Automatic Speech Recognition Implementations in Healthcare

By

Richard Tobin

A dissertation submitted to the University of Dublin, in partial fulfilment of the requirements for the Degree of Masters of Science in Health Informatics

2005

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other University.

Signed: _____

Richard Tobin

2005

Permission to lend and/or copy

I agreed that Trinity College library may lend or copy this dissertation upon request.

Signed: _____

Richard Tobin

Acknowledgement

I would like to thank Mr. Derek Bennett MCh, FRCSI, FRCS (TR&Orth) and

Professor Jane Grimson, computer science department, Trinity College Dublin for their help and advice throughout this project.

I wish to thank my wife for her patience and constructive input.

Table of Contents

Declaration **2**

Permission to lend and/or copy **3**

Acknowledgement **4**

Table of Contents **5**

Table of Figures **7**

Summary **8**

Acronyms **9**

1.0 Introduction **11**

1.1 Limitations of spoken communication *11*

1.2 Evolution of Communication Technology *12*

1.3 Enhanced Information Transfer *14*

1.4 User Groups *15*

1.5 Streamlined Workflow *17*

1.6 Users with Disabilities *17*

1.7 Automation of Repetitive Tasks *18*

1.8 Innovators and Manufacturers *19*

1.9 Focus of this Thesis *19*

2.0 State of The Art **21**

2.1 Brief History of Speech recognition systems. *21*

2.2 Linguistics in Speech Recognition *22*

2.3 Dysfluencies *24*

3.0 The Use of Speech Recognition in Disabilities **26**

3.1 Physical disability and Speech recognition *27*

3.2 Dysarthria and Speech recognition *29*

3.3 Speech Recognition for Speech Practice and Comprehension *34*

3.4 Speech Recognition and Learning Disabled Students *37*

3.5 Future Research and Disability *39*

4.0 Speech-enabled Technology and security **40**

4.1 Advantages of voice authentication *41*

4.2 Disadvantages of voice authentication *42*

4.3 Categories in accuracy of biometrics 42

4.4 Preventing intruder access 43

5.0 Interactive Voice Response IVR and ASR **44**

5.1 Speech Application Development and Standards **45**

5.1.1 Voice Extensible Markup Language Voice XML 46

5.1.2 Speech Application Language Tags SALT 47

5.1.3 Controlled language 48

5.1.4 Benefits of controlled language in speech implementation 48

5.2 IVR systems and capabilities 49
5.2.1 Continuous Availability of Service 50

5.2.2 A Health System in Boston and IVR System application 52

6.0 Transcription in Healthcare and ASR **53**

6.1 Voice input devices 57

6.2 Assisted transcription and backend processing 62

6.3 Pervasive computing and ASR 63
6.3.1 Multi-modal interface 64
6.3.2 Mobile speech implementation 64
6.3.3 Digital signal processors 67

6.4 Successful ASR transcription in Healthcare 68
6.4.1 ASR Transcription in Cardiology 68
6.4.2 ASR Transcription in Radiology 68
6.4.3 Other Healthcare Specialities and ASR 79
6.4.5 Ten years experience of ASR transcription and EPR 80

7.0 The future and automatic speech recognition in healthcare **84**

Bibliography **87**

Appendix 1 *VR Operational Implementation Plan – MGH Department of Radiology* **92**

Glossary **97**

Table of Figures

FIG 1: STANDARD HEADSET 58
FIG 2: WIRELESS HEADSET WITH RECORDER 59
FIG 3: DIRECTIONAL DIGITAL ARRAY MICROPHONES 60
FIG 4: SPEECH ENGINE OF THE SERVER 65
FIG 5: SPEECH ENGINE OF THE CLIENT 66
FIG 6: SPEECH ENGINE – CLIENT SERVER 66
FIG. 7: TRADITIONAL TRANSCRIPTION REPORT WORKFLOW 70
FIG 8: ASR TRANSCRIPTION GENERATED REPORT PRODUCTION WORKFLOW 71

Summary

Advantages in using the technology of automatic speech recognition in healthcare and particularly its use in transcription appear axiomatic. Timely and accurate documentation of medical records has the potential for improvement in standards of record keeping, healthcare delivery and cost saving. Speech to text capabilities enables ease of entry into electronic medical record systems achieving rapid direct data entry and reducing data redundancy. Enhanced consistency in documentation and potential for error reduction in data recording and transfer, have particularly significant implications for healthcare records. Application of these technologies has important implications for the future development of healthcare. In a domain noted for funding shortfalls, the benefits of improved workflow, service delivery and cost saving will become critical to all health care providers and institutions. Successful implementation of speech recognition based technologies requires adherence to a planned implementation in association with structured review of user and technology based issues.

Automatic speech recognition applications in healthcare extend beyond transcription. Interactive voice response systems using automatic speech recognition provide for the capability to streamline many of the communication intensive aspects of healthcare. Such voice based technology enhances the potential for deployment of information systems which currently lack coherent integration.With continuing developments and advances in speech recognition this technology will increasingly become part of mainstream healthcare systems. The use of mobile computing coupled with voice input and automatic speech recognition provides for particular benefits in the healthcare domain. Despite apparent benefits relatively low levels of penetration of these technologies currently exists and the factors contributing to will be addressed.

This thesis is intended to provide an up to date review of the automatic speech recognition technologies in the context of healthcare. Future trends in the use of these technologies in healthcare will be discussed.

Acronyms

ASR- Automated Speech Recognition

CDC-Centre for Disease Control

DBMS - Database Management System

DES - Data Encryption Standard

DSL -Digital Subscriber Lines

DTMF-Dual Tone Multi Frequency

ECU- Environmental Control Unit

EPR-Electronic Patient Record

ETSI- European telecommunications standards institute

EMR-Electronic Medical Record

FTP - File Transfer Protocol

GUI- Graphic User Interface

HIS - Hospital Information System

HTTP -Hyper Text Transfer Protocols

IVR- Interactive Voice Response

ICT- Information and Communications Technology

I. E. E. E. -Institute of Electrical and Electronics engineering

LIS - Laboratory Information System

NLP-Natural Language Processing

PIN-Personal Identification Number

PKI -Public Key Infrastructure

ROI -Return on Investment

RIS- Radiology Information System

SALT-Speech Application Language Tags

SSL – Sound Source Localisation

TTS-Text to Speech

URL – Uniform Resource Locator

USB - Universal Serial Bus

NIST - National Institute of Standards and Technology

MGH- Massachusetts General Hospital

VUI- Voice User Interface

VoIP= voice over Internet Protocol

W3C-World Wide Web Consortium

PBX-Private Branch Exchange

HMT-Health Management Technology URL:http://www.healthmgttech.com/

VR-Voice Recognitions

XML-Extensible Markup Language

1.0 Introduction

Mankind depends on speech as the most natural and instinctive means of communication. Acquisition of the skills necessary for human speech begins in infancy but the development of skills necessary for comprehension and expression of information are improved and refined throughout life. The ability to communicate through the spoken word has been fundamental to the evolution of sophisticated human society which thrives on exchange of ideas, collaboration and transfer of information between individuals. The ability to communicate through the spoken word to one or more other individuals has been an essential component of advances in our technological and knowledge based society.

1.1 Limitations of spoken communication

Speech as a modality for enduring communication has been hampered largely by its transience and limited scope of communication to listeners in the speakers vicinity. This limitation was recognised by mankind thousands of years ago. Attempts to overcome this shortcoming by the development of pictorial and symbolic communication have ultimately evolved into what we now recognise as written language. Hieroglyphics and other graphic media have been with us for thousands of years. Several ancient civilisations developed symbol based representations which allowed for manual recording of information, concepts and ideas. The capacity to record written words allowed the transfer of information to a wider audience and more significantly allowed transfer of

information in an enduring and efficient manner. Although we routinely speak conversationally at up to 300 words per minute, assimilation of information content while scanning written text is recognised as the most efficient means of extracting the most important detail.

1.2 Evolution of Communication Technology

The labour intensive nature of early manuscripts represented a great limitation in terms of time and limited volume. The potential audience for the written information was limited by the number of copies which could be produced by hand. The ability to mass produce printed material was developed over 500 years ago. This represented a quantum leap in man's ability to communicate. Not only could accounts of events and concept of logical argument be disseminated to large numbers of people, but the production of printed text made it possible to easily create a relatively permanent record of the information. This technology required a high degree of skill on the part of the printer responsible for the type setting. The technology was further limited by the complexity and relative cost of the equipment required to produce printed text. In comparison with modern text production, early printing was slow, expensive and limited in volume of output.

The ability to produce text advanced significantly further with the introduction of keyboard based entry over 120 years ago. The development of the typewriter, and in particular the mass production of typewriters, made printed text more widely available in a standardised format. It became possible for a skilled typist to produce as much printed

text in one week than could be produced by a skilled calligrapher in a lifetime. Simultaneous advances in printing technology enabled the production of multiple copies of printed text at a more attainable economic and labour cost.

The ready availability of high volume printed text facilitated many of the advances civilised society now takes for granted such as mass media communication and transfer of information to multiple recipients in local and remote locations. These developments by providing exchange of information and ideas provided a major impetus to the development of commerce.

Over the past 40 years the widespread use of computer based word processing via keyboard entry has been realised. Again this has represented a huge improvement in the accessibility of production facilities for printed material. It has also dramatically reduced the cost effort and human time required to produce and disseminate large volumes of text based information. Early computer word processing brought numerous benefits such as the ability to manipulate and modify text. But a significant drawback was the requirement for mastery of the keyboard to be able to produce and modify text. This represents a continued limitation on man's ability to produce, disseminate, modify and discuss written information. The spectrum of keyboard skills which exists in modern life today mean that although the most skilled keyboard aficionados can produce text at in excess of 100 words per minute, it has been reported that the average office worker types at just 30 to 40 words per minute. Many office workers, particularly those who often operate at senior management level, have a typing rate far below this level. Since we

routinely speak conversationally at 150 to 200 words per minute this limitation clearly represents a significant impediment to widespread production of text based information for transfer and exchange of ideas.

1.3 Enhanced Information Transfer

The huge gulf that exists between speed of production of conceptualised text and the speed at which this text can be entered on a keyboard to produce a semi-permanent and electronic record has given rise to a number of attempts to overcome the difficulty. Speech perception is not exclusively an auditory phenomenon. In the normal conversational context input from other modalities such as lip and facial movements visually increases clarity in a noisy environment by an equivalent magnitude of twenty decibels[1, 2]. Automatic speech recognition has not yet evolved to a level capable of utilizing audiovisual speech cues and developments to date focus on speech processing and analysis as the primary goals.

At its simplest level the spoken word can be recorded either in analogue or digital format. This allows expressive communication to take place at the same rate as normal conversation. However it does limit receptive communication to a maximum speed of the same rate as voice production. This rate is much slower than the human mind's ability to assimilate information in written form. However this method of data storage does add the advantage that a relatively permanent record of the spoken concepts can be created and stored. In order to speed up the transfer of information however, it is necessary to convert the recorded audio file into written text. At its simplest level this can be achieved

by a transcriptionist. There are many drawbacks to this technique. It is highly labour and equipment intense. It incurs a significant financial cost. The stored data may become corrupt if proof reading is not performed or is performed poorly.

Over the last 10 years technology has evolved which allows both expressive and receptive communication to become automated. High quality synthetic speech can read text and data on demand allowing users to listen to e-mail or other materials while they are otherwise engaged. The advent of commercial large vocabulary continuous speech recognition software products within the last 10 years formed the basis for today's desktop speech capabilities. In addition to or instead of typing users can now use speech effectively as an input modality. Users speak and their computers can take appropriate action on all verbal demands and more significantly can immediately transcribe natural speech into arbitrary text on their screen as they speak. While far from perfect, further research and product development will continue to improve system performance and expand the markets and user groups adopting this technology.

1.4 User Groups

Not surprisingly the medical and legal professions have most strongly embraced electronic voice recognition technology. These groups must generate large volumes of text and must do their work in a time sensitive environment. Like many knowledge based workers these professionals must communicate their skills, opinions and recommendations by means of printed text. Time taken to produce a document, the cost of the production and the accuracy of the document are all key factors for these groups. Individual lawyers or doctors may generate many thousands of pages of text annually.

There is a huge variation in the production rate of text documents among these professionals. Some documents are produced in less than a day and others may take many days or even weeks. In this scenario any technology which improves throughput and reduces turnaround time will result in dramatic human and economic cost savings. Large medical and legal professionals typically have their own staff who may work on a 24 hour basis to carry out transcription duties. Certain organisations have achieved increased efficiencies and cost savings by use of offshore transcription services. Recording of speech on electronic media allows instantaneous transmission of the spoken data by e-mail to a remote site with a cost base per trained worker being much lower. Other smaller organisations may "contract out" all or a significant part of their typing. In each case the significant factors are turnaround time, cost and accuracy of the transcription. A two-stage record then transcribe process can intrinsically give rise to errors which may not be detected by the original author in later reviews and proof reading. Common transcription errors may give rise to serious adverse consequences. Omission of a single word can often radically change the meaning of a sentence. In contrast to this an electronic front-end speech recognition system allows immediate feedback, proof reading and correction of errors. The healthcare professionals are not required to undertake further rechecking and proof reading documents which have been produced directly with ASR transcription. Errors correction quality relates to the recency of assessments by the health professional.

Delays in the production of transcribed documents may result in critical errors in patient care. This is particularly relevant when a range of information from different healthcare

practitioners is required in an emergency situation. It is also significant when multiple doctors are conferring on a single patient. Delays in production of documents may also result in delays in payments from health insurers where the production of satisfactory patient data form is mandatory in order for payment to take proceed. In urgent clinical situation where all relevant information is not immediately available duplication of work already undertaken can have clinical and economic impact. Resulting costs may not be recoverable until the complete data can be submitted to the third party health insurer.

