Designation: E2364 – 04 (Reapproved 2010)

Standard Guide to Speech Recognition Technology Products in Health Care[1]

This standard is issued under the fixed designation E2364; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This guide identifies system types and describes various features of speech recognition technology (SRT) products used to create the healthcare record. This will assist users (health information professionals, medical report originators, administrators, medical transcriptionists, speech recognition medical transcription editors (SRMTEs), system integrators, support personnel, trainers, and others) to make informed decisions relating to the design and utilization of SRT systems.

1.2 This guide does not address the following items:

1.2.1 System and data (voice and text) security.

1.2.2 Administrative processes such as authentication of the document, productivity measurements, etc.

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E1902 Specification for Management of the Confidentiality and Security of Dictation, Transcription, and Transcribed Health Records

E1985 Guide for User Authentication and Authorization

E2084 Specification for Authentication of Healthcare Information Using Digital Signatures

E2184 Specification for Healthcare Document Formats

E2185 Specification for Transferring Digital Voice Data Between Independent Digital Dictation Systems and Workstations

E2344 Guide for Data Capture through the Dictation Process

2.2 *Other Documents:*

Resource Interchange File Format (RIFF) Standard

3. Terminology

3.1 *Definitions:*

3.1.1 *acoustic model, n*—phoneme map of user.

3.1.2 *authentication, n*—the process of confirming authorship of an entry or of a document, for example, by verifying with a written signature, identifiable initials, computer key, or other methods.

3.1.3 *author, n*—person responsible for content of text file.

3.1.4 *back-end system, n*—delayed processing for document completion.

3.1.5 *compound file, n*—a file containing recorded voice with its transcribed text.

3.1.6 *context, n*—a long list of vocabulary words and phrases used for the particular subject matter, with their spellings and pronunciations, statistical information about usage of each word alone and in combination. For example, the context may include the number of times that "right," "Wright," "turn right," "right turn," "right hand," and "Mr. Wright" occur in a body of text. It also includes grammar and style information. Language model, lexicon, topic, and vocabulary are terms that are all used synonymously with context.

3.1.7 *digital signature, n*—data associated with, or a cryptographic transformation of, a data unit that allows a recipient to prove the source and integrity of the data unit and protect against forgery, for example, by the recipient.

3.1.8 *edit, v*—to review the document while listening to the originator's recorded voice and reading the associated transcribed text (compound file), checking for recognition errors and correcting document formatting and other inconsistencies. When the SRMTE is not the originator, the SRMTE may need to flag the document for originator/author clarification of unclear content or intent.

3.1.9 *encryption, n*—the process of transforming plain text (readable) into cipher text (unreadable) for the purpose of security and privacy.

3.1.10 *front-end system, n*—a system incorporating real-time recognition and may include real-time self-editing by the originator.

3.1.11 *language model, n*—context specific to medical specialty, user, or practice setting.

3.1.12 *lossless compression, n*—a lossless compression reduces the amount of data required to represent the original voice file but has no impact on sound quality. The original file can be replicated precisely at any time.

[1] This guide is under the jurisdiction of ASTM Committee E31 on Healthcare Informatics and is the direct responsibility of Subcommittee E31.15 on Healthcare Information Capture and Documentation.

Current edition approved March 1, 2010. Published August 2010. Originally approved in 2004. Last previous edition approved in 2004 as E2364–04. DOI: 10.1520/E2364-04R10.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

3.1.13 *lossy compression, n*—a lossy compression loses some information, resulting in degradation of the sound quality inherent in the original voice file and an inability to precisely regenerate that original file.

3.1.14 *microphone, n*—an instrument whereby sound waves are caused to generate or modulate an electric current usually for the purpose of transmitting or recording sound (as speech or music).

3.1.15 *microphone element, n*—the component within the microphone that does the actual conversion from sound waves to electrical signals.

3.1.16 *natural language processing, n*—method used in artificial intelligence to process and derive interpretation of human language.

3.1.17 *networked system, n*—system connected to a network.

3.1.18 *"normal" dictation, n*—routine phrases or paragraphs.

