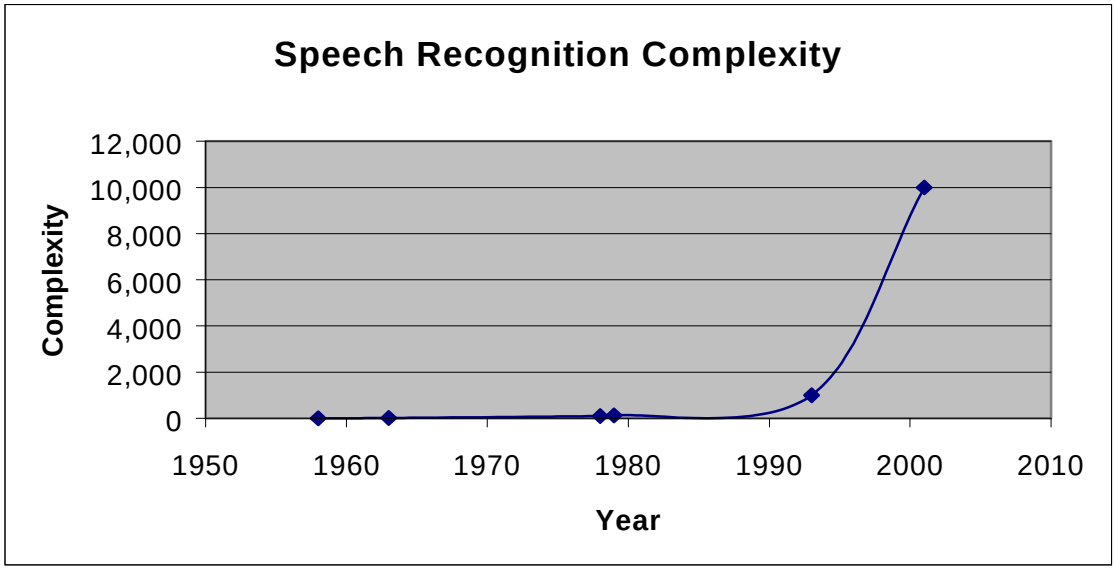

Figure 4. Speech Recognition Complexity

Figure 5, Speech Recognition Timeline, shows the timeline of the events used to generate the graph in Figure 4. This data is derived from a presentation made at the 3rd Annual Telephony Voice User Interface Conference, February 6-9, 2001. (proceedings, p.147).

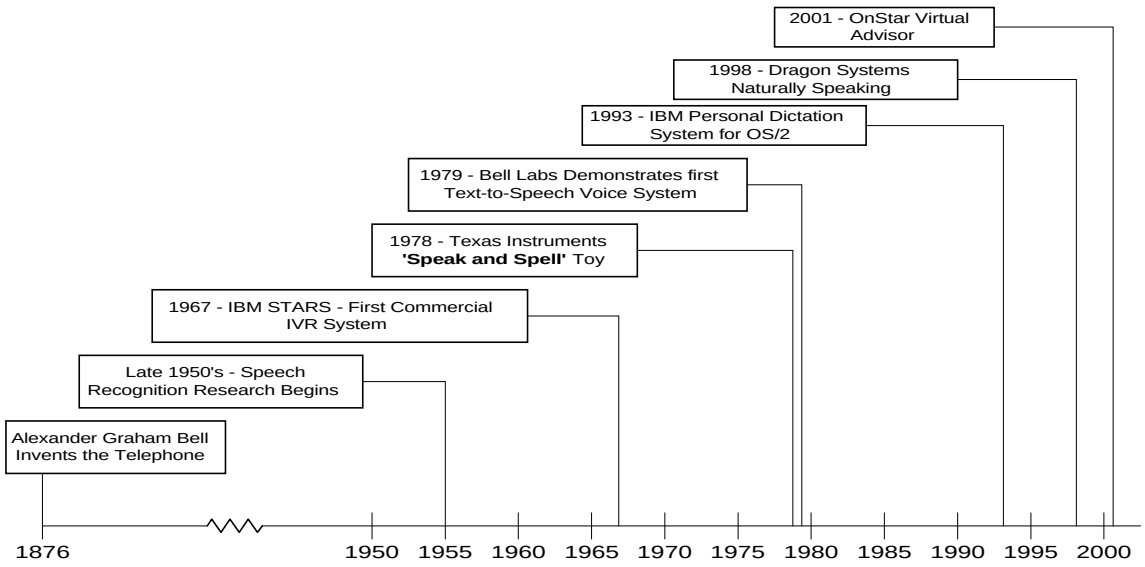

Figure 5. Speech Recognition Timeline

As processing power increased and became more generally available, the pace of development in speech recognition began to accelerate. With the development of the microprocessor, which enabled high-speed servers and PCs, development accelerated dramatically. The result of all of this is the large number of speech recognition products that are in the marketplace today. And, since processing power is still increasing at an ever-faster pace, the products that are available today are constantly being improved and released into the marketplace at a quick rate.