1.5 Streamlined Workflow

A further major advantage in translating text and data immediately into electronic form is the ability to develop streamlined work processes and to integrate these with the document production process. This facilitates the timely sharing of information with hospital information systems and electronic patient records and associated healthcare databases. A system which centralises clinical information in an electronic patient record improves the integrity, consistency, availability and tractability of information and also reduces potential sources of errors and delays. There are currently many commercial entities in the business of dictation systems production who have integrated computer dictation capabilities into their dictation products.

1.6 Users with Disabilities

The development of speech recognition software has also been of significant use to people with disabilities. For many of these people the availability of these systems has

enabled them to pursue a career or further their education. Many different types of disabilities can be helped through the use of this automatic speech recognition. These include neurological conditions such as Motor Neuron disease and cerebral palsy. Many large companies will now facilitate the recruitment of employees with disabilities by making voice recognition software available for computer control and data input.

Expressive communication software has also been developed to benefit people with disabilities. The ability to convert text to the spoken word allows people with visual impairment to access computer based information. For others with disabilities which impair their ability to speak, synthetic speech can make effective communication possible. Disabilities have in reality been the inception of many of today's major office speech based applications.

1.7 Automation of Repetitive Tasks

Form filling operations are well suited to voice input. Forms typically include a combination of well defined fields which require the insertion of specific and restricted types of data. Forms may also include fields for free text entry and allow the capture of data in a structured manner. Growing numbers of mobile workers routinely record data and reports into high quality hand-held digital voice recorders. On returning to their desktop PC audio file data is downloaded through the computer and can be automatically transcribed. Clinical notes and memoranda which are recorded contemporaneously are more accurate and complete than later recollections. [3]

1.8 Innovators and Manufacturers

Many large corporations have now become involved in the increasing market for voice recognition products. The industry leaders include Scansoft, IBM and Microsoft. IBM has been developing this technology focusing primarily on speech recognition since the early 1970s. There is an extensive range of Via Voice[4] products which are available in more than ten different languages from IBM. One of the early leaders in voice recognition technology, Dragon Systems, which developed the Dragon voice recognition product range, was acquired by Scansoft[5] in 2002. Dragon products are also available in many languages and are also available with a range of vocabulary sizes. Professional version products are also available with domain specific specialist vocabularies particularly in the medical and legal arenas. Philips[6] and other value added retailers VAR's supply addition transcription solutions.

1.9 Focus of this Thesis

The focus of this study relates specifically to the use of automated speech recognition in healthcare. The early adopters of this technology for data entry using ASR transcription applications within the medical domain have often been medical professionals who have limited direct patient contact. Those medical professionals who undertake their work in a quiet office setting such as radiologists have distinct advantages in the production of text using voice recognition. In contrast to this, many clinicians are required to document their information in a much less conducive environment. Very often consultations with

patients take place in environments having multiple professionals present and associated background noises. The clinical contact with the patient is often less structured than the contact in radiology or laboratory specialist area. Aside from the technological considerations relating to automatic speech recognition there are many other factors which determine its acceptability in a given settings. Ambient noise level, multiple sources of simultaneous information, repeated interruptions and the need to communicate with patients and other staff while documenting important data is representative or the reality for many, The acceptability of using automatic speech recognitions application in these clinical scenarios will be addressed. Additional factors considered include the types of equipment used and functionality of the software and hardware involved in the production ASR applications. Issues relating to successful implementation of automatic speech recognition in the clinical scenario for transcription solutions will be detailed.

ASR application use in healthcare is not limited to STT transcription. Applications of ASR technology in the area of disabilities are also addressed. Voice user response systems are capable of impacting in many diverse areas of healthcare and particularly in rationalisation of communications and information access.

2.0 State of The Art

2.1 Brief History of Speech recognition systems.

Speech recognition systems have continuously advanced over the last 35 years. In 1972 the first speech recognition system integrating dictation and word processing systems were described[1]. These systems were limited to discrete speech dictation where pauses between every word spoken were required for the speech input processing. Most speech recognition programs today have the capacity to handle continuous speech where the speaker talks naturally, without the need to pause between every word. Continuous systems do not make ongoing adaptations to the user's speech whereas discrete systems do. VoicePad Platinum is one of the last programs using discrete word recognition. Continuous speech recognition systems use stored templates of the users speech profiles which are not automatically adjusted in continuous systems however templates can be updated. Training sessions involving reading of text passages are a prerequisite for effective use of user dependent speech recognition. The capacity for vocabulary size while initially limited to approximately 40,000 words now incorporates vocabularies of more than 100,000 words in an increasing variety of specialized domains. Speech recognition programs may be speaker dependent or speaker independent. The former requires the user to train the system to develop a recognition template of words. This template is user specific and is accessed each time the user activates the system. This is the typical scenario in commonly used continuous speech recognition packages .Speaker independent systems, on the other hand, use previously stored templates that are provided

by the manufacturer. Interactive voice recognition systems exemplified by called centre applications represents a case in point.

The ability to directly convert speech to text enabling direct electronic data storage facilitates centralization of information. This improves consistency and integrity of data recording and enhances traceability. At the same time it reduces redundancy and improves processing and information access. Consistent data entry minimises sources of error and thereby realizes significant time saving. Centralizing information directly improves its integrity, consistency and availability while reducing redundancy, multiple sources of errors and time delays.

2.2 Linguistics in Speech Recognition

While the concept of automatic speech recognition is intuitive and appealing, its entails a series of complex and integrated steps in order to achieve functionality as a technology. A fundamental knowledge of speech and language structure is central to this technology. The primary input for any such system is language production by the user providing input to the automatic speech recognition system. Factors which will influence speech production will impact on this input and includes considerations such as cultural background of the user. In the case of the automatic recognition of spoken English, native or non-native speaker status is significant.

In order to process voice input the speech recognition applications must reflect an integrated knowledgebase of speech and language which incorporates language production or phonetics. It must also be capable of adapting to non native users. Most

speech to text applications are speaker dependent systems using adapted recognition profiles to enhance accuracy in automatic speech recognition for a specific user. Speech engines are continuously improving so that this limitation is likely to be eliminated in the future, allowing recognition to be achieved on a user independent basis.

Morphology represents the capturing of information regarding the shape and behaviour of words such as singular or plural enunciations and is an integral part of speech engines which achieve high levels of recognition. In order to process meaning within spoken text, knowledge of compositional semantics is required. This feature is currently more a frequently incorporated into interactive voice recognition system applications rather than speech to text applications. As this capability evolves it will result in further performance enhancements of automatic speech recognition.

ASR engines systems must incorporate some or all of these modalities of language analysis in order to achieved functional automatic speech recognition. Language analysis processing by automatic speech recognition systems incorporates natural language processing and takes account of:

- Phonetics and phonology: knowledge of linguistic sounds
- Morphology: knowledge of the meaningful components of words
- Syntax: knowledge of the structural relationship between words
- Semantics: knowledge of the meaning of words.
- Pragmatics: knowledge of how language is used to accomplish goals
- Discourse: knowledge of the use of larger linguistic units in a single utterance

Speech applications must incorporate a balance of these key language recognition elements to achieve the highest performance both in the speech to text applications and in interactive voice recognition. Natural language processing NLP is essential for speech recognition in the context of normal spoken language. In different settings as in dysarthric speech, language analysis would play little part and success depends to a large extent on phonetics.

2.3 Dysfluencies

People of all nationalities naturally intonate pauses in their own languages as an integral part of casual speech. In conversational English as spoken in the United Kingdom they say *uh* as common context filler but spell it *er*, when documented in writing.

The French say something that sounds like *euh*, and Hebrew speakers say *ehhh*. In Dutch and German you can say *uh, um, mmm* and in Norwegian, *e, eh, m* and *hm*. These disruptions pose special challenges to speech recognition systems in particular and researchers have increasingly been turning their attention to *uh* and *um* (among other so-called dysfluencies)."If someday you want machines to be as smart as people, then you have to have machines that understand speech that's natural, and natural speech has lots of dysfluencies in it," said Liz Shriberg, a research psychologist at S.R.I. International, a research company based in Menlo Park, California. *Uh* and *um* might tell a computer about a speaker's alertness or emotional state so the system can adjust itself and let people speak naturally to speech-to-text programs.

Frieda Goldman-Eisle a psychologist in London in the 1950's, who was looking for a way to make psychiatric interviews more efficient and established the study of dysfluencies by developing instruments that counted pauses in speech and measured their duration. She concluded that fifty percent of a person's speaking time was silence. She also hypothesized that a speaker planned his next words for the length of the *uh* or *um* and this was further studied by another psychiatrist George Mahl at Yale university around this time. He counted *uhs* and nine other speech dysfluencies in order to measure a person's anxiety level. He concluded that 85% of dysfluencies comprised restarted sentences, repeated words and uh and um. A slip of the tongue comprised only 1% of errors in normal speech. The frequency of dysfluencies as outlined by doctor Mahl who calculated that these occur every 4.4 seconds in spontaneous speech emphasizes the importance of language analysis as part of effective speech engines.

It has been hypothesised by Herbert Clark, a psychologist at Stanford, and Jean Fox Tree, a psychologist at the University of California, Santa Cruz, that *Uh* signals a forthcoming pause that will be short, while *um* signals a longer pause,. They consider that *Uh* and *um* are not acoustic accidents, but full-fledged words that signal a delay yet to come. Research of this nature is vital in clarifying linguistic analysis of spontaneous speech which will advance the performance of futures speech engine implementation.

3.0 The Use of Speech Recognition in Disabilities

The pace of development of computer technology is remarkable and continues to accelerate exponentially. This is particularly true of speech recognition technology. As soon as a new "state of the art" speech recognition product is brought to market, another developer announces the development of a newer, more sophisticated offering. Today, most PCs are sold with integrated speech recognition capabilities. Speech recognition systems for people with disabilities are not new technology and have been developed over several decades

Four main areas of interest have received attention by professionals working with speech recognition and disabilities. The initial area researched was use of speech recognition systems by people with physical disabilities without speech impairment. Investigation in a second area expanded to include use of speech recognition by individuals with impaired or dysarthric speech. Speech recognition was seen as having the potential to improve recognition of dysarthric speech, which is difficult to understand. The third area explored was use of speech recognition for speech practice. Using this technology was considered potentially beneficial for people with motor speech disorders, as well as those with hearing impairments by improving intelligibility using speech recognition programs that provided feedback. The most recent area studied by researchers is use of speech recognition by students with learning disabilities as an aid to improve the effectiveness of written comprehension.

3.1 Physical disability and Speech recognition

Speech recognition systems were first used by severely disabled individuals with normal speech. The goal was to promote independence whereby the technology was used to convert human speech signals into actions. Frequently, speech is the only remaining means of communication left for these individuals. The first voice activated wheelchair with an environmental control unit (ECU) was developed in the late 1970s at Rehabilitation Medicine in New York[8] The user could operate multiple items including the telephone, radio, fans, curtains, intercom, page-turner and more. A group of individuals with cerebral palsy rated the wheelchair as superior to breath control systems because it eliminated the need for scanning, allowing the user quicker access by directly selecting the desired function with voice.

The VACS is composed of a microphone, microprocessor and feature electronic display, a teletype and a relay interface. It was a speaker dependent system that the user trained with a capacity for 99 words. The visual display allows the user to confirm all commands before they are executed. It is composed of three functional modes: control, type, and calculate. Only words valid for the current mode were accepted by the system; all other words were rejected and an indication light alerted the user of an invalid function. Control mode activated up to 12 automated devices that were interfaced with the VACS. The type mode allows the user to compose written text using the phonetic alphabet. There was a capacity for up to 20 vocabulary words to be pre-stored and entered by a single utterance. All other words were spelled letter by letter, which was a fatiguing and time consuming process. If the user attempted to speak a word that was stored for use in another mode, it would be

rejected by the system. Finally, calculate mode executed addition, subtraction, multiplication and division. Although not perfect, the introduction of this device in the 1970s gave continued promise that technology could enhance the lives of individuals with disabilities. Application of Electronic Control Units in the home was later investigated. This was first considered by Damper 1984 with the development of the Voice Activated Domestic Appliance System (VADAS) [9]. The VADAS is a speaker dependent, isolated word recogniser with the capacity to control up to 16 household appliances .A similar notion was the use of speech to control robotic arms. Research was initiated in this area in 1981 and more recently by the Palo Alto Veterans Administration Medical. The goal of the robotic arm was to assist individuals with household tasks including food preparation, recreational tasks and even therapeutic skills. Few complaints were voiced about the actual functioning of the system; rather greater concern was with frequent misrecognitions of speech. Finally, a more recent consideration of speech recognition use by individuals with severe disabilities occurred in the medical setting. In this environment speech recognition could serve two purposes. First, it was used to activate functional electronic stimulation (FES) systems. This allows the patient to maintain a routine as directed by the physical therapist without the consistent involvement of a nurse. This can be motivational to the patient, providing a degree of control in their rehabilitation recovery process. Secondly, the use of computers to assist patients with disability to complete surveys or questionnaires was also been considered. Verbal responses would eliminate the need for writing, a difficult and sometimes impossible task for many individuals with physical disabilities. While systems that combine SR with ECUs alleviated many frustrations for users, several issues still required resolution. Frequent misrecognitions and

non-recognitions were unresolved problems. A temporary solution was for the user to make note of repeated failures and make changes to the stored template as necessary. As fatigue increases accuracy could decrease. Perhaps the systems should be used for limited time periods initially, to maximize accuracy, rather than using the systems for lengthy periods but with multiple errors. As endurance improves, length of use could be gradually increased. This could positively influence the user's perspective in relation to the quality and accuracy of a system. Another contributing factor to mis-recognitions was background noise. While placement of the microphone close to the mouth could help to alleviate this to some degree, it is not avoided entirely. In addition, running speech recognition systems required mastery of a complex set of rules and sequences, especially to switch between modes as in the VACS. Maintenance and upkeep of expensive devices or programs was also a consideration. Unfortunately, according to respondents of a survey conducted in 1997, occupational therapists recommended ECUs for fewer than 25% of their clients. The primary reasons cited for lack of referral were high cost and lack of third-party payer reimbursement. While this was an alarming statistic, the use of SR, with or without interfacing an ECU, was not be ruled out as an option for individuals with physical disabilities as a way to re-establish and maintain their independence.