3.1.19 *originator, n*—person who provides oral input or dictation, not necessarily the person responsible for the content.

3.1.20 *phoneme, n*—smallest unit of sound in a spoken language.

3.1.21 *prompts, n*—reminders provided in order to complete a task.

3.1.22 *real-time recognition, n*—simultaneous speech-to-text transcription.

3.1.23 *RecO*—speech recognition error

3.1.24 *RIFF file, n*—Resource Interchange File Format (RIFF) is self-descriptive; that is, the voice file format is defined within the file.

3.1.25 *speech recognition, n*—computerized transcription of speech to text.

3.1.26 *speech recognition medical transcription editor, n*—medical transcriptionist who edits compound files and/or the SRT language model.

3.1.27 *SRT engine, n*—speech recognition processor.

3.1.28 *standalone system, n*—system not connected to a network.

3.1.29 *synchronization, v*—having voice and text matched such as in a point-and-play manner.

3.1.30 *text file, n*—a file that contains text message.

3.1.31 *voice enrollment, n*—the process whereby a user reads aloud selected text so the SRT software can map or record the user's speech sound pattern (phonemes).

3.1.32 *voice file, n*—digitalized audio message representing voice input.

3.1.33 *voice macros, n*—stored keystrokes that are activated by a voice command.

3.1.34 *WAV, n*—voice file format.

3.2 *Acronyms:*

3.2.1 *MT*—medical transcriptionist

3.2.2 *SRMTE*—speech recognition medical transcription editor

3.2.3 *RIFF*—resource interchange file format

3.2.4 *SRT*—speech recognition technology

4. Significance and Use

4.1 This guide is intended to provide general guidelines toward the design and utilization of SRT products used for healthcare documentation. It is intended to recommend the essential elements required of SRT systems in healthcare.

4.2 This guide will not identify specific products or make recommendations regarding specific vendors or their products or services.

4.3 A well-edited SRT document may result in improved quality over current methods of documentation, that is, handwritten notes and improved productivity over traditional dictation and transcription.

4.3.1 Faster turnaround times.

4.3.2 Legible documentation over handwriting has many advantages:

4.3.2.1 Improved patient care communication.

4.3.2.2 Enhanced patient safety.

4.3.2.3 Reduced malpractice risks.

4.3.2.4 Facilitation of appropriate reimbursement.

4.3.3 For the medical transcriptionist and/or SRMTE, decreased repetitive stress injuries, such as neck, arm, wrist, and heel pain.

4.3.4 Facilitation of cost controls related to document completion.

4.3.5 Better utilization of medical language skills of MTs as productivity is not limited by keyboarding skills.

5. Speech Recognition Technology Systems

5.1 Speech recognition technology (SRT) is designed to capture voice and transcribe that speech into text. This can be done by a single user working at a standalone computer or by a large group of users working on a network. Another method is processing a pre-recorded digital voice file through an SRT system, with the resulting text and/or SRT engine being edited by the MTE.

5.2 Speech recognition technology system workflow.

5.2.1 Front-end speech recognition process involves:

5.2.1.1 Recording the voice.

5.2.1.2 SRT transcription of the voice file to text.

5.2.1.3 Editing may be done by the originator and/or SRMTE.

5.2.1.4 Compound file may be saved as an option.

5.2.1.5 Text file can be printed, archived, transmitted, or integrated into an electronic health record.

5.2.1.6 Update the SRT context for RecOs and new terminology.

5.2.2 Back-end speech recognition process involves:

5.2.2.1 Recording the voice.

5.2.2.2 Transmitting the voice file to the speech recognition engine.

5.2.2.3 SRT transcription of the voice file to text.

5.2.2.4 Saving the voice and text as a compound file.

5.2.2.5 Routing the compound file to the SRMTE.

5.2.2.6 Editing done by the SRMTE.

5.2.2.7 Saving the text file.

5.2.2.8 Returning the edited text file to the originator for authentication.

5.2.2.9 SRMTE updates the SRT context for RecOs and new terminology.

5.2.3 *Standalone SRT System:*

5.2.3.1 Only one person at a time can use a standalone system.

5.2.3.2 Context is limited by the hard drive space.

5.2.3.3 Editing is done locally, at the point of input, either by the originator or by the SRMTE.

5.2.3.4 Input devices.

(1) Noise-canceling SRT microphones.