In addition to the rapid pace of development in speech recognition capabilities, the physical size and power requirements of the devices that are capable of speech recognition is constantly diminishing. The kitchen example used earlier of "Oven, set bake temperature to 350 degrees and tell me when the pre-heating is done" is very close to becoming a reality. The limitation is no longer based on what the technology can achieve, but on what the marketplace is willing to accept and the price it is willing to pay for it.

3.2 Today's Technology

3.2.1 Convergence of algorithms and processing power

The advances in algorithms and increases in available processing power at ever declining price-performance levels have combined to create today's technological environment where speech recognition, in its various forms, can be made available to the commercial marketplace on a realistic level. These Commercial Off The Shelf (COTS) packages and systems come in two basic forms: (1) PC based / Continuous / Speaker-Dependant software packages; and (2) Server based / Discrete & Continuous / Speaker-Independent systems, some of which include proprietary or specifically tuned, bundled server and software packages. At the low end of the price-performance scale are the single-user speaker-dependent PC-based systems that typically cost in the range from $100 to $500 dollars per seat. Server-based multi-user speaker-independent systems can easily cost upward of several hundred thousand dollars just for the platform and software, with application development often costing the same amount again. But their payoffs are much greater, as will be discussed later in this paper, and quite often have a return on investment of six to nine months, or even less.

PC-based speaker-dependent continuous-speech recognition systems cost considerably less than speaker-independent systems. It should be noted that the limitations in their capabilities are not with respect to their recognition abilities with their single user, but rather their design limitation of being a single-user or speaker-dependent system. Installation of these systems is very straight forward, as it is with just about any PC package. Once the system has been installed on a PC, the user is walked through a series of setup steps to adjust the microphone and speakers/headset, and to learn a starter set of words, or beginning vocabulary. This is done by having the user read a preset group of text. The system uses this to match the vocalization of the speaker against the known

words of the text. From that point on, the system learns the speaker's vocabulary and vocalizations by having its misinterpretations corrected by the user. Within a short period of time, from days to a few weeks, the system will be operating with an accuracy of 95 percent or better, and recognition accuracies of 99 percent are commonly reported. The more the speech recognition system is used, the more accurate it becomes.

Most versions allow for verbal command recognition and facilitated correction of recognition errors. Using such capabilities, typical users can develop skills to be able to enter from 40 – 70 words per minute into their PC applications, as verified in tests described by Ziff-Davis at www.ZDNET.com.

Server-based speaker-independent continuous-speech recognition systems currently are the high-end systems of speech recognition. As such, their installation and customization are far more complex than the single-user PC-based systems. Whereas a single-user PC-based system can be installed in a matter of minutes and be usable within a few hours of installation, speaker-independent (multi-user) server-based systems can take months of customization and application development work before an application is ready for deployment. The main reasons for their being server-based is the processing power needed to run the applications for many concurrent users, data storage for the much larger databases required, and the processing power to run the more complex matching algorithms against the databases. Along with this comes a much higher initial price tag for the systems themselves. The advantage they bring with them, however, is the power of the applications that they make possible.

Speaker-independent continuous-speech-recognition systems enable an entirely new set of applications to be fielded, as well as improving the efficiency of many currently operating public-use systems. One example of an application that it was not previously feasible to fully automate, and which is applicable to a SESA, is a change of address system. Currently, such systems are implemented using IVR capabilities to collect user information such as an account number and PIN, and determining the action a user wants to take, such as changing their address of record. Once the user has selected "change of address," the application either routes the call to a CSR or simply records the user's old and new addresses in a voice mailbox and associates that voice mailbox with the user's account information as collected via TouchTone® input. In the latter case, an individual in an administrative or account management area of the company then calls up this information on a terminal. That individual then verifies the old address information, and keys in the new address information to updates the record(s) on file.

Using speech recognition eliminates the manual process of reviewing and entering the changed information. Once the IVR system has determined that the caller wants to do an address change, it prompts the user for their current address. This is processed by the speech recognition system and verified against the account record. If the data has been verified as correct, the user is prompted to speak in their new address information. This data is then converted into text by the speech recognition system and stored in the user's

record in the database. The process is now completed at the time of the user's call and requires no manual processing.

Another example, this time of an application that is completely new and enabled by speech recognition technology, is a traffic report, locator, and driving directions service for cellular and PCS telephone users. Airbiquity Corporation and Televigation Corporation have jointly announced:

> *"The first Global Positioning System (GPS)-enabled, voice-activated, location-driven application in a complete, end-to-end solution. . . . This application enables location services, such as traffic reports, turn-by-turn driving directions and nearby business searches, based on the actual position of the user. The GPS-based application resolves the most common issue for wireless callers with location-sensitive services: knowing their exact position, which is not obvious when driving on a highway or in an unknown city. Location capability is provided by Airbiquity's GPS Accessory, a retrofittable device consisting of GPS technology embedded in the battery of Nokia wireless phones, the most widely used mobile phones on the market today. Voice activated navigation services are provided by Televigation. Users can find the nearest ATM, restaurant, or gas station, get driving directions to an address, or receive personalized traffic reports, based on his/her current position, without the need to input an address, zip code, caller ID, or other inconvenient or less accurate means of determining the origin. . . . The wireless user [interacts] with Televigation's server via voice to access information such as driving directions, yellow page search and traffic reports. Connecting Airbiquity's GPS Accessory to Televigation's voice- activated navigation server gives wireless phone users the ability to receive spoken location information with the push of a button."*
>
> *(**e-Blast**, Speech Technology Magazine's e-mail newsletter, January 24 2001 – Vol. 1, Issue 4)*

In addition to change-of-address type of applications that were not previously feasible to fully automate, and integrated wireless telephony applications such as the Airbiquity/Televigation example above, a third category of speech recognition capability is just beginning to emerge in the marketplace. This is the dedicated special-purpose speech recognition capability. We already have speech recognition teddy bears and the like, as mentioned earlier. The wireless dial-by-name speech recognition capability, currently implemented as a hosted service by the wireless provider, will soon be implemented within the wireless handset itself. This capability will be found in cellular, PCS, and home cordless telephones. It is already present in some of the soft-phones

found on PCs. As processing power becomes even smaller and more powerful, and as speech recognition algorithms are refined and also become more powerful, speech recognition will spread to even more uses.

All of these advancements in speech and voice recognition are being driven by three interrelated factors. The first of these is advances in processing power, both in general-purpose computing platforms (PCs and Servers) as well as dedicated built-in processors from teddy bears to appliances. The second factor is advances in algortihms, making the processing of voice and speech faster and more accurate. And the third factor is the marketplace, which ultimately determines which products will make it into users' hands, and whether or not they survive in the long-term.

3.2.2 Many vendors in the marketplace

Because of the progress that has been made in the past few years with speech recognition, the following two tables of vendors and their products are restricted to continuous speech recognition products. While the information in these tables was current at the time of this paper's publication, the market is very dynamic and products, pricing, and even the vendors themselves, are constantly changing. For these reasons readers are advised to perform their own market surveys of available products.

Table 2, Speaker-Dependent Speech Recognition Vendors and Products, presents the vendors of PC-based speaker-dependent systems. Table 3, Speaker-Independent Speech Recognition Vendors and Products, presents the vendors of speaker-independent, multi-user, server-based speech recognition systems. Since these types of systems vary broadly in their complexity and ease of application development, a thorough investigation of vendor offerings must be made before selecting any one for a given suite of applications.

When looking at speaker-independent multi-user systems one must consider a number of factors in addition to price-point. Not the least of these factors is the number of ports the base package supports, the maximum number of ports that can be handled on a single server, and the number of ports per incremental upgrade and the cost of each incremental-port upgrade. Likewise, a careful consideration needs to be given to the application development environment: how easy is it to work in; how readily available are developer resources within the organization, from the vendor, and in the general marketplace; and are any standards, either industry or de facto, being adhered to.

Table 2. Speaker-Dependent Speech Recognition Vendors and Products

Vendor	Product
Dragon Systems*	Naturally Speaking
Lernout and Hauspie	Voice Xpress
IBM	Via Voice

*Recently purchased by Lernout and Hauspie

Table 3. Speaker-Independent Speech Recognition Vendors and Products

Vendor	Product
Conversay	CASSI
IBM	Embedded ViaVoice, WebSphere Voice Server
Lucent	Lucent Speech Server
Natural MicroSystems	HearSay
Nuance	Voyager, V-Builder, & others
Philips	SpeechWave, SpeechPearl, SpeechMania
Phonetic Systems	Voice Search Engine (VSE)
Sound Advantage	SANDi
SpeechWorks	SMARTRecognizer
viaFone	OneBridge
VocalPoint	VoiceBrowser

4 APPLICABILITY TO CALL CENTERS

4.1 Call Centers in General

Speech and voice recognition are useful tools in call centers in a number of ways. Not all of these tools and uses are exclusive to the call center environment and could be applied to other locations within a SESA. They also make use of both single-user speaker-dependant PC-based systems as well as multi-user speaker-independent server-based systems.

One of the more generic uses of single-user speech recognition is as an aid to Customer Service Representatives (CSRs) and others to reduce or eliminate repetitive stress injuries that result from keyboard and mouse use, particularly carpal tunnel syndrome. By replacing an individual's keyboarding of information with a speech recognition interface, repetitive stress injury complaints decrease significantly. This has proven to be the case in a number of large call centers, one in particular being that of the L. L. Bean Company.

A second single-user application is to enable job positions to be filled by physically handicapped individuals. Both L. L. Bean and the State of Michigan UI Appeals division have used speech recognition applications to enable paraplegic individuals to work at jobs they would otherwise not be able to do. This has been accomplished by replacing the keyboarding of information with speech recognition applications.