3.2 Dysarthria and Speech recognition

A second goal is to use speech recognition as an interface to type or send signals to a speech synthesizer that would translate difficult to understand dysarthric speech into a more recognizable form. Using speech recognition in this manner could benefit individuals with cerebral palsy, survivors of stroke and traumatic brain injury, and those with degenerative

neurologic diseases such as Parkinsonism and ALS. Speech recognition may also eliminate or minimize the challenges that persons with motor disorders face when attempting to manipulate controls of augmentative communication devices. Additionally, speech recognition could allow improved interactions by improving the rate efficiency of responses. The greatest barrier to successful use of speech recognition by individuals with dysarthria is inconsistency. Severity of dysarthria not only varies across individuals; it can vary for a single speaker depending on the time of day, fatigue, stress or other personal and environmental factors. Thus, the effectiveness of speech recognition may vary at any time or any place. Research has demonstrated some success of speech recognition with speakers with dysarthria, but reports indicate a rapid decrease in performance for vocabulary sizes exceeding 30 words. Coleman and Meyers (1991) [10] used a structured set of stimuli to compare computer recognition capabilities for dysarthric and non-disabled speakers. A total of 23 Australian subjects were used, 10 dysarthric speakers with cerebral palsy (CP) and 13 non-disabled speakers ranging in age from 20-53 years. Severity of dysarthria for subjects with CP was established using the Assessment of Intelligibility of Dysarthric Speech. Average intelligibility for words was 63.6% and 52.0% for sentences. The Shadow VET/2 speech recognition system used was installed on an Apple II computer. Stimuli included 12 consonants paired with a neutral vowel, all 12 Australian vowels in an h-d environment, 12 hard words and 12 easy words from Tikofsky's list. For example, one of the hard words was platform and an easy word was chant. Each list was randomized and made up a set of stimuli. During training, participants repeated each item five times. In recognition testing, each item was verbalized three times in random order. The tester read stimulus items in random order and each subject repeated the item. The tester documented

whether the system recognized the intended item correctly. If another item was recognized, a note was made. Training and testing of each stimuli set occurred during the same session. The researchers concluded that while the total number of correct recognitions was fewer for the dysarthric speakers, a similar pattern of recognition was present for both groups. Both had significantly less correct recognition for consonants than for vowels; in addition, errors of place were greater for both groups as compared to recognition errors associated with manner and voicing. These similar patterns give hope that general improvement in speech recognition systems to help nondisabled speakers will also improve recognition for dysarthric speakers. At present, for speech recognition to be successful with dysarthric speakers, Coleman and Meyers indicate a need for adjustments to be made to the input signal or the instrumentation of the recognition system. Future research needs to address the specifics of each and determine what adjustments may facilitate improvement. Ferrier, L.J., Jarrell, N., Carpenter, T., Shane, H.C.[11] (1992) did a case study of a dysarthric speaker using the DragonDictate Voice Recognition System (Dragon Systems, Inc.). They hypothesized that there is a range of speech intelligibility that is most recognizable by the DragonDictate Voice Recognition System. Specifically, their research questions were as follows. What is the potential recognition level that can be achieved by a dysarthric user with cerebral palsy compared to normal speakers? How much time is involved in reaching the maximal level of recognition? What are the speech and voice features associated with lower levels of recognition? What phonetic characteristics predict lower recognition levels? Does word length affect recognition? Finally, is there a difference in a subject's rate and accuracy when using DragonDictate compared to a manual access computer? A longitudinal study of recognition levels in five subjects, one male with a diagnosis of

cerebral palsy (CP) and mild dysarthria, and four normals, two male, two female, was done first. This was followed by analysis of the speech and voice features of the unrecognizable words spoken by the dysarthric speaker. Results to the research questions were as follows. It took six sessions for the user with CP to reach 100% intelligibility where it took normals two to four sessions. The overall learning pattern for speech recognition of the dysarthric speaker was similar to that of the normal speakers. After the first session intelligibility was between 85-95% for the dysarthric speaker, where it reached between 90-100% for normals. Baseline intelligibility, using the Computerized Assessment of Intelligibility of Dysarthric Speakers (CAIDS)., was 84%, slightly below the performance obtained at the second session using DragonDictate. This is a positive indication that the CAIDS may be an accurate predictor of how successful a user will be with the DragonDictate (Dragon Systems, Inc.). Imprecise consonants, low loudness, hypernasality, insufficient prosody, slow rate, equal stress and final consonant deletion were all characteristics associated with lower recognition scores. Doyle, et al. (1997)[12] compared the recognition of dysarthric speech by a computerized voice recognition system and non-hearing impaired human adult listeners. Intelligibility ratings were obtained for six dysarthric speakers and six matched controls. The researchers were interested in patterns of recognition rather than accuracy alone. The IBM VoiceType recognized non-dysarthric speakers with greater accuracy than the age and gender matched dysarthric speakers; however, the learning curves between both groups were not significantly different. Gradual improvements were made at each of the five sessions across both groups of speakers. Had training continued beyond five sessions the pattern of increasing accuracy of recognition would be likely to occur. The human adult listeners were 100% accurate for stimuli produced by control speakers.

Intelligibility scores for dysarthric speakers were 94-96% for mildly dysarthric, 90-94% for moderate and 18-85% for severely dysarthric speakers. Of interest is that the human listeners judged the mild and moderately dysarthric speakers to be quite similar. This is contradictory to the results obtained on the CAIDS used to classify the subjects. The presence of this discrepancy needs to be further investigated. No trends of improvement observed over sessions occurred with human listeners as occurred with the VoiceType recognition system. Overall results indicate that the IBM VoiceType system gradually improves in recognition accuracy across session while human adult listeners judgments of intelligibility remain stable. A different perspective using speech recognition with people with severe dysarthria is the use of a small set of utterances to elicit reliable recognition. The long-term goal is to improve the individual's ease of computer access and execution[13] . The user does not have to speak recognizable words. Rather, a set of vocalizations is recognized by the system to improve the performance of job related tasks.

Finally, a similar study[14] examined the use of speech recognition in combination with scanning to increase the rate of computer input by individuals who are functionally non-speaking. As previously mentioned, not all speech recognition systems require intelligible words be used as stimuli. Specific vocalizations can be assigned to perform direct selection tasks. The goal was to increase the rate of access while minimizing any additional cognitive processing demands. Two methods of access, scanning alone and scanning combined with speech recognition were compared for six participants ranging in age from 5-21 years old. Results for a 12-year-old boy are described. He could consistently produce 3 discrete, repeatable vocalizations. The productions "ma", "hey", and "heya" were assigned either: delete the last selection, skip to the second half of the rows or second half of the columns

and to select the row of verbs. Results indicated that the participant made 3.6 errors when scanning alone and 3.9 errors using scanning combined with SR. Vocalizations were repeated on average 1.6 times before they were recognized. After five sessions, the participant was making only 2.5 selections on average per minute using scanning only as opposed to 3.4 selections with scanning and voice combined. Thus far, it seems there are gains to be made when using a combination of scanning and speech recognition even with limited vocalizations. Clearly, further research is required with a variety of individuals to draw any conclusions about the significance of such a program.

The use of speech recognition by speakers with dysarthria has great potential for growth. Continued research assessing the effectiveness of various speech recognition systems with speakers of varying severity and intelligibility of speech is needed. In addition, replication of all studies discussed across larger numbers of individuals is necessary before any concrete conclusions can be drawn regarding the benefits of using SR with speakers with dysarthria of varying severity.

3.3 Speech Recognition for Speech Practice and Comprehension

Another area of interest is computer-based training for impaired speech. The goal is implementation of a low cost speech training aid using commercial technology. Research was initiated over 25 years ago with the focus toward improving the speech of hearing-impaired individuals[15]. A study[16] cited linguistic, cognitive and attention components as the sources of initial failures. With those issues in mind, several researchers have developed

taxonomy for classification of computer based speech training systems over the last 10 years. Classification systems and deliberations of assessment and intervention of various training systems are discussed. A taxonomy[17] for speech training systems classified three categories based on the kinds of knowledge about speech that they incorporate. In class A systems, text display can be used to demonstrate a relationship between acoustic or physiologic measures and recognition accuracy. These systems also rely on knowledge or perceptions obtained from listeners. Class B systems incorporate calibrated analytical displays, that would be more typical of engineers. Physical signal attributes can be obtained but not information related to accuracy of the perception of the signal. Finally, Class C systems consist of a speech signal-to-visual transformations but lack explanations for the speech production accuracy. A Class C system may be used to teach volume control where the display shows the user level of intensity as defined by colour. The Visi-Pitch by Kay Elemetrics is one of many Class C systems on the market. Two interrelated computer-based speech training aids developed[16]: the Speech Training Station (STS) for assessment and intervention in the clinic, and the Speech Practice Station (SPS) for independent practice in the home incorporating a game format. These would be Class A systems according to Bernstein's taxonomy. The goals in mind during development of these systems were two-fold: 1) assessment of skills through an objective measure 2) practice through drill in a game format. The STS had the capability to provide the therapist with feedback of physiological parameters that was not incorporated into the SPS. Six games were generated to teach vocalization, production of repeated syllables and control of voice intensity and fundamental frequency. A limited clinical evaluation of STS and SPS was completed over fifteen months. Fifteen subjects participated in the evaluation. All children routinely wore

hearing aids and had no other known handicaps. Either individual or small group treatment occurred twice a week for 20 minutes by one of two clinicians. Subjective and objective observations of the STS were as follows. Both clinicians reported the system as easy to learn and found its capacity to individualize to the needs of the child as favourable. Two factors contributed to inconsistent reliability ratings by the computer system and clinicians. First, over-sensitivity of the computer program enabled it to detect continuous sound that the clinicians could not hear. Secondly, the computer focused on a single attribute at one time while the clinician could provide feedback about other speech characteristics regardless if it was the target of the game. Because a fixed placement of the microphone was not used initially, this resulted in variations of loudness levels mixed with ambient noise. Misreading occurred because of this inconsistency and it was felt the observations of the children's behaviour indicated a positive response to the system. Children practiced independently and used the computer even when supervision was not provided. It was determined that the children spent a greater length of time practicing on their speech than they might have otherwise had games not been incorporated. The capacity to individualize to each child's needs minimized frustration yet presented a challenge. Several clinical benefits of the aid were noted. It could be used with a wide age range. Time devoted to developing fun therapy activities was no longer required. The aid could be used alone or combined with other, more traditional therapy techniques. If two or more children worked together, pragmatics such as turn taking, were indirectly enhanced. The objective measures of fundamental frequency duration and intensity levels provided feedback to both child and clinician. Because visual displays were based on a single speech parameter, the clinician

learned to focus on the target. Both child and clinician seemed to benefit from the format of the training aid.

Supplemental use of the SPS in the home was found to enhance results of speech production that otherwise would not have occurred. Use of the practice station ranged from 82 minutes to 185 minutes. The ability to adjust the parameters of the system allowed for consistency between training and home practice activities. The entire family frequently participated in practice sessions[18]. This is of significance because family interest is critical for facilitation of spoken language[16]. Although the clinicians observed the children to be thoroughly engaged in the activities during training, parents, on the other hand, felt a greater variety of games should be provided. When interpreting the results, it should be kept in mind that parents were not trained on what behaviours and responses to observe. The existence of any parental concerns warrants the need for further investigation of the effectiveness of speech training if they are to be used on a regular basis for home based practice.

3.4 Speech Recognition and Learning Disabled Students

The use of speech recognition as a compensatory strategy for individuals with learning disabilities did not receive much attention until the early to mid 1980s. Speech recognition can enhance the learning disabled (LD) students composition. These students are prone to making more spelling, punctuation and capitalization errors. Continual pausing to correct these frequent errors during content formation interrupts the train of thought. This may result in forgetting the initial message the student wanted to express, leaving him/her frustrated. Additionally, it then takes the LD student a longer time to generate written work

as opposed to peers who are not learning disabled. LD students use a simplified vocabulary when writing to avoid spelling errors even though they may want to use a more difficult word. Finally, these students tend to have a negative attitude about writing in general[19]. Utilization of speech recognition may allow the student to focus on the planning and content generation of text rather than the mechanics of writing. In addition, speech recognition has the potential to increase the rate of production of written material and positively enhance the overall writing experiences that LD students traditionally avoid.

The most recent research is by De La Paz[19]. The purpose of her study was to provide a rationale for using dictation with LD students as a means to enhance their written composition skills. She included several suggestions for improving the quality of writing via dictation. She also emphasizes the need for advanced planning. Ideas or key words should be generated in an outline format or as notes to refer to when dictating.