(2) Handheld digital recorders.

(3) Digital dictation systems.

(4) Telephones.

5.2.3.5 The following scenarios are offered to give the reader examples of how these systems work. They are not intended to represent every possible scenario for these systems.

(1) A radiologist (originator) dictates into a microphone connected to a personal computer running an SRT program. The voice is translated to text in real time. The originator edits the text and/or the SRT context.

(2) A family practitioner dictates into a personal computer throughout the day. Each compound file is saved and then, using the same computer, the SRMTE edits the text, listening to the recorded voice as necessary for clarification. The SRMTE may also be responsible for editing the SRT context.

(3) A group of cardiologists dictate into handheld digital recording devices throughout the day. The voice files are transmitted from the recorders to a computer and recognized by the SRT engine, using the cardiology context and each physician's acoustic model. Once recognized, each text file is edited by the SRMTE. The SRMTE may also be responsible for editing the SRT context.

5.2.4 *Networked SRT System:*

5.2.4.1 On a networked system, all files containing recorded dictation (voice files) are transmitted to a server, where the files are queued up for recognition. The compound files are then routed to the SRMTE for editing.

5.2.4.2 A networked system is designed to allow multiple originators and SRMTEs to work simultaneously. The voice files are recognized on a server or at the workstation(s) and the resulting compound files are routed to the SRMTE for editing.

5.2.4.3 Contexts.

(1) The networked system may be programmed for a single medical specialty or subspecialty, such as radiology, pathology, family practice, physical therapy, or emergency medicine.

(2) A networked system may also be programmed with many contexts or language models so originators from many different medical specialties can use it to improve speech recognition accuracy.

5.2.4.4 Editing may be done in the same facility, or the compound files may be sent to a remote SRMTE.

5.2.4.5 Input devices.

(1) Noise-canceling SRT microphones.

(2) Handheld digital recorders.

(3) Digital dictation systems.

(4) Telephone.

5.2.4.6 The following scenarios are offered to give the reader examples of how these systems work. They are not intended to represent every possible scenario for these systems.

(1) Six radiologists simultaneously dictate at individual workstations. Each voice file is routed to a recognition server, or the processing may take place on each workstation, with information regarding the originator's specialty and identification, allowing the recognition server to load the corresponding acoustic model and context. The voice file is processed by the SRT engine and the resulting compound file (voice and text files) is routed to the SRMTE for editing. The SRMTE may also be responsible for editing the SRT context.

(2) A hospital has 300 healthcare providers dictating into portable handheld digital recording devices from the hospital and several remote satellite clinics, or dictation may take place on individual workstations. The voice files are encrypted and securely transmitted to the digital dictation system of a contracted transcription company. Each voice file is routed to a recognition server, or the processing may take place on workstations, with information regarding the originator's specialty and identification, allowing the recognition server to load the corresponding acoustic model and context. The voice file is processed by the SRT engine and the resulting compound file (voice and text files) is routed to the SRMTE for editing. SRMTEs working both in the office and remotely receive recognized compound files via encrypted Internet transmissions. The editing is performed on standalone computers and the encrypted text files are returned. The SRMTE may also be responsible for editing the SRT context.