The other major area of speech and voice recognition use in call center environments is through the use of multi-user speaker-independent server-based systems. These systems are often used in conjunction with existing or new IVR applications. In either case, they serve to extend the capabilities of the IVR applications. As described previously, speech recognition can automate a manual CSR task such as address changes. It also allows a user to use a previously established password or pass-phrase in place of entering a PIN via the TouchTone® keys on the telephone. This password/pass-phrase application makes use of both speech and voice recognition. The password or phrase is used via speech recognition, and the caller's vocal characteristics are matched against a reference database for voice verification. This process can even be hidden from the user by simply having their pass-phrase be their name. When the system answers their call it simply asks them to identify themself by speaking their name. This process also eliminates the problem of users forgetting or losing their PINs. After all, it is very rare for a person to forget their own name. If the system has difficulty verifying their identity, a keypad-entered PIN can always be used as a backup, with a CSR being the final fallback support. It should be noted that using a PIN as a backup to voice verification decreases the overall security of the system, since a stolen PIN can simply be used to bypass the voice verification process. The overall security aspects of the system will determine if a PIN

backup is an acceptable means of gaining access to the system or any particular application.

As with change of address information, multi-user speaker-independent server-based speech recognition allows a many applications to be developed and implemented that heretofore required a CSR. While most useful for simple multiple-choice selections, similar to an IVR menuing system, it also enables the entry of free-form text by the caller in a very simple, straightforward, and natural way. Whether this information is used to directly update a system's databases, or if it is first reviewed by support staff prior to being released for system updating, is a matter of management choice. Certainly, during the early phases of a new application's introduction it might be both cautious and helpful to review any updates before they are made.

4.2 Cost Savings of Speech Recognition

While the single-user PC-based implementations described above do not have a direct major impact on revenue and cost factors, they certainly do contribute to an organization's performance and productivity. And they come at a very low investment cost to the organization as well. The multi-user server-based applications do play directly to the organization's cost of delivery of services, however. And while they do have a significant investment cost, from initial acquisition to customization and implementation, they also have very significant returns on that investment. In general, companies using speech recognition in call center environments are seeing significant cost savings as compared to CSR-handled calls. Typical costs for CSR-handled calls are in the $1.00-$1.50 per call minute range. Typical costs for speech recognition handled calls are in the 10¢-30¢ per call minute range. [Proceedings of the Third Annual Telephone Voice User Interface Conference Feb. 7-9, 2000, Tarzana, CA]

These costs are achieved by diverting a high percentage of calls from CSRs to a speech recognition-enhanced IVR system that can obtain the caller's information or answer their questions without a staff person involved. The cost of the call consists of the telephone charges plus the amortized and allocated cost of the speech recognition with IVR system on a per minute basis[1]. The fundamental question for SESAs is the percent of calls that can be diverted to the system.

Figure 6, CSR vs. Speech Recognition Call Cost Per Minute, clearly shows the differences in per minute call costs between CSR handled calls and Speech Recognition handled calls of a like type. Using the most conservative approach and taking the lowest CSR-handled and highest speech recognition handled call costs, which is a 70 percent reduction in per call minute costs. Using the most aggressive comparison, this gives a 93

[1] <http://www.nuance.com>The Business Case for Speech Recognition, Nuance Communications, Inc. Menlo Park, CA 94025

percent reduction in per minute cost of a call. With cost savings like these, it is clear to see why speech recognition is being adopted for call center use as rapidly as it is.

And not all of the benefits resulting from implementing speech recognition in the call center environment are of a direct financial nature. Customer surveys done of the users of speech recognition applications have shown that the users are very satisfied with using the systems, and even prefer using the speech recognition systems to using an IVR system or speaking to a CSR.