Combining speech recognition with speech synthesis is another way to help LD students learn how to correct their own errors. Programs such as the Kurzweil 3000[20] (Kurzweil Educational Systems, Inc.) can read text that has been scanned into its system. The auditory component may improve the student's awareness of grammatical and spelling errors that would otherwise be overlooked when silently reading. Most importantly, it must be emphasized that speech recognition is not a substitute for learning the rules of written grammar. These skills need to be mastered; speech recognition is simply a supplementary device to make the writing process less fearful, more appealing and motivating to the LD student.

3.5 Future Research and Disability

Overall improvements in speech recognition have the capacity to positively effect the lives of all individuals who use it. Regardless of the nature of the disability, improved technology and design give potential to greatly improved accuracy and rate of dictation. Present fluctuations are not solely a result of the engineering of the system itself, but may be due to the inconsistencies in the user's speech. Therefore, the user's preferences and expectations seem vital in the development and research of new tools. Suggestions and future research related to the technology cannot be associated with one genre of users. Ongoing research is needed to investigate ways to reduce the effects of voice drift, accelerate the development of simplified instructions, and adopt improved hardware designs. If use of speech recognition systems by people with cognitive deficits and learning disabilities is to prove beneficial, additional research must be done to determine if and how the cognitive load can be minimized. Assessment of effectiveness, efficiency and independent execution across diagnoses, age ranges, and multiple systems is needed as well. The amount of training required, frequency of training sessions and methods to promote widespread application of speech recognition to activities of daily living pertinent to each user are also in need of investigation.

4.0 Speech-enabled Technology and security

Biometrics refers to the use of unique physical characteristics as a means of identifying an individual. It has been utilized extensively in security applications. Biometrics can be used as a means of controlling physical[21] access to a restricted area or by providing electronic access to user specific areas and levels of authorization. All security systems based on user based privileged access, typical of modern computer network environments, are based on access levels uniquely assigned to a specific user. Biometrics as an authentication tool is very powerful because of the enduring nature of identification features used. This is in contrast to password or swipe card security methods which are prone to recall errors or loss. When deployed in conjunction with speech recognition and text to speech businesses can build a range of cost-effective applications which can automation and improve the end user satisfaction

Accuracy is fundamental to the use of biometrics which can use a variety of technologies. Retinal scans are one of the most consistent methods available because of the enduring retinal patterns observed using low level infrared scanning. Implementation requires the user to present their eye to specialised scanners as distinct from the simplicity of microphone input used in voice biometrics.

All biometric methods are prone to method specific errors as in the case of retinal scans where lenses may cause specific difficulties. Fingerprint biometrics is susceptible to aging and wearing of skin. Handwriting biometrics relies on consistent speed and pressure and variations result in increased error rates.

Voice biometrics is based on the unique tone pitch and rhythm of speech which is used to generate a unique voiceprint. Illness particularly with hoarseness, background noise and variations as a result of input devices may cause problems with authorization and voice identification.

4.1 Advantages of voice authentication

Because no special hardware requirements are necessary and voice interfaces exists in the normal computing environment there is little cost implication in this regard. Alternate inputs can be received from telephone or microphone.In general voice authentication is easy to use and acceptable to users. Speech represents the most natural of input methods unlike scanning biometric methods.Because speech is readily transmitted using a variety of options, one major advantage in relation to voice biometrics is the ability to undertake remote authentication independent of location.

Voiceprints require initial enrolment whereby the speaker is asked to speak a number of set words or phrases. This is used as the basis for voice print generation and is generally between two and eight seconds in duration.

Voiceprint storage sizes are small and depend on the degree of security required. Typically they vary between 1Kb and 6Kb in size. Database performance in retrieving user identities is rapid as a result and this increases user acceptability. Voice- prints are not voice recordings and are very difficult to reverse engineer if unauthorized access is attempted. A study completed by TouchStone consulting in April, 2003 determined that 88% of participants found voice- authentication to be more are equally convenient to DTMF

personal identification number entry. 74% felt voice- authentication was equality or more secure compared to PIN entry.

4.2 Disadvantages of voice authentication

Voice biometrics alone do not present the most secure biometric methods although their error rates can be adjusted to reach the levels in excess of 99% accuracy. Decreasing tolerance for potential errors degrades the performance of the system and may result in unacceptable delays in authentication for the user. Age related changes in the quality and character of voice may compromise the performance of voice up authentication, although this can be overcome by repeating the enrolment process intermittently. The accuracy of voice and authentication depends on the quality of audio input. This will vary depending on the device used for example whether a near speaker microphone compared to telephone input. Environmental factors may also lead to errors in accuracy.

4.3 Categories in accuracy of biometrics

Failure to enroll can occur where users have failed to complete enrolment as required by the voice authentication application.

- False acceptance rates reflecting the frequency of authentication of impostors.
- False rejection rates reflect the frequency whereby the authentic user is rejected by the system.

Adjusting the false rejection rate is likely to increase the false acceptance rate also and a graphic plot of these two variables is demonstrated in the figure below. Where these two rates intersect is known as the equal error rate or crossover error rate. The lower the latter

the better the system is performing. These rates depend on the application used for voice identification.

Voiceprints provides the data sample for analysis by various algorithms which use combinations of tone and pitch to determine matches between stored and input voice prints. This can be “text “based whereby the two sets of data are compared based on verbalization of the same words or phrases.An alternative is “text independent” which does not depend on specific text input. In this instance the common pattern between test and reference voice input determines the match and requires a larger data sample to reliably undertake such comparison. The speech engine used in this scenario is not synonymous with that used in voice to text recognition however the two technologies can be used synergistically and can result in highly accurate authentication through the combination of voice- with knowledge authentication.

4.4 Preventing intruder access

One obvious way in which to attempt to exploit voice identification would be to replay pre recorded input from an authorized user. Text dependent methods are more susceptible to this in their simplest form. A challenge response system helps to overcome this approach to unauthorized access through the use of randomized request for words and phrases. Knowledge based voice biometrics which use automatic speech recognition provides high levels of protection against unauthorised access.

5.0 Interactive Voice Response IVR and ASR

After twenty years of development the accuracy of speaker independent directed speech recognition was greater than 95%. This represented a threshold level of accuracy as it approximated to that achieved by many operator based systems. Recent developments have achieved accuracy levels of 98% on the speaker independents deployments of voice user interface. The ability to integrate existing systems with voice recognition systems has enhanced their functionality. Standardizations in the design tools of VUI systems including voice XML enhanced the capability for development of products in this area.

Voice user interfaces provides the ability to assist users in information retrieval through the use of automatic speech recognition. This may be achieved entirely independent of operator assistance and capability described as “barge in” allows users familiar with the VUI structure to respond without having to wait for full system prompts.

In terms of the extent to which these applications fulfil the desired function, their usability of systems has been extensively studied.The principle measures used in assessing usability includes percentage task to completion as a whole. Further analysis of percentage completion by individual task, mean duration of all and individual tasks together with analysis of recognition failure in each instance provide a satisfactory measure of the overall functionality of interactive voice recognition systems. Other assessment methodology employed use video recording in assisting the analysis VUI systems. In these audiovisual analyses there is an inherent experimenter based bias,

however overall it system provides valuable data useful in predicting the real world applications of such systems.

The end user ultimately determines the usability and acceptability of any system and their experience will determine the viability of such systems together with the likelihood of further used. As automated speech recognition continues to improve with respect to recognition accuracy in particular, the end user experience is likely to continue to improve and the applications of VUI systems will undoubtedly permeate many healthcare and normal daily activities. These systems are increasingly designed to incorporate learnability and effectiveness. Learnability reflects the extent to which an application conforms to the laws of learning psychology such that it teaches users to use it effectively. It is evident that if an application can teach its users, user satisfaction will be greater and designers will find it more effective.

5.1 Speech Application Development and Standards

The compatibility and effectiveness of interfaces will inevitably become increasingly significant with the use of speech recognition technologies. Many of the early systems were closed architecture design which inhibits the transfer and flexibility needed in the field. Restrictions such as proprietary languages significantly restrict the development of speech recognition systems and increased greatly the expense involved in the use and expansion of such applications. Many of the operation platforms on which automatic speech recognition systems are developed lack specific speech enabled capabilities such as automatic speech recognition or text to speech.The ability to integrate areas such as telephony with

information and communication technology systems would be greatly limited because of compatibility issues where standardized application interfaces did not exist. Inevitably maintenance and upgrades comes with an extremely high price in this scenario.

5.1.1 Voice Extensible Markup Language Voice XML

Open standard based implementation is critical for the promotion of voice web developments. XML based languages which are fundamental to the development and implementation of web based applications relying greatly on open XML based language standard. The standard markup language for speech enabled web applications is voice XML.

Voice XML[22] is a language allowing human computer dialogue using speech with HTML based speech tags. Because of widespread familiarity with the HTML based applications it provides a standard development environment which readily transports to other applications. Many application developers are familiar with the structure of Web-based markup languages. Voice-XML is undergoing continuous development expanding its capabilities. It has been used in interactive voice response programs extensively.

Some of the limitations of voice XML include limited call control and limited multi-modal capabilities. Additional XML base languages like call control XML CC-XML, have helped to overcome these limitations and can be used in voice XML based applications.

5.1.2 Speech Application Language Tags SALT

A more recent speech development language is speech language application tags or SALT[23]. This was developed after the initial use of voice XML aimed at enhancing development capabilities. It has been developed as a collaboration between a number of major players in the software application, networking and speech technology domains. These include Cisco, Comverse, Intel, Philips, SpeechWorks and Microsoft.

SALT design is specifically compatible with existing and developing applications by Microsoft. It incorporates enhanced support for multi-modal applications specifically which facilitates data entry capabilities through combinations of voice and alternative input methods. Having the backing of large software developers with previous experience of standards development is likely to be enhanced the use of this language. Speech-language application tags provides a platform independent development environment. It is designed to readily integrate with existing infrastructure and presents a design format familiar to many programmers.

Many companies who have implemented voice- solutions have done so using proprietary legacy systems in which they have made substantial investments. These companies will be anxious to simplify development and maintenance costs by migration to open a language based developments such as voice XML or salt. In moving to open standards consideration for support of emerging technologies such as VoIP must be assured. Support for existing modalities of access such as DTMF based input needs to be given ongoing support to ensure a smooth transition with the existing users. Other benefits including enhanced call control and transfer should be assured before migration.

5.1.3 Controlled language

Controlled language is a subset of a more widely spoken language such as English, with specific grammar, vocabulary rules and style. They have been used extensively in specific environments which have requirements for limited vocabulary requirements and are overall aimed at improving readability and comprehension. Style consists of the use of short sentences, active verbs and personal pronouns. They have been implemented to standardize and simplify such applications as maintenance manuals as far back as the nineteen seventies where agreement on the structure of this language was overseen by the European association of aerospace industries

5.1.4 Benefits of controlled language in speech implementation

Controlled language will limit the extent to which misinterpretations may occur by virtue of their built-in standardization in terms of vocabulary scope and structure. The European telecommunications standards institute oversees the controlled language used in communication devices and services. This is a multilingual body ensuring standardization in English, French, German, Italian and Spanish.

Application development particularly in interactive voice response systems use controlled language, as their use of simplified syntax and avoidance of compound sentences for example improves overall design and performance. Voice- user interfaces should use natural contemporary sounding language which comprises another aspect of controlled language together with standardized short consistent responses and prompts In the

development of telephone control for example the ETSI common commands enhance integration with ASR and TTS engines.

5.2 IVR systems and capabilities

Theoretical capabilities of systems designed for real-world implementation are not relevant if not reflected in actual use. VUI based programs are used in a wide variety of applications outside of healthcare. Applications such as highly accurate text to speech playback of voice dialed companies or individuals are in extensive use. Automated directory assistance will be familiar to an ever increasing number of users. Some of the major players in this area include to Nance, Scansoft and Targus information systems.

In the future, development of the best dictionaries for use within voice user interface applications will improve through increasing use of auto-attended systems. Improved dictionaries will reflect both the scope relating to the number of possible entries and breath, which takes account of the variations in possible pronunciation. This will enhance overall accuracy. Further enhancements of the systems will be achieved through refinement in the rule based structures. The application of computer algorithms in the analysis of automatic speech recognition based system performance is an area of ongoing research which promises to automatically develop rule sets without human intervention. This approach is based on statistical language modeling.

Voice automation can be used in any setting where human interaction is not required and where the decision-making process involved lends itself to the rules based system. Many of the normal communications in daily healthcare represents such instances.

Communications such as contact with the exchange operators for transfer to specific individuals or departments within a healthcare center or hospital have the potential to be incorporated into rules based implementation as part of a call steering application.

Similarly the potential for designing systems capable of connecting with specific departments to enable information retrieval is equally amenable to implementation using this technology. Information retrieval applications previously implemented have used dual tone multi frequency [DTMF] or touchtone based methods. Performance of the systems has been limited by difficulty in navigation and the necessity for alphanumeric input which is difficult to handle.

Communication applications are familiar to those who use voice-activated dialling which is incorporated into most contemporary mobile telephones. This application of voice automation can provide ready access to individuals and departments. In addition it can easily access e-mail, faxes and voice mail using a phone for example. Inquiry such as billing which represents a transaction application using automated speech recognition can assist the user in retrieving details related to outstanding financial transactions.

5.2.1 Continuous Availability of Service

Apart from information access this technology also provides continuous services independent of fulltime operator requirements coupled with the ability to incorporate security through voice related biometric technologies. Given that many areas in healthcare must by its nature be available at all times, services systems which are continuously available have particular appeal. As with other automated speech technologies interactive voice recognition systems have been

demonstrated to achieve significant returns on investment within short time frames. Cost savings are ongoing insofar as once implemented the systems supplants the need for the level of manpower required in the non-automated environment. As a rule most interactive voice response installations are used in replacing repetitive task based functions provided by personnel and have been shown to repay investment costs within six to twelve months. Cost savings are realised through a combination of factors including, improved productivity, revenue generation and cost reduction and avoidance. A leading analyst firm Giga estimates that there is a twenty to sixty percent increase in use of interactive voice response systems compared to touch tone DTMF based access. This reflects the enhanced user experience in IVR versus DTMF.