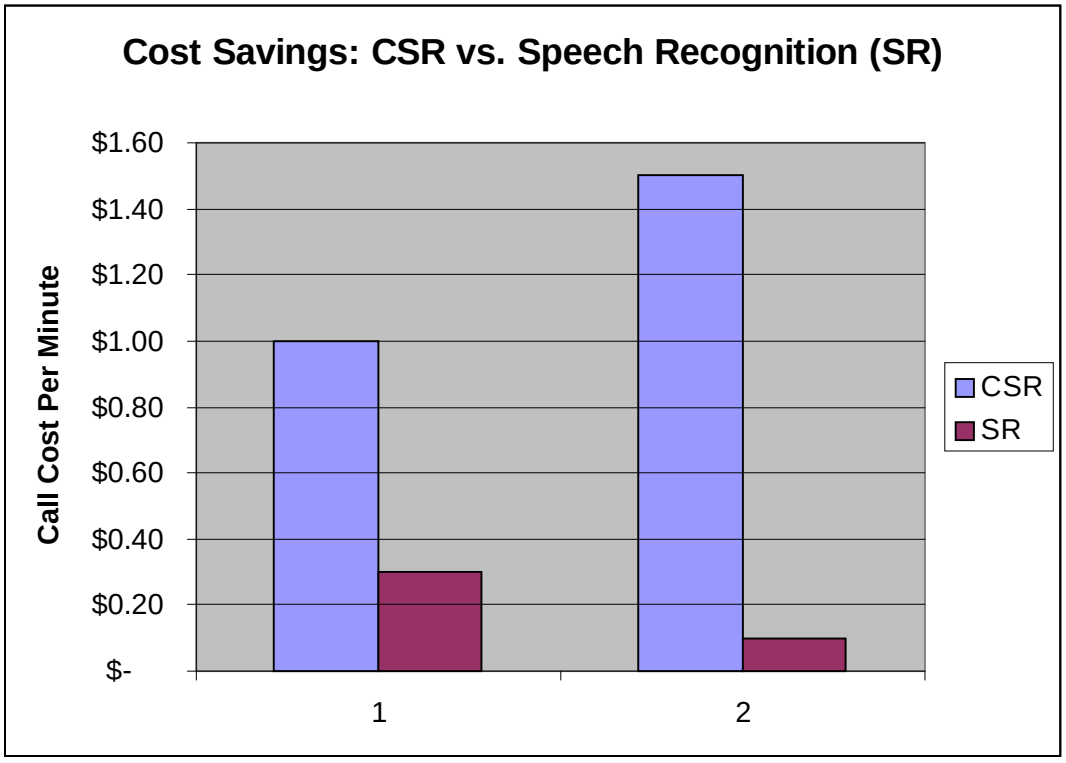

Figure 6. CSR vs. Speech Recognition Call Cost Per Minute

4.3 SESA Specific Factors

Within the SESA environment speech recognition can be utilized to perform the same functions as described above. Single-user speaker-dependent PC-based applications can be implemented to mitigate or eliminate repetitive stress injuries and to enable handicapped persons to fill positions that they would otherwise not be able to. In addition, it can streamline other positions, such as adjudications and appeals, where text

is normally entered by keyboard or tape transcription for later keying by administrative staff.

With respect to multi-user speaker-independent server-based systems, SESAs can take advantage of speech recognition in numerous ways. The most obvious use is to enhance IVR applications, coupled with the use of speaker verification in place of the use of PIN numbers. With speech recognition it is possible for claimants to completely file an initial claim without ever having the need to speak with a CSR. In addition, by using speech recognition for automated initial claims filing, the claimant can establish their pass-phrase while doing their initial filing and then use that as their claimant verification for later use. Speech recognition can also be coupled with Text-To-Speech (TTS) capability for claim status information and continuing claims data entry. Likewise, applications can be set up to handle the filing of additional and re-open claims using speech recognition, taking that burden away from the CSRs.

Server-based speech recognition can also be used as a replacement for IVR applications that provide the public with general claims filing information. Rather than keying the information in on a TouchTone® telephone keypad, the caller can simply speak the required information into the phone. This process is actually faster than using the phone's keypad, thus shortening call durations. This results in reduced port requirements as well as lower 800-number costs.

Overall savings from use of server-based speech recognition systems for UI call center operations will depend upon the number of CSR-supported minutes of caller time that can be diverted t the speech-enabled IVR. A white paper from Nuance[2] shows a 3.6-month pay back for installing and operating a standard size speech recognition system when applied to a 100 CSR call center. The percent of call minutes that are assumed to be saved is 80 percent. Typical results for SESAs may be subsequently lower, perhaps at 50 percent Even at a reasonable 50 percent their analysis would show a six month system payback for a CSR SESA call center.

Just as speech recognition can be used to improve existing automated applications and add new ones for the claimant population, it can also be used for employer applications that have previously been done either manually or via IVR. The most obvious of these are employee wage reporting and separation reporting.

The use of speech recognition applications can benefit SESAs in a number of ways. First and foremost they can reduce the need for call center agents by automating routine tasks that they handle today. This will enrich the CSR's jobs by removing the routine tasks and leaving them the more challenging calls to handle. Another strong positive to using speech recognition is that it is far easier and less costly to add telephone lines and ports to

[2] <http://www.nuance.com>The Business Case for Speech Recognition, Nuance Communications, Inc. Menlo Park, CA 94025

a speech recognition system to handle sudden surges in caller demand, rather than trying to find, hire, and train qualified CSRs. Likewise, when demand falls off, call center management is not faced with the problem of having excess staff on board.

5 CONCLUSION AND SUMMARY

Speech recognition has developed very rapidly over the last ten years. This has largely been the result of advances in processing power, which has enabled correspondingly rapid advances in speech recognition algorithms. The speech recognition systems that are on the market today are the embodiments of these new algorithms. What was once the property of the R&D lab and science fiction is now a reality.

It is often referred to, as speech recognition is actually two separate forms of voice input processing. Speech recognition is the conversion of spoken input to text or commands to an application or the computer operating system. Voice recognition, on the other hand, is the identification or verification of an individual's identity using speech as the identifying characteristic.