After implementation these systems provide the capability for enhanced management, data mining and report generation. Data collection allows for ongoing adjustments and tuning of recognition parameters within a given application which further enhances efficiency and usability.

Technology development in the IVR systems requires the normal processes of application design, development testing and deployment. Throughout development awareness of the needs of the end user must be paramount and continuously reviewed as in standard graphic user interface based developments. Voice application development differ significantly from standard applications specifically as they utilize a voice user interface which requires a completely different skills set from developments normally based on graphic user interface GUI.

IVR systems coupled with ASR are currently in relatively early development phase. As speech recognition continuously improves voice-enabled technologies will inevitably permeate all areas in healthcare and all spheres of normal living through the use of such systems.

5.2.2 A Health System in Boston and IVR System application

Caritas Christi Health System in Boston Institution implemented an interactive voice response system incorporating intuitive speech based telephone connections with departments and individuals. In place of previously fulltime staffing of telephone exchanges, users of this system can access the IVR system either internally or externally and utilize speech based requests for connections, for example "connect me to the nurse's station on the second floor". In similar manner connections to employees pagers can be achieved through spoken request. This system is operational 24 hours a day and has tools that have allowed the integration of multiple data sources providing continuous updating of contact databases. These can be accessed through a web interface and accessible from any external portal. The system has been made available to both public and staff. A spin-off benefit in this scenario has been the ability to dispense with expensive frequently outdated telephone directories. Inter hospital connections utilizing existing data connections have also resulted in substantial savings. A 30% reduction in switchboard calls has been noted and in this institution to date and constitutes in excess of 350 fewer calls per day.

6.0 Transcription in Healthcare and ASR

In the United States almost half of medical charting is generated from speech, with the other half being handwritten[24]. Because compositional speech outputs at up to 300 words per minute, even complex examination details can be dictators in minutes. In a study conducted by the Centre for Disease Control (CDC), 72 percent of Americans visit office-based settings for ambulatory care. In 2000, the number of Americans visiting office-based healthcare settings averaged 756.7 million. Medications were provided in three quarters of all visits with an average of 2.3 drugs per prescribed case.The CDC recorded 30.8 million non-institutionalized adults in America who suffer from hearing loss. Efforts have been made to open the communication gap between healthcare givers and patients who are hearing impaired or deaf. Using software that allows the healthcare provider to speak into a microphone system which enables speech to be converted into a number of formats including sign language, patients can view the translated communication on a computer screen. Speech represents the most effective means by which we record knowledge. It is estimated that the cost of transcription services per line is 15¢ and a typical forty line report would costs $6.00 based on data from 2000. Between $10 and $24 billion per year are spent in the USA[25.]with ongoing growth in transcription of the order of 30% per year. Consideration of the current workflow costs including handwritten documentation, editing and chart retrieval all have bearing on the decision to undertake automatic speech recognition solutions in the Medical context. Turnaround time which represents the time from initial audio recording to the production and dispatch of the final report or charting constitutes one of the major costs involved in medical

recordkeeping. Potential for reduction in turnaround time has been identified as one of the prime motivators in the adoption of speech recognition technology in radiology. Delays in completing transcription can directly impact on patient care. The ability to generate timely and detailed communication helps to ensure coordinated and consistent patient care.It has been estimated as of June, 2000 that 5% of Physicians use speech recognition in the USA[27]. In the setting where the doctor directly dictates using automatic speech recognition the potential cost savings are significant by eliminating the need for third party transcription completely.

Cost savings considerations must take account of the additional time required for a physician to undertake proofing. Current speech recognition solutions in healthcare allow professionals to achieve recognition rates of the order of 97 to 98% in speciality specific domains such as radiology, orthopedics and ophthalmology. This nonetheless implies that in a typical 300 word report, approximately six to nine errors will require correction with an inherent time and cost implication. The Physician must locate, identify and undertake suitable correction for each error.

Program manufacturers will recommend specific hardware requirements, which are frequently aimed at entry level specification. Use of higher performance hardware can significantly improve overall success. Use of domains specific speciality language models which rank the likelihood of word combinations are used in professional versions of automatic speech recognition and significantly enhance recognition rates. Error free documentation despite flawless recognition cannot be achieved with normal speech in practice as result of mispronunciations and dysfluencies.

Adhering to the proper training protocol and dictation style contribute to higher recognition rates. The use of voice activated macros and templates can very significantly improve overall documentation. Entry of details such as names and addresses for example is often best achieved through data retrieval through a database management system, but can be achieved with high degrees of accuracy using voice recognition systems structured for such inputs.Assistance with training installation and support forms part of the usual package available from certified speech dealers and contributes greatly to the likely outcome of success. Understanding the importance of individual user's settings and voice backup requirements is paramount in ensuring ongoing success. Use of certified speech dealers is particularly desirable for those who have little computer based experience.

Automatic speech recognition providing transcription solutions also incorporates navigation capabilities where voice- input provides desktop control. This is specifically useful in the context of disability as previously described**.** Acceptability of speech recognition systems within the healthcare domain increases with the increases in performance, speed accuracy and large specialized vocabularies using continuous speech recognition.All major computer operating system producers such as Microsoft, Apple, IBM and Sun incorporated speech recognition as a standard feature in their operating systems.

User preferences in composing text can depend on many factors including the user's preferences in normal composition of text. Characteristics such as linear or nonlinear cognitive styles can play a significant role in their use of speech-enabled systems.[28] Graphic user interface and voice-input may further the efficiency of structured data entry in practice.Structured data entry is most readily realized by defining relevant entities

which are linked with corresponding attributes and values. The Galen approach utilizes the structured data type entry model and also incorporates the ability for more easily documented compositional concepts. It provides this through a terminology server which retrieved all possible relevant compositional options. Ongoing structural and design concepts in the GALEN project are part of the telematics program of the European Union. Its primary initial focus has been in the area of diabetes.

Healthcare professionals are familiar with computer-based developments of area specific database implementations. The ability through the use of compatible application programming interfaces to use speech recognition technology directly within such applications greatly enhances the appeal for speech deployment in these areas.Examples include EPR, Laboratory information systems, hospital information systems, radiology information systems and DBMS in primary care. Modern database are amenable to voice-input without the need for the user to adopt his style of the input to a format with which they feel unfamiliar or limited.

6.1 Voice input devices

Quality of acoustic input is a critical component of successful speech and language processing. Systems are increasingly more tolerant of poor quality input as the technology advances. High quality audio inputs however maximise the performance of automated speech recognition. Most high quality headsets utilize noise counselling microphone elements. In lower frequency ranges [100 Hz] these microphones can be highly effective reducing ambient noise by 25 decibels relative to the speaker. As frequencies increase however noise cancelling effectiveness is attenuated significantly and typically for noise above three kilohertz [3 kHz] frequency range attenuation is lost. The latter frequency range is typical of cooling fan noise associated with computers. Frequency response between 100Hz and 6 kHz corresponding to normal speech is the range needed by speech engines. Close-talking microphones attempt to maximize the primary voice based audio signal based on close proximity to speech source. Room acoustics have also been demonstrated to impact speech recognition accuracy. Sound reflection particularly off hard floors, walls and ceilings result in reverberation and echo effects which will degrade recognition accuracy.

There is a variety of options through which sound input can be achieved for processing in automatic speech recognition scenarios.

Standard Headset

Fig 1: Standard headset

The most traditional sound input is achieved through the use of near user microphones and a wired headset connecting through a sound input unit or soundcard to a personal computer. The headset may also comprise the sound processing unit as in the case of USB based headsets which connect via the high-speed data input ports capable of supporting high data throughput allowing high quality audio input can be achieved. This configuration necessitates a physical connection between user and PC with high visibility which may limit acceptability in normal circumstances. Arrangements for the cable management to avoid disruption of desktop are a further consideration.

Wireless Headsets

Fig 2: Wireless headset with recorder

With the advent of increasing use of wireless technology, the use of wireless headset is increasingly being used and will ultimately replace the more traditional and intrusive wired headset. These afford the user greater freedom and depending on the precise wireless technology used do not require close proximity to the PC. The practicality where proofing is undertaken at dictation may reduce the benefits.

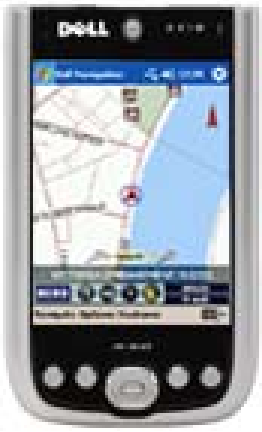
The use of digital handheld recording devices with wireless capabilities offers another means of sound recording which capture speech recording. This allows for later transferred to wireless network of other device. This arrangement with wireless headsets maintains the advantage of a near sited microphone with consistent proximity to the speech recording source.

Directional Digital Array Microphones

Fig 3: Directional Digital Array Microphones

The use of desktop mounted microphones in the form of single or multiple directional microphones arrays is a further option allowing the user to dictate without a near sited microphone.. This configuration like a wireless headset use allows recording to occur without physical connection but in this instance without any physically attached near sited microphone. This configuration has particular potential application especially with circular array setup in the setting of meetings, and refinements in this technology are continuously advancing in their tolerance of pickup of in phase conversation even in the setting of background noises and interference. Sound source localization SSL by these devices can be highly accurate (Ref) enhancing the capability for automatic speech recognition applications to transcribe multiple user input scenarios.

Digital recorders and Handheld Devices



Using handheld recorders is familiar to many healthcare professionals and is used as the preferred method for recordings for transcription. It provides an eyes free setup which allows the user to concentrate on the subject matter. Most modern recorders allow for index marking which greatly facilitates the ability to navigate to a dictation of specific interest. Digital recording permits rapid playback and insertion features not possible using audio tape. Sound quality can be an issue when using handheld recorders as they do not specifically filter background noise. Distance from the speaker can degrade voice quality and subsequent automatic speech recognition performance.Their current is primarily in backend transcription solutions as described. Many handheld devices including phones and pocket PCs or PDA integrate voice recording of sufficient quality for ASR.

A wireless capability of many handheld devices has simplified transfer of audio files to networks or local devices for further processing.

The choice of sound recording technique will be determined by consideration of the workflow in the environment in which documentation or transcription is undertaken. The model setting for direct front-end input to the personal computed by voice is the quiet intrusion free environment with negligible ambient noise. The setting of traditional radiology represents a good example of this scenario and is one area in which automated speech recognition systems have been extensively implemented. Similarly in the area of histopathology data recording is undertaken in environments conducive to direct speech recognition.

6.2 Assisted transcription and backend processing

For many Healthcare scenarios the uses of front-end ASR speech to text input involving the professional directly sitting in front of a PC may not represent the ideal or indeed practical solution in practice.Possible dictation solutions include handheld devices and high quality telephony links to the speech servers over DSL or other broadband connection. These devices generally do not achieve the levels of acoustic accuracy of headsets with near sited microphones. These apparent disadvantages can be overcome through the combination of back end processing using automatic speech recognition to assist the transcriptionist. In this context the transcriptionist essentially proofs the document by listening to the audio file while overseeing the correction of any recognition errors by automated speech recognition. Experienced operators in specific domains can achieve very high levels of accurate output using this format. Using ASR transcription applications in this setting does permit changes in the user's voice profile to be updated

based on effort correction undertaken by the transcriptionist. This results in continuing improvement in performance without any direct input from the narrator.

The use of telephone with encryption allows for secure transmission of dictated medical details amenable to this method. This allows for backend processing over the Internet by a remote or locally based transcription service provider. Outsourcing provides the opportunity to free local resources and the choice of implementation model will depend on cost considerations to a large extent in back-end ASR based transcription.

6.3 Pervasive computing and ASR

Pervasive computing refers to the concept of mobile communications associated with location independent connectivity and is becoming a key part of all information intensive environments.The necessity for continuous access to information which is independent of location and access device used , is becoming increasingly a reality in business and healthcare.

The ready availability of personal digital assistants with wireless capabilities including WiFi, Bluetooth and wireless broadband demonstrate the reality of pervasive computing in everyday use. Smartphones are part already of this technological implementation in general use providing web and e-mail access in addition to capabilities for linkages to corporate and wide area networks. The addition of speech-enabled input to devices used in pervasive computing overcomes many of the disadvantages of their reduced size, which limits ease of data entry. The hands free, eye three aspects of speech- input represents a major advantage as these technologies are refined.

6.3.1 Multi-modal interface

Multi-modal interfaces are those which are not limited to a single input modality allowing a number of different options for data input which may include voice input with speech recognition, visual and tactile input through for example in built microphone, keypad, touch-screen or mouse. The concurrent use of these different modalities of input can significantly enhance data entry on mobile devices. Multi-modal entry may simplify tasks such as data retrieval, form filling or document review by including a speech based command and entry capability.

In the Healthcare environments the ability to access information on the move and independent of location has clear benefits. The ability for example to enter details such as an electronic prescription into a handheld device with immediate wireless relay has obvious potential benefits. The integration of such devices and telephony provide for the transmission of electronic data to a variety of destinations including pharmacies in the case of electronic prescribing. The ability to retrieve patient information independent of location through the use of pervasive computing will become increasingly available both at institutional and community level.