Speech recognition systems fall into two main types: speaker-dependent continuous-speech PC-based systems, and speaker-independent continuous-speech server-based systems. Older systems typically were discrete-speech systems in which each word had to be spoken separately. These systems have since been largely supplanted by the newer continuous-speech recognition systems.

Speech recognition represents a very real option for user input, whether that is a single user system, or one fielded as an adjunct or replacement for an IVR system. Recognition accuracies for both single-user PC and server-based systems are now in the 95 percent range and better, with many high-end systems consistently delivering 99 percent accuracy. As good as these systems are, the entire field of speech recognition is still evolving, and very rapidly. As these systems continue to evolve they will function both faster and more accurately, and more languages will be added to their repertoires. Additionally, as with all technology products, the price-performance of these systems will continue to drop as they evolve and penetrate the marketplace further.

Just as speech recognition is bringing competitive and workplace advantages to many companies in the private sector, it can provide distinct workplace and service advantages for SESAs. Many of these advantages also carry with them performance and efficiency improvements, helping to reduce the cost of delivery of the required services, and speeding up the time it takes to process initial and continuing claims.

Of great significance in the call center environment is the use of multi-user, speaker-independent speech recognition systems. While carrying significant up front costs for platforms, software, and application development, they also have significant cost avoidance or reduction numbers. Per call minute savings typically range from 70¢ to $1.40, and from 70 percent to 93 percent call cost reductions. Appropriate applications

of speech recognition systems to SESA are likely to result in a return-on-investment pay back period of less then twelve months.

Speaker-dependent single use speech recognition software provides a viable option for SESA staff to replace typing requirements using PCs with voice input and voice-activated commands with reasonable facility, speed and accuracy

APPENDIX A - BIOMETRICS

Definition

Any definition of the term biometrics is dependent on the intended scope of that definition. Cornell University's Department of Biometrics <www.biom.cornell.edu/faq.html> provides us with a definition of biometrics at its broadest, most encompassing level:

"Biometrics is the application of mathematics and statistics to problems with a biological component, including problems in the agricultural, environmental and biological sciences as well as medical science."

This is all well and good, but entirely too far-reaching and general for use within the restricted scope of speech and voice recognition. Within this restricted scope, a far more useful definition is the one put forward by The Biometrics Consortium <www.biometrics.org/html/introduction.html>. (The Biometric Consortium serves as the US Government's focal point for research, development, test, evaluation, and application of biometric-based personal identification/verification technology.)

"Biometrics: Automatically recognizing a person using distinguishing traits."

We can narrow that definition even further for our purposes:

"Speech and voice recognition biometrics is the automated process of recognizing a person using distinguishing vocal and speech characteristics."

Types of Biometrics

Speech and voice recognition biometrics today are primarily restricted to voice recognition, as described earlier in Section 2.1, What Is Speech Recognition? Any words or phrases used in the biometric process are simply ways to restrict the reference sample and simplify the recognition process.

Voice biometrics can be placed into one of two specific categories: identification and verification. Identification consists of determining who an individual is. The "Who am I?" question. Verification consists of determining whether or not an individual is who they say they are. The "Am I who I say I am?" question. Of the two, verification is the much simpler and more reliable process.

In voice identification biometrics, the vocalization characteristics of an individual are compared to a database of reference samples. These reference samples are pre-recorded words or phrases that are stored for later comparison to a live sample. If the comparison process finds a match between the live sample and an entry in the reference database, the individual is considered successfully identified.

In voice verification biometrics, the vocalization characteristics of an individual are compared to a "specific" reference sample in the database with a resulting match/no-match condition. In order to allow for false negative results - possible no-match conditions resulting from normal variations in a person's vocalization patterns that result from a cold, laryngitis, or other reasons - most voice verification systems allow for a keyboard-entered password as an alternate means of verification.

The reality of today's technology with respect to voice biometrics is that voice verification is a real capability, whereas voice identification for large population groups is not. Voice identification does work reasonably well for population groups numbering in the tens of individuals. Once the population group becomes in the hundreds the False Positive Rate (FPR) begins to climb. Also, the amount of time it takes to perform a comparison climbs accordingly. To make matters worse, the more one tries to tighten the discrimination characteristics to reduce the FPR, the longer it takes to perform each comparison of the live sample to the database of reference samples. Just as speech recognition and voice verification have become a reality as the result of increases in processing power and improvements in algorithms, so too voice identification will become more reliable and practical as further improvements in are made these areas.

Technical Aspects of Voice Biometrics

As an example of the complexity of a voice sample, Figure 7, "Good Morning" Audio Waveform, below shows the voice waveform of the simple phrase "Good Morning." This waveform is just ¾ of a second in duration. It is fairly obvious from this diagram that the waveform data is complex and, at some points, such as the middle of the word 'morning,' very dense in information.