6.3.2 Mobile speech implementation

Connectivity of applications on mobile devices provides one of the most powerful applications in mobile computing. Clients are wirelessly connected to servers which can deliver data to mobile applications through compatibles services. Connectivity is achieved

through WAN, IEEE 802.11 based wireless local area networks or Bluetooth depending on proximity to the data server.

In common with client and server database configurations, three main possibilities exist for speech-enabled mobile database connectivity:

Speech engine on the server

In this scenario a voice channel is established to the server on which speech recognition and others speech-enabled technologies run. The voice input device or voice portal solution is typically a phone with no client side application requirement.

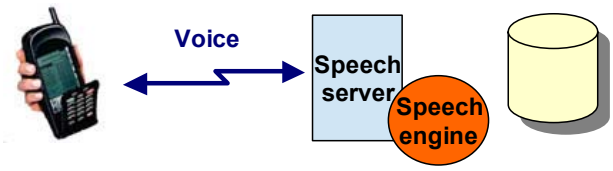


Fig 4: Speech Engine of the Server

Because no processing requirements exist at the voice portal this represents a thin client setup. Vocabulary engines and automatic speech recognition processing with natural language processing can be run on the server side. Dependency on the quality of the speech channel is inherently part of this configuration.

Speech engine on the client

This configuration requires that the speech engine runs on the client device which therefore requires significant processing power and memory capabilities. It can be implemented in an "always on" type connection often involving HTML data transfer using a voice XML application on the client device. Where the client is intermittently connected voice processing occurs using the local speech engine prior to bidirectional synchronization of data during connections.

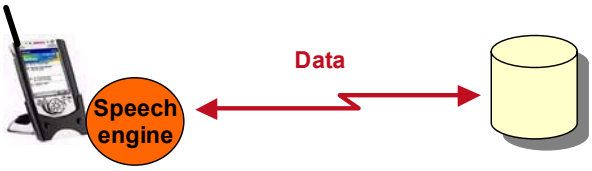


Fig 5: Speech Engine of the Client

Strategic speech engine – client server

As in standard database configurations in a distributed setting processing occurs on both the client and server sides. The main recognition work is on the server side where processing capabilities can be maximized. This can be used in the speech recognition STT context or in the TTS.

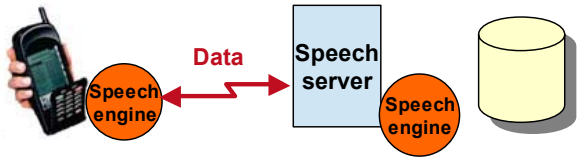


Fig 6: Speech engine – client server

6.3.3 Digital signal processors

Interactive voice response systems can be deployed using ASR through a hardware solution based on embedding a digital signal processor DSP as a dedicated speech engine based enhancement. This can improve the performance of solutions both for the developer and user. This contrasts with the commoner implementation in which the central processing unit CPU undertakes the automatic speech recognition function representing a software based solution. When utilized higher levels of accuracy and functionality are likely to be seen in voice-driven systems. In the configuration using DSP based automatic speech recognition the requirements for speech recognition processing occur at hardware level enabling the realization of smaller more compact systems. Currently DSP capabilities can be added as an ad in without disturbing existing hardware setups and freeing up CPU based processing for application channels. By porting DSP capabilities into existing IVR systems, existing PBX systems, and dedicated messaging platforms and others similar systems can enhance performance through the addition of ASR capability.

6.4 Successful ASR transcription in Healthcare

6.4.1 ASR Transcription in Cardiology

Caritas Christi Health System in Boston has implemented speech based solutions in a number of areas with significant reported benefit to date using ASR transcription.

In the cardiology setting a system whereby the physician's assistant dictates details during ward rounds onto a laptop is used. Dictation is undertaken in the nurse's station providing a superior acoustic environment, and recognition accuracies of 98% have been reported in this scenario. Through wireless connection with the hospital database, following review by the cardiologist, text based data entry can be completed into the existing electronic patient record. Benefits reported include enhanced level of detail in documentation which impacts patient care. Improved documentation of procedures augments the efficiency of billing and thereby the overall cost saving and return on investment.

6.4.2 ASR Transcription in Radiology

It is clear that specific specialties in medicine have implemented automatic speech recognition based transcription more than others. Radiology provides an environment highly conducive to the use of ASR by virtue of its favourable acoustics, relatively intrusion free nature and relatively limited specialist vocabulary. Specialized vocabularies optimize the performance of speech recognition and quiet environments improve the quality of audio signals on which recognition is based.

Many reports of cost savings have been documented for systems deployed in radiology.[29,30] As of July, 2003 30% of radiology practices have either implemented voice-recognition systems or are in the process of planning implementation[31]. The increase in radiologist's time in direct involvement with transcription has however been reported to prolong report generation in some instances.[32]. Factors requiring consideration are addressed detail in the successful implementations below. The changes in workflow from traditional and automatic speech recognition based transcription are outlined in Fig 7 & 8. These workflow stages readily transfer to many other disciplines in healthcare.

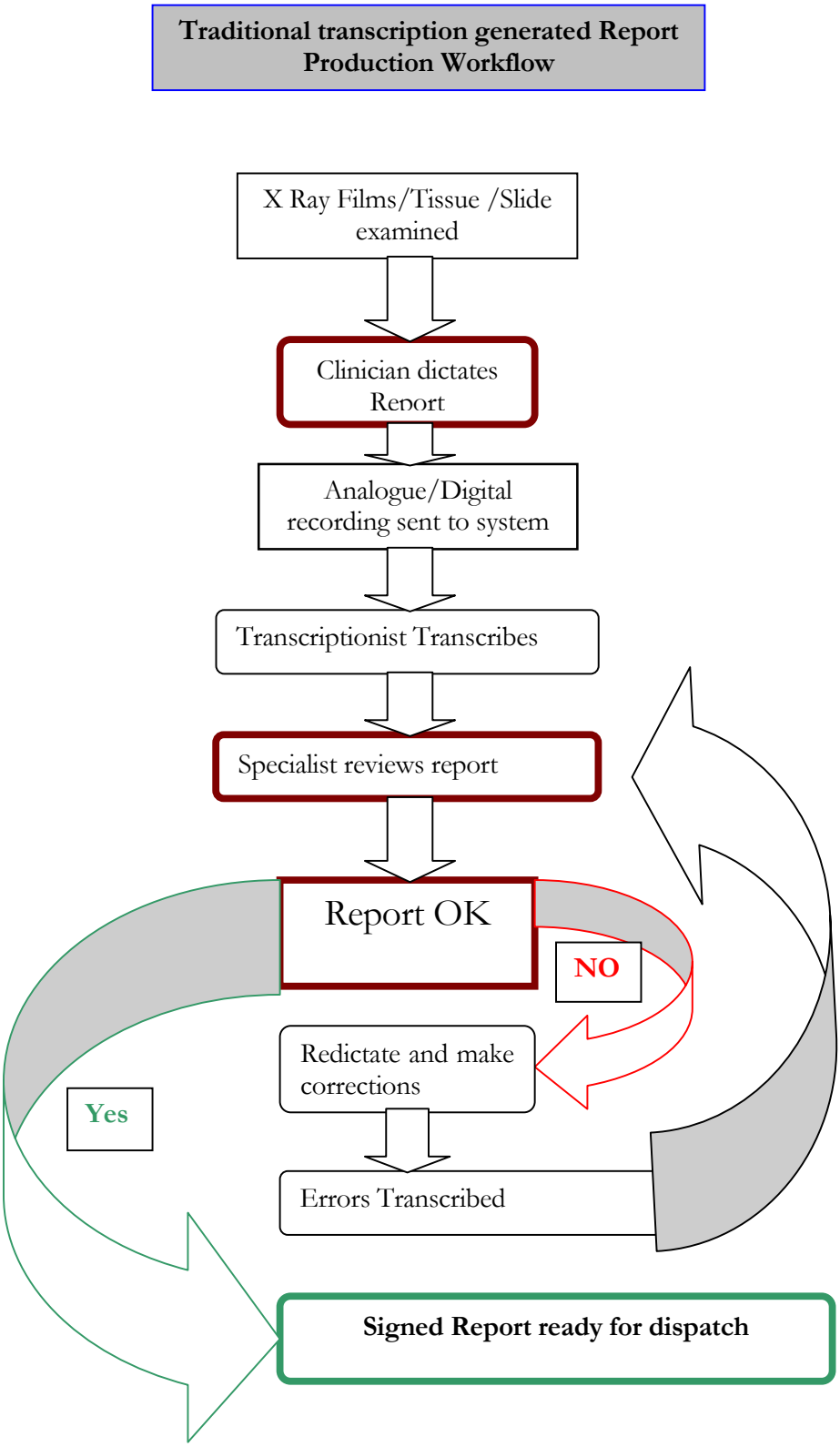


Fig. 7: Traditional Transcription Report Workflow

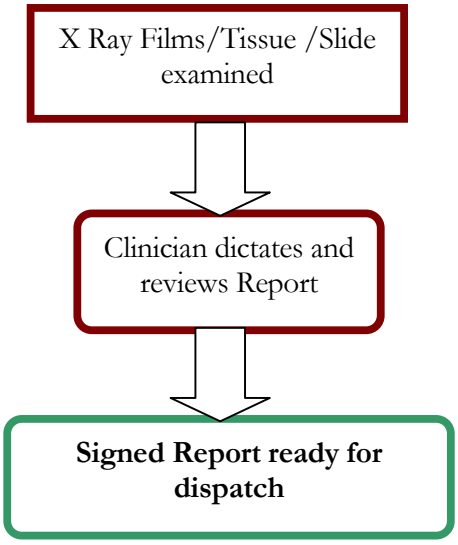


Fig 8: ASR transcription generated Report Production workflow

6.4.2.1 ASR transcription at Massachusetts General Hospital in Boston[29]

Project planning

Hospital administrators are faced with the reality of rapidly changing technology within the healthcare environment. Many claims relating to improved patient care and efficiencies are made and decision-making must be based on sound analysis of patient care considerations balanced with return on investment ROI. Technology implementations require the acquisition of hardware and software, but equally acceptance of the effective use of technology depends on the user interface. Because of the major implications on workflow, a detailed and comprehensive assessment of vendors is essential. In the case of radiology the primary focus is on the radiologist as user together with and ancillary staffs involved in the changes in workflow. Minimising disruption and maximizing efficiencies of the technology are the goal.

Massachusetts general hospital in Boston implemented voice-recognition over a two year commencing 1998. An annual saving a 530,000 dollars and a 50% decrease in report turnaround was achieved. It was concluded following study of this implementation that four areas in particular contributed to success.Project planning involving a plan which clearly outlined clear goals achievable with voice-recognition coupled with a timeline which was to be closely adhered to.

Through this approach 75% of departments adopted the use of voice-recognition, reduced turnaround times from 4.3 to 2.0 days and achieved more than half a million dollars savings per year.

Training and support were considered vital to the success of Voice-recognition. Training materials to assist all instructions most be developed and follow-up implementation surveys were recommended as regular time intervals. Appendix 1. Measurement quantifying the degree of success or failure with voice-recognition was found to provide positive and negative feedback to the team responsible for voice-recognition implementation. He provided information as to where retraining resources should be focused. It demonstrated details of the actual cost savings and provided upper level management with future recommendations.Sustainment was essential in order to maintain success achieved using voice-recognition. Feedback relating to success is problems and changes were communicated to all department levels to provide inclusive input from all users.

Senior departmental leadership was seen as essential in the implementation of the Massachusetts general hospital voice-recognition project. They were part of the project leaders designation process since leaders must oversee project deadlines, focus on all operational areas in order to achieve high satisfaction with minimum conflicts. This was achieved through the voice-recognition team who provided support education and guidance as required.

Education and pocket guides were developed to assist radiologists in implementing voice-recognition. Appendix 1. A dedicated training specialist was available as necessary and updated teaching materials were distributed. Technical support systems were made continuously available with a standard response time of twenty minutes.

The operational implementation required consideration of hardware setups and methods of providing continuing support. Estimations of dates and locations requiring voice-recognition were implemented. These areas were targeted individually

Hardware specifications which had been established by the voice-recognition team including the number and configuration of workstations together with total compliance with manufacturer's specifications was scrupulously undertaken. It was necessary that all workstations were networked to the radiology information system to allow patient based data integration. Client configuration required software installation together with voice-recognition application hardware and microphones. Removal of existing audio typing systems was undertaken. The initial implementation was undertaken in the Teleradiology department. Following this it took six months to train 72 staff radiologists, 23 clinical fellows and 29 residents. 60 workstations were deployed throughout the department for the purposes of voice-recognition.

It was appreciated that the tangible benefits including reduced cost and decreased turnaround were not immediately evident to radiologists. It was however incentivised by management on the basis that cost savings could be traded for other benefits.

The training undertaken in Massachusetts general hospital radiology department was pivotal in determining the success of the automatic speech recognition project. Training sessions lasted about one and a half hours and consisted of user enrolment, education and additional practice. Individual steps in creating a dictated report were addressed during training.

Additional training reinforces the effectiveness of the technology to the user and continuing use results in increases in speed and accuracy. Once adequately trained users could then provide assistance to other learners.

It was agreed that in a crisis situation where dictation using voice-recognition was not achieved within 45 minutes that temporary use of the conventional dictation system was allowed.

Accuracy of dictation depended on satisfactory enrolment whereby users names and passwords were initially assigned. A minimum of 50 or maximum of 200 pre-planned sentences were used. The system created a voice profile based on the enrolment process at this time.

The use of macros and templates greatly decreased the time necessary to generate reports. Trigger words or phrases were used to insert appropriate text based macros. A list of more than 100 institutional macros was compiled particularly for normal reports of for example chest or spine. In MGH this was achieved by the radiologist transferring a macro to their personal macro list. Templates were similar accessed and provided a form layout without pre entered text.With the above supports in place practice was the means by which radiologists could further enhance their skills. The professionals were more likely to use the technology with firsthand experience of its efficacy.

A methodical analysis of the performance of the system was measured on an ongoing basis using six primary areas. Appendix 1

- names, numbers and percentages of radiologists trained on VR,
- technical and educational VR problems,
- penetration by division and radiologist,
- percentage compliance with VR vs. the former system,
- report turnaround time, and
- Annual cost savings.

These reports were run on a six monthly basis and statistical analysis performed. A database of the training status of physicians was maintained which provided an assessment of progress and voice-recognition penetration. The problem log was also maintained and if difficulties were found to be a training issue additional appropriate support was arranged.

Penetration

Penetration was measured on an individual radiologist and division basis. Analysis guided the requirements for further training by focusing on low penetration areas. In this analysis that percentage of reports generated using voice-recognition together with the details of the physician and division were collected. The twentieth upper and lower centiles where then reported again allowing focus of training and additional resources on those achieving poorest (and best)voice- recognition based success.

A satisfaction survey was undertaken Post-implementation with a view to identifying problem areas requiring improvements in training support and functionality. Appendix 1. The questionnaire was assessed on the four-point ordinal scale aimed at eexamining key

areas relating to training and materials. Response times for problems experienced were assessed and open ended questions permitting clear responses of approval or disapproval were included. Frequency of use of macro/ template was assessed as this was considered to relate to utilization rates.

Valuable information was gathered with regard to wide variety of aspects to the implementation of voice-recognition. Survey results were prepared in graphic format, distributed widely and contained detailed explanations and proposed future actions

In June, 1998 75% of radiologist felt that training and certification was effective. And 43% approved of the shorter turnaround times, while 21% liked the ability to immediately sign of reports.

By grading the voice-recognition on a four-point ordinal scale no respondents were totally satisfied, 29% were mostly satisfied, 48% somewhat satisfied while 16% were not at all satisfied. There has been very significant improvement in automated speech recognition products since 1998.

By July, 1991, 900 of 1200 reports were dictated by voice-recognition., with some departments achieving 93% report generation by ASR. By 2000 58.06% of radiologists dictated over 75% of their reports using voice-recognition. 42.7 percent dictated more than 90% of their reports using the technology.

Report turnaround time

This is a paramount consideration for radiologists given that the speed of dispatch of reports has implications for patient management. Pre Implementation data provided a benchmark in February, 1997 it took 4.3 days reducing to 2.19 days by July, '99 representing a 51% reduction.

6.4.2.2 Lehigh Valley hospital [30]

In a study at Lehigh Valley hospital, Allentown, Pennsylvania, USA the compelling reason voice-recognition was implemented was the fact that initially in November, 1998 only 41% of reports were completed within 24 hours. Every voice recognition system on the market was reviewed bearing the radiologists workflow in mind. A business plan and funding was arranged. This implementation of involved fifteen locations as the Lehigh Valley Hospital and Health Network (LVHHN) which was a multi-site centre that performs over 360,000 procedures annually. It's incorporated 40 radiologists and 200 fulltime employees FTE's. Planning and testing took approximately four months and estimates of cost saving and reduction in FTE's exceeded expectations. Transcription staff were not displaced but were reassigned to other transcription duties. The proposed ROI period was seventeen 17 months but was in fact less than twelve months in practice. The goal in reducing outsourcing services by $670,000.00 over three years realized a saving in excess of $900,000.00 in practice.

At the time of publication report completion in using voice recognition within 24 hours varies between 85 and 92% which represented a major impact in the time the preparation of reports.

6.4.3 Other Healthcare Specialities and ASR

In other clinical setting the documentation process differs significantly and occurs coincident with contact with the patient or very soon thereafter. Use for example in an accident and emergency setting is in sharp contrast to that of office based documentation. In 2003 a survey of all emergency departments in the USA indicated uptake of ASR based transcription was only 7%.[33]Effective implementation has been realized in the Accident and Emergency setting.34 Timely recording coincident with patient contact is essential to ensure consistency and reliability record keeping. Delays[3] in recording details correlate directly with the probability of omission of significant details. Poor quality records are a frequent a source of clinical errors in terms of treatment and management.

Considerations such as the privacy and confidentiality[35] of the detail recorded must be appreciated in most clinical scenarios. These factors will inevitably impact and on the practicality and usability of automatic speech recognition. The latter however does not in any way necessarily indicate the potential for use of this technology. The ability to and individual words was part of the system implement it.

Training certification was undertaken and considered important because it allowed compilation of an accurate database of all Physicians denoting their strengths and weaknesses. This was completed for each user after every training session.

6.4.5 Ten years experience of ASR transcription and EPR

Automated speech recognition based transcription has as intuitive and instinctive feel based on its use of speech as the primary input modality for record and document generation. Through this technology conversion of speech to text achieves near instantaneous realization. Expectations of speech recognition technology essentially failed the reality of use, particularly in the early nineties when the only speech recognition applications available consisted of discrete word recognition software packages. It has been stated that we have been within two years of full implementation of automatic speech recognition based transcription for the past ten years.

Limitations in our experience of more than ten years and ASR impacting on the use of this technology were not only related to software design. Hardware posed significant challenges through limited processing capabilities and complex configurations. Coupled with this, standard entry level configurations required significant enhancement to cope with the processing capabilities required for automated speech recognition. These enhancements including high performance sound cards and particularly additional memory requirements. This presented significant additional expense at a time when random access memory capacities were limited and expensive. Compatibility of headsets with sound cards constituted one aspect of difficulties in the early implementation. Solutions to such configuration problems were frequently empirical. Hardware compatibility issues between computers and sound cards presented another layer of potential conflict and frustrated the attempts of many to implement speech recognition. Device drivers compatibility issues

with hardware and operating systems together with hardwired jumper settings challenged even the most experienced and enthusiasm computer user.

Initial experience with discrete word recognition packages, while demonstrating the capability for speech to text conversion, were readily recognised as technology with future rather than current potential. In using discrete word recognition software we observed prepositions and single syllable words were frequently difficult to achieve consistency with and keyboard entry was often utilised in reality thus negating the benefit of automated speech recognition. It was observed nonetheless that at this point the technology could provide benefits to those who may have impaired or limited abilities in using keyboard entry. It was clear that although tedious and time intensive document generation using discrete word recognition was achievable. Comparisons in terms of time input were based on one's own typing .

It was following the advent of continuous speech recognition software availability that full implementation in practice was implemented.. In keeping with successful implementations outlined earlier in this paper, following initial experience with discrete work recognition packages, a review of the principle voice-recognition packages available at this time was undertaken. These included DragonNaturally speaking, Voice Express from L. & H, IBM ViaVoice professional edition and later speech recognition bundled with the office XP and Microsoft XP professional. Undertaking this review did present significant financial issues but the initial potential of the technology, coupled with anticipation of cost saving with respect to additional staffing, was the primary incentive in persevering.

While working in a busy practice the absolute necessity for reliability and accuracy in speech recognition was imperative. In assessing the most viable transcriptions solutions the principal determinants were accuracy of recognition coupled with ease of correction techniques. Speed of the processing was a paramount consideration as rapid recognition rates were essential if this technology was to be integrated into a demanding real world healthcare environment.

In the late nineteen nineties automatic speech recognition based transcription completely supplanted all other methods of normal document preparation comprising referrals, formal legal reports and all detailed document preparation. It was not widely used for data entry relating to normal clinical encounters in primary care. Up to 2002 patient encounters were documented using a DOS based DBMS. In general normal brief clinical encounters were documented through keyboard use incorporating abbreviations and keyboard macros implementations. . It was our experience that use of all voice-recognition applications was very limited in performance in DOS based applications. The ability to paste text into this or any other application, effectively overcame this limitation. It was noted that documentation within the native word-processing application bundled with the voice-recognition application that he's the best recognition rates with greatest performance of correction techniques. Text was then easily transferred from this word-processing application to the practice management database.This was being implemented within a practice setting which had been using a database management system as the sole recordkeeping modality for the preceding ten years. In this way it was possible to link electronically prepared documents with corresponding patient files.

6.4.5.1Use of voice macros

The standardization of voice macros in terms of the phrases used to invoke the macro has simplified use. While the voice macros used was essentially limited to text based and formatting input, the ability to incorporate macros which interface directly with application programming interfaces of a variety of applications is already established and reliable. Standardizing macro terminology included the use such as the use of "Referral to Doctor Jones, referral to accident and emergency, referral to orthopaedic outpatients", allows the use of intuitive terminology to invoke macros. This is in contrast to keyboard macros where key combinations frequently have little or no intuitive relation to actually use. There are a host of dedicated keyboard based macro applications. Many of these provide powerful integration within commonly use applications. The user with access to automated speech recognition has now the option to utilize any or all of these macro technologies for timesaving and automation of tasks with ASR.

7.0 The future and automatic speech recognition in healthcare

Automatic speech recognition based transcription constitutes the leading application of speech recognition technology in healthcare to date. Changes in work practices are inevitable when implementing voice recognition and this creates a natural resistance to the introduction of such new technology. It is not surprising that users with ready access to experienced and competent transcription services are unlikely to feel motivated to change. Practical and financial benefits in introducing a new technology must be evident to the end user if successful implementation is to be achieved. They are a number of areas which warrant further study in the context of the ASR based transcription solutions. Direct front end versus back-end solutions ASR implementations differ markedly. Given the variety of factors in clinical settings which interact in determining the ultimate viability of ASR based transcription further work is needed to clarify key issues. The existing workflow sequence in terms of how, where and when clinical documentation occurs deserves special consideration.

Review of the literature over ten years reveals a trend towards realisation that ASR technologies have steadily evolved from potential to real world applications. Its viability in selected areas conducive to the use of speech recognition particularly radiology has been repeatedly demonstrated. Implementation models in these areas provide a template for its migration into many other healthcare areas. It is evident that significant improvements in this technology will continue to increase it ease of use and thus acceptability and penetration.

Voice input is an intuitive and fundamental means of communication and will supplant other modalities of data entry and command and control functions. Multimodal entry will complement voice enable functions as the primary input mode. This has particular relevance when used in association with overcoming limitations associated with disabilities.

Benefits in healthcare delivery, communications, data integrity and costing saving through the use of transcription and interactive voice response systems and ASR have been outlined. The decision to undertake implementation of such technology must take account of the challenge presented to the end user in addition to the need for on site ICT knowledge and support for speech applications. The broader implications for healthcare are highly significant. This is especially so in the context of healthcare organisations and government based departments with responsibility for maximising efficiency and return on investment. Transcription and IVR systems using automatic speech recognition are for the most part a non recurring cost while the cost saving are perpetual.

The greater volumes of healthcare services are delivered outside the hospital environment and are widely dispersed. Continuous secure access to medical information provided in conjunction with pervasive computing can deliver mobile access to multiple information sources independent of location. Through voice verification technologies security can be integrated with data access. This is not exclusive of existing network based security and access protocols. Such configurations permit highly secure infrastructures through which medical information access and exchange can be realised on an “always on” basis.

The convergence of automatic speech recognition and voice enabled technology as modalities for secure data processing and control in a hands free, eyes free environment assure its central role in healthcare into the future.

Bibliography

1) Sumby and Pollack (1954Sumby, W. H., & Pollack, I. (1954).
Visual contribution to speech intelligibility in noise.
Journal of the Acoustical Society of America, 26, 212–215.

2) McGurk and MacDonald (1976) McGurk, H., & MacDonald, J.(1976) Hearing lips and seeing voices Nature, 264, 746–748.

3) Data should be acquired as close to the source as possible.
Van Bemmel J, Musen MA. Handbook of medical informatics. Heidelberg: Springer, 1997:34.

4) URL:http://www-306.ibm.com/software/voice/viavoice/

5) URL: http://www.scansoft.com/naturallyspeaking/

6) URL: http://www.speechrecognition.philips.com/

7) VoicePad Platinum [computer software]. (1997). Waltham, MA: Kurzweil Educational Systems, Inc.

8) 8 Youdin, M., Sell, G., Reich, T., Clagnaz, M., Louie, H., & Kolwicz, R., (1980).
A voice controlled powered wheelchair and environmental control system for the severely disabled. Medical Progress Through Technology, 7, 139-143.

9) Damper, R. (1984). Voice-input aids for the physically disabled.
International Journal of Man-Machine Studies, 21, 541-553

10) Coleman, C. & Meyers, L. (1991). Computer recognition of the speech of adults
with cerebral palsy and dysarthria. Augmentative and Alternative Communication, 7,34-42, 1991.

11) DragonDictate [computer software]. (1990). Newton, MA: Dragon Systems.
Ferrier, L.J., Jarrell, N., Carpenter, T., & Shane, H.C. (1992). A case study of a dysarthric speaker using the DragonDictate Voice Recognition System. .

12) Doyle, P., Leeper, H., Kotler, A., Thomas-Stonell, N., OíNeill, C., Dylke, M., & Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. Journal of Rehabilitation Research and Development,34, (3), 309-316.

13) Trapagnier, C. & Rosen, M.J. (1991). Towards a method for computer interface design using speech recognition. Paper presented at RESNA 14th Annual Conference, Kansas City, MO.

14) Treviranus, J., Shein, F. Haataja, S., Parnes, P., & Milner, M. (1991, month unknown). Speech recognition to enhance computer access for children and young adults who are functionally non-speaking. RESNA 14th Annual Conference, Kansas City, Mo.

15) Watson, C., Reed, J., Kewley-Port, D., & Maki, D. (1989). The Indiana Speech Training Aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality. Journal of Speech and Hearing Research, 32, 245-251.