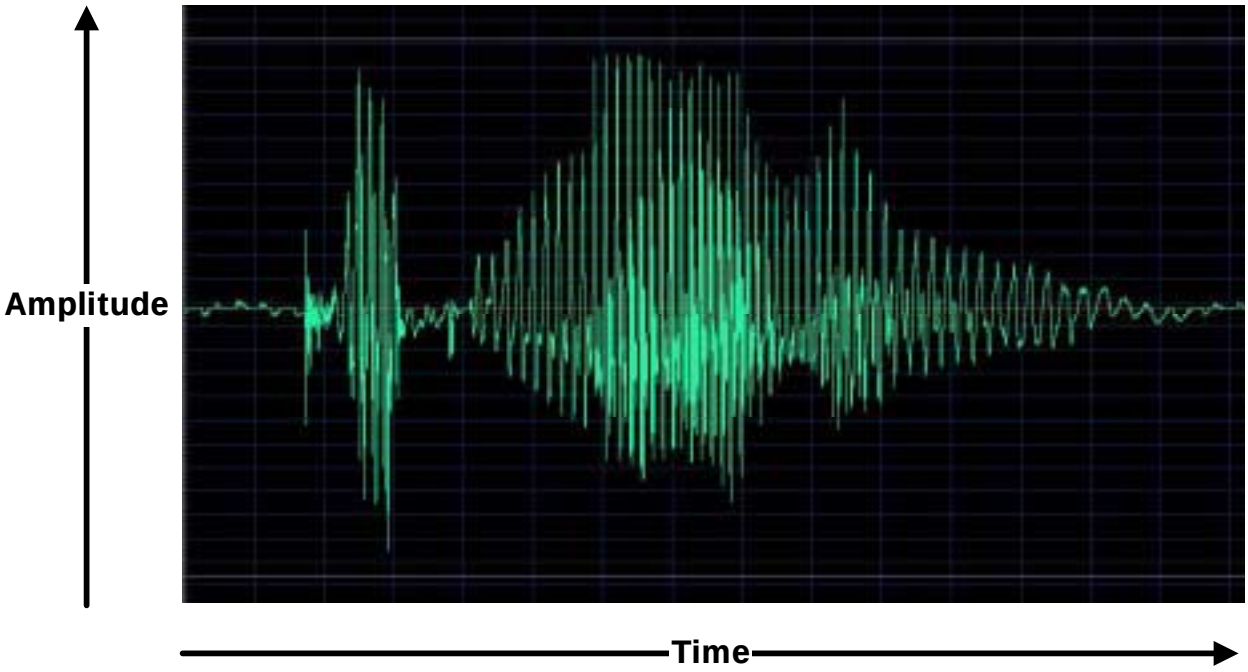

Figure 7. "Good Morning" Audio Waveform

Figure 8, "Good Morning" Audio Spectral Distribution, below, shows an alternate way of looking at voice information, known as a spectral diagram. A spectral diagram is most useful for analyzing audio data to see which frequencies are most prevalent throughout the sound file. The more abundant a frequency (the greater a signal's amplitude component within a specific frequency range), the brighter the display colors. Dark blue colors indicate that next to no frequencies exist in this range. Bright yellow colors indicate that very strong frequencies are in this range. Lower frequencies are displayed near the bottom of the display, and higher frequencies are displayed near the middle or the top. The display is linear and the top of the display represents frequencies that are just below the Nyquist frequency, which is the frequency that is half of the sampling rate for the voice data being viewed. This represents the highest frequency that can be reliably reproduced for that sampling rate. As an example, if the voice signal was sampled at 22 KHz, the Nyquist frequency is near 11 KHz. The voice sample shown in Figures 7 and 8 was digitized at a sampling rate of 22,050 Hz.

A more in-depth treatment of voice identification and verification as it pertains to SESA operations, claimant identification, and fraud prevention is contained in a previous white paper published by the ITSC, UI Claimant Recognition Technology Assessment, December 1997.