16) Mahshie, J., Vari-Alquist, D., Waddy-Smith, B., & Bernstein, L. (1988). Speech training aids for hearing-impaired individuals: III. Preliminary observations in the clinic and childrenís homes. Journal of Rehabilitation Research and Development, 25,69-82.

17) Bernstein, L. (1989) Computer-based speech training for the profoundly hearing impaired: Some design considerations. Volta-Review, 91,19-28.
Bernstein, L., Goldstein, M.,& Mashie, J. (1988). Speech training aids for hearing-impaired individuals. Journal of Rehabilitation Research and Development, 25,53-62.

18) Ling, D. (1976). Speech and the Hearing Impaired Child: Theory and Practice. Washington DC: The Alexander Graham Bell Association of the Deaf, Inc.

19) De La Paz, (in press). Composing via dictation and speech recognition systems: compensatory technology for students with learning disabilities. Learning Disabilities Quarterly.

20) Kurzweil 3000 [computer software]. (1998). Waltham, MA: Kurzweil Educational Systems, Inc.
Lange, H. (1993). Speech synthesis and speech recognition: Tomorrowís human-computer interfaces? Annual Review of Information Science and Technology (ARIST), 28,153-185.

21) Markowitz, Judith. “Voice Biometrics - Are You Who You Say You Are?” December 2003. http://www.speechtechmag.com/issues/8_6/cover/2751-1.html

22) http://www.voicexml.org/

23) http://www.saltforum.org/

24) Source: eHealthCoach.com, November 2000, Vol. 1, No.10

25) Source: Medical Transcription Industry Association.

26) Source: Robert Lowes, Ed. Bits & Bytes. Medical Economics. 5 June, 2000

27) Bergeron BP. Voice recognition: an enabling technology for modern health care? In: Cimino Joe. AMIA Annu Fall Symp: Hanley & Belfus, Inc, 1996:802-806.

28) Van Ginneken AM. The Structure of Data in Medical Records. In: van Bemmel JH, McCray AT, eds. Yearbook of Medical Informatics 1995,

29) Projects Planning, Training, Measurement and Sustainment
Sharon Antiles, M.P.H; John Couris, M.S.M.; Alan Schweitzer, M.E.E
Date:01/01/2000 http://www.ahraonline.org/about.htm American Healthcare Radiology Administrators

30) Radiology Management 2001 May-Jun;25(3):42-9.
Comment in: Radiology Management 2001 Mar-Apr;23(2):23-5;discussion 26-7.Success with voice recognition.
Sferrella
Lehigh Valley Hospital, Allentown, Penn., USA. sheila.sferrella@lvh.com
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12817421&dopt=Abstract

31) J Thorac Imaging. 2003 Jul;18(3):178-82.
Voice recognition.
A, McLoud TC.
Harvard Medical School, Massachusetts General Hospital, 55 Fruit Street, Boston,MA 02114, USA.

32) Radiology Management 2001 Mar-Apr;23(2):18-22.
Comment in:Radiology Management 2001 Mar-Apr;23(2):23-5;discussion 26-7.
Radiology report production times: voice recognition vs. transcription.
Gale B, Safriel Y, Lukban A, Kalowitz J, Fleischer J, Gordon D.
State University of New York Health Sciences Center at Brooklyn, USA.
briangalemd@pol.net

33) Acad Emerg Med. 2003 Aug;10(8):848-52.
Information technology in emergency medicine residency-affiliated emergency departments.
Pallin D, Lahman M, Baumlin K.
Departments of Emergency Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02215, USA. dpallin@partners.org
Am J Emerg Med. 2001 Jul;19(4):295-8.

34) Voice recognition software versus a traditional transcription service for physician charting in the ED.
Zick RG, Olsen J.
Lutheran General Hospital, Department of Emergency Medicine, Park Ridge, IL,USA.

35) J Med Pract Manage. 2001 Jan-Feb;16(4):213-5.
Voice recognition in clinical medicine: process versus technology.
Bergeron BP.
Department of Anesthesia & Critical Care, Cln 309, Massachusetts General Hospital, Fruit Street, Boston, MA 02114, USA. archetype5@aol.com

Appendix 1 *VR Operational Implementation Plan – MGH Department of Radiology*

Figure 1
VR Operational Implementation Plan - Department of Radiology

The following is an operational implementation plan that detail all aspects to be considered and acted upon when targeting and implementing VR. This plan also includes the hardware implementation.

Operational area: Pediatric radiology (1unit) COMPLETE

Strategic relevance: Increased teleradiology and Shriners CR support

Implementation (hardware) target date: December 22, 1997

Operational implementation (training and certification completed) date: January 12, 1998

Actions Necessary for Implementation

Hardware/physical plan	Training/certification	Operational follow-up	Removal of prior system
1. ID physical space for additional unit. 2. Request network connection. 3. Install units.	1. ID physicians that need to be re-trained through total certification process. 2. Train and certify physicians in area.	1. Two week follow-up on physician competency on system. 2. Make any necessary changes to operation or additional training/	1. Has been removed.

Steps needed to complete or in process to move to the next operational area

1. Ensure that all staff in the area are trained and certified.
2. Validate that the system is integrating well within operations.
3. Ensure that the hardware and network connections are ready for the next operation area.

Operational area: Neuro (4units) CT/MR NEURO COMPLETE

Strategic relevance: Limited filming increased report availability on PCIS, could potentially assist with the referring physicians in transitioning to filmless.

Implementation (hardware) target date: December 22, 1997

Operational implementation (training and certification completed) date: May 22, 1998

Actions Necessary for Implementation			
Hardware/physical plan	**Training/certification**	**Operational follow-up**	**Removal of prior system**
1. ID physical space for additional unit. 2. Request network connection. 3. Install units.	1. ID physicians that need to be re-trained through total certification process. 2. Train and certify physicians in area.	1. Two week follow-up on physician competency on system. 2. Make any necessary changes to operation or additional training/	1. Has been removed.

Figure 2
Radiologist Quick Reference Pocket Guide

Logon Procedures

- Type your user ID
- Press the TAB key and enter password in lower case characters.

1. Select microphone and area type, check microphone volume level. Adjust so NO red. Green is good.

- Press the ENTER key.

2. Residents must select attending who will be last signer from the list

Accession Number Entry

1. Use the bar code reader to scan the accession number into the order number field or type the accession number.
2. Repeat above step to enter multiple accession numbers.

- Say "continue" or click the CONTINUE button

Microphone

- To activate microphone, use "F4" key or click on the microphone icon.
- The white microphone icon indicates on. The gray microphone icon indicates off.
- The microphone will automatically turn off after 30 seconds of uninterrupted use.

Dictation

- Say "begin dictation" or click icon

1. The screen will turn blue and a beep will sound indicating dictation can begin.

Figure 3
Competency-Based Evaluation Form

Steps	**Demonstrated**	**Evaluator's Initials**
1 Logging onto the PC		
Initiating the VR system		
1 Double click on VR icon if not active		
2 Entering user ID and password		
3 Enter examination order number		
Dictating a Report		
1 Test microphone volume		
2 Dictate new report		
3 View report in queue		
4 Turn on microphone (2 methods)		
5 Knowledge of F4 commands		
6 Knowledge of verbal commands		
7 Ability to dictate body of report		
8 Ability to pause the dictation while in dictation mode		
9 Putting microphone to sleep/wake up		
10 Create macros/templates		
11 Insert macros/templates		
12 Knowledge of how to edit customize screen		
13 Promptly saying "end dictation" when dictation is finished		
14 Turn off microphone (3 Methods)		
Editing a Report		
1 Highlight text to be edited		
2 Dictate or type corrections		
Completing a Report		

1 Differentiating between save as incomplete/preliminary		
2 Accepting and signing a report		
3 Logging off the system:"exit application"		
4 Knowledge of accessing "last text dictated" if VR problem		
Trainee's Name:	**Date:**	

Figure 4
Report on Penetration by Radiologist

Radiologist	Division	Reports VR	%VR	BI & Card	% BI &Card	Reports PS*	%PS*	Total VR+ PS
1	Musculo-Skeletal-Staff	384	98.7	0	0	5	1.3	389
2	Fellow	615	98.2	0	0	11	1.8	626
3	Resident	620	94.5	0	0	36	5.5	656
4	Pediatric Radiology-Staff	132	98.5	0	0	2	1.5	134
5	Neuroradiology-Staff	37	80.4	0	0	9	19.6	46
6	Chest Radiology-Staff	93	90.3	0	0	10	9.7	103
7	Musculo-Skeletal-Staff	257	99.6	0	0	1	0.4	258
8	GIGU Radiology-Staff	70	98.6	0	0	1	1.4	71
9	Mammography Staff	34	27.9	88	72.1	0	0.0	122
10	Resident	307	90.6	0	0	32	9.4	339
11	Resident	301	85.3	0	0	52	14.7	353
12	Fellow	168	95.5	0	0	8	4.5	176
13	Resident	352	97.5	0	0	9	2.5	361
14	EW Radiology-Staff	301	99.3	0	0	2	0.7	303
15	GIGU Radiology-Staff	125	100.0	0	0	0	0.0	125
16	Musculo-Skeletal-Staff	288	100.0	0	0	0	0.0	288
17	Resident	368	99.7	0	0	1	0.3	369
18	Fellow-Mammography	132	71.0	0	0	54	29.0	186
19	Neuroradiology-Staff	143	87.2	0	0	21	12.8	164
20	Fellow	444	99.8	0	0	1	0.2	445

Note:*Prior or alternative dictation/transcription system

Glossary

Biometric-based technology (also **biometrics**) technology that verifies or identifies individuals by analyzing a facet of their physiology and/or behaviour (e.g., voices, fingerprints)

Command and control using voice to issue commands to computer software, an IVR (interactive voice response system), or another piece of equipment (e.g., an automobile).

Continuous speech speech recognition is able to process naturally-spoken utterances: utterances that do not have pauses between every word.

Conversational interaction speech recognition that is designed to interact with a human through the use of voiced prompts. This contrasts with command and control.

Discrete speech speech recognition that requires a speaker to pause after each word. Also called *isolated word speech*.

Equal error rate a threshold setting in a product or system that results in an approximately equal percentage of false acceptance errors and false reject errors.

False acceptance when a verification system allows an impostor to get in. It may be the result of one match or it may be the result of more than one match within a single authentication interaction. For example, if a system prompts for a password and rejects the speaker the first time but then re-prompts for a second attempt which is successful.

False match refers to a single comparison of voiceprints. In verification it occurs when the voiceprint of an impostor is sufficiently similar to the voiceprint of the claimed identity to allow the system to erroneously accept the claim as authentic. In identification, a false match occurs when the system, in error, finds a voiceprint in the database that it deems sufficiently similar to that of the target voiceprint to believe that it has found a matching voice.

False non-match refers to a single comparison of voiceprints. In verification it occurs when the voiceprint of a legitimate user is judged sufficiently different from the voiceprint of the claimed identity to allow the system to erroneously reject the user as an impostor. In identification, a false non-match occurs when the system fails to find a voiceprint in the database for an individual when, in fact, there is one. In such cases the system erroneously reports that the person is not in its database.

False rejection when a verification system rejects a valid user. It may be the result of one match or of multiple attempts within a single interaction.

Keyword spotting speech-recognition technology that looks for specific words in what a person has said. For example, an IVR system to allow a naïve user to say something like "Give me the loan department please" when all it wants is the word "loan." Synonymous with “Word Spotting”

Speaker adaptive speech recognition cannot be used effectively by a speaker until that speaker has provided a sample of speech to the system. Virtually all dictation technology is speaker adaptive. This is typical of dictation technology.

Speaker dependent speech recognition cannot be used effectively by a speaker until that speaker has trained every word in the system. This is typical of low-end voice-activated dialing systems.

Speaker independent speech recognition can be used by people who have not first enrolled with a system or otherwise trained it to recognize their speech. Most speech-recognition systems not used for dictation are speaker independent.

Speaker identification generally, the process of finding the identity of an unknown speaker by comparing the voice of that unknown speaker with voices in a database of speakers. It entails a one-to-many comparison.

Speaker recognition An ambiguous term. 1. A synonym for speaker identification. 2. A generic term referring to many spoken technologies applied to speakers, including speaker identification and speaker verification.

Speaker verification the process of determining whether a person is who she/he claims to be. It entails a one-to-one comparison between a newly input voiceprint (by the claimant) and the voiceprint for the claimed identity that is stored in the system.

Speech to text generally, a synonym for speech-recognition dictation technology. It converts spoken words into printed text.

Spoken language understanding technology and systems that incorporate elements of artificial intelligence to process spoken input. Sometimes this is used to refer to speech recognition systems that process "conversational speech" even if they do not use artificial intelligence.

Text-dependent A variant of speaker verification that requires the use of a password, pass phrase, or another pre-established identifier (e.g., the speaker's name).

Text-independent A variant of speaker verification that can process freely spoken speech (an unconstrained utterance).

Text-prompted A variant of speaker verification that asks users to repeat random numbers and/or words. A typical prompt might be "Say 25 84." Some developers consider text prompting to be a kind of text-independent technology. It is also called *challenge response*

Threshold the degree to which a new speech sample must match a stored voiceprint before a system will accept the claim of identity. In most products the threshold is adjustable and may be set at different levels depending upon the security requirements of the application, user, and/or organization.

Voice ID a generic term covering speaker verification, speaker identification, and speaker separation. A synonym of speaker recognition and voice biometrics. Not widely used

Voice recognition synonym for speech recognition. Also used as a synonym for speaker verification

Voiceprint a sample of speech that has been converted to a form that a voice biometrics system can analyze