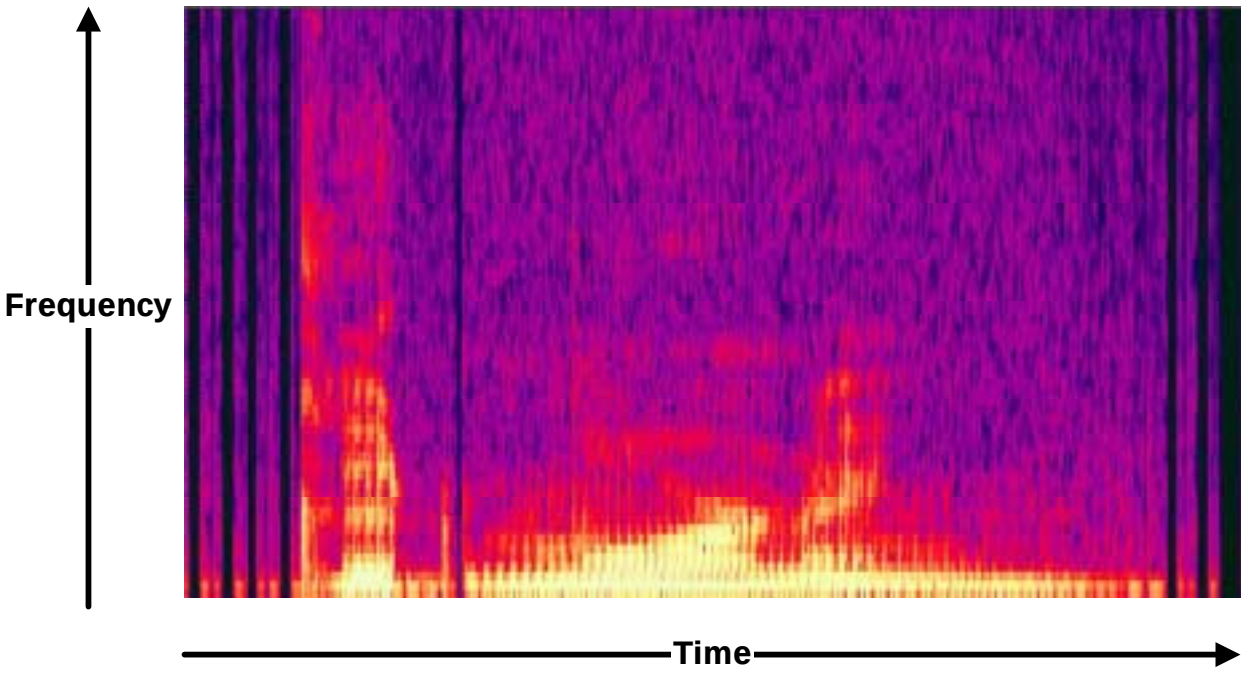

Figure 8. “Good Morning” Audio Spectral Distribution

Note: The images shown in Figure 7 and Figure 8 were obtained from the PC sound editor package CoolEdit, by Syntrillium Software Corporation, Phoenix, AZ.

GLOSSARY

Auditory Characteristics	The unique aspects of a particular sound or sequence of sounds, such as frequency and duration.
Biometrics	(1) The application of mathematics and statistics to problems with a biological component, including problems in the agricultural, environmental and biological sciences as well as medical science. (2) Speech and voice recognition biometrics is the automated process of recognizing a person using distinguishing vocal and speech characteristics.
Continuous	Not broken up by spaces or silence; conversational-sounding.
Discrete	Speech that is broken down into separate words and each word is separated from the next by a discernible period of silence.
Interactive Voice Response	Automated systems that use speech to prompt a user, who has called into the system via a telephone, to answer or indicate their choice of options by pressing one or more keys on a TouchTone® telephone's keypad.
IVR	See **Interactive Voice Response**.
Multi-User	A system designed to be used by more than one user at a time.
PBX	A Private Branch Exchange (PBX) is a piece of telephone equipment that provides the same essential telephone services as a telephone company's Central Office. It resides within a business' property or premises and handles the routing of incoming phone calls to specific numbers/telephones within that business or building. It also handles the routing of outgoing calls from any given telephone under its control to the outside public telephone network, as well as between any two or more telephones within the business or building.
PCS	Personal Communications Service, an all-digital wireless telephone service, that is a specific type of cellular telephone service.

GLOSSARY (cont.)

Phoneme	The smallest unit of speech that differentiates one utterance from another in any spoken language or dialect.
Phonetic	Of or pertaining to spoken language or speech sounds and based on the principle division of speech sounds into phonemes.
SESA	State Employment Security Agency
Sound-Stream	A continuous segment of an auditory signal or speech.
Speaker-Dependent	A speech recognition system that requires training to a specific speaker's voice that can only be effectively used by that individual.
Speaker Identification	A system that identifies an individual solely by their vocal or speech characteristics.
Speaker-Independent	A speech recognition system that does not have to be trained to an individual user's voice characteristics and which can be used by anyone form a general population of people.
Speaker Verification	The verification, through the use of voice characteristics, that an individual is who they have identified themselves to be.
Speech Recognition	The automated process of identifying and processing speech.
Transcription	The recording of speech and conversion of that speech to written text.
Vocal Characteristics	The specific attributes of a person's voice, such as amplitude, frequency, duration, pitch, and timbre.
Voice Identification	See **Speaker Identification**.
Voice Recognition	The automated process of identifying or verifying a speaker's identity through analysis of their vocal characteristics and speech patterns.
Voice Verification	See **Speaker Verification**.

REFERENCES

The Biometrics Consortium, < www.biometrics.org/html/introduction.html >1/30/01

Cornell University, Department of Biometrics.< www.biom.cornell.edu/faq.html> 1/30/01

e-Blast, Speech Technology Magazine's e-mail newsletter, January 24 2001 – Vol. 1, Issue 4

Moore's Law. < www.intel.com/intel/museum/25anniv/hof/moore.htm > 2/26/01

Proceedings of the third annual telephony voice user interface conference. February 7-9, 2001. William Meisel and TMA Associates, P. O. Box 570730, Tarzana, CA 91357-0730. < www.tmaa.com >