6. Training

6.1 *Originators:*

6.1.1 Voice enrollment and proper position of microphone and proper placement of microphone element.

6.1.2 Build customized language model.

6.1.3 Build "normal" dictations per user.

6.1.4 Develop skill sets.

6.1.4.1 Proper correction technique for a RecO.

6.1.4.2 Navigation/mobility skills for moving around in the document.

6.1.4.3 Editing skills specific to SRT products.

6.1.4.4 Editing language model.

6.2 *Speech Recognition Medical Transcription Editor:*

6.2.1 Voice enrollment and proper position of microphone and proper placement of microphone element.

6.2.2 Build customized language model.

6.2.3 Build "normal" dictations per user.

6.2.4 Develop skill sets.

6.2.4.1 Proper correction technique for a RecO.

6.2.4.2 Navigation/mobility skills for moving around in the document.

6.2.4.3 Editing skills specific to SRT products.

6.2.4.4 Editing language model.

6.2.4.5 Start and stop audio file.

6.2.4.6 Identify a RecO.

7. Realities of Speech Recognition Technology

7.1 Originators with good dictation habits will more likely be successful using SRT. See Guide E2344.

7.2 Originators with exceptionally heavy guttural accents may have more challenges. However, speakers of English as a second language are not necessarily precluded from using SRT.

7.2.1 Depending upon the SRT context it may not resolve the differentiation of homonyms.

7.3 SRT will not overcome dictation errors, improper grammar, incomplete or disorganized dictation, or incorrect punctuation.

7.4 Excessive background noise in the environment may cause problems in recognition accuracy. However, this may be overcome with noise-canceling SRT microphones and proper microphone position and placement.

7.5 Editing may be necessary, whether by the originator or by the SRMTE.

7.6 SRT will not eliminate the need for all medical transcriptionists.

7.7 The MT who has good English language skills, understands the medical language, can identify a RecO, and can edit the SRT context may become the SRMTE.

7.8 Appropriate training will facilitate better results with the SRT system.

7.9 SRT may not be the method of choice for data entry in some environments.

8. Features of SRT Systems

8.1 SRT begins with the input of a voice file and ends with the output of an SRT compound file. The following features are aspects that an SRT system could have utilizing other software applications. Unless noted, features may be found on both standalone and networked systems.

8.2 *System Administration:*

8.2.1 Set up users.

8.2.2 Map context/vocabulary to users.

8.2.3 Security/access limitation may be accomplished by the SRT engine or by the server/workstation environment.

8.2.4 Manage users' voice profile files.

8.3 *Speech Input:*

8.3.1 Pre-formatted normals, forms, and templates.

8.3.2 Personalized voice commands and voice macros.

8.3.3 Prompts for headings.

8.3.4 Context prompts.

8.4 *Document Output:*

8.4.1 Add digital signature through voice macro (See Specification E2084).

8.4.2 Save in optional text formats.

8.4.3 Voice macros may be used for document distribution.

8.4.3.1 Fax.

8.4.3.2 Print.

8.4.3.3 Secure e-mail.

8.4.3.4 Point-to-point electronic transmission.

8.4.3.5 Inclusion in designated record set.

8.5 *Editing:*

8.5.1 *Text Editing*—The system should have the following capabilities:

8.5.1.1 Text version controls.

8.5.1.2 Synchronization. The voice and text should be synchronized so that the SRMTE can point to a portion of the text and hear the voice played back at the same time (such as with a point-and-play technique).

8.5.1.3 Flag/bookmark/electronic "sticky note."

8.5.1.4 Word processing features, such as cut, paste, delete, undo, redo, text formatting (bold, italics, etc.), search, find, replace.

8.5.1.5 Ability to enter corrections with keyboard or voice.

8.5.1.6 Abbreviation expansion.

8.5.1.7 Alternate (pop-up) word lists.

8.5.1.8 Medical and English spellcheckers.

8.5.1.9 Electronic medical dictionary.

8.5.1.10 Flexible file formats: rich text format (.rtf), text-only (.txt), Microsoft Word (.doc), or Corel WordPerfect (.wpd).

8.5.2 *Voice Playback:*

8.5.2.1 Controls.

(1) Play, rewind, fast forward, auto rewind.

(2) Speed.

(3) Volume.

(4) Pitch.

8.5.2.2 Devices.

(1) Foot pedal.

(2) Keyboard.

(3) Mouse.

(4) Headset and/or speakers.

(5) Microphone.

8.6 *Voice:*

8.6.1 *Voice File/Voice File Compression*—Sound is an analog wave that can be converted into a digital form for computer storage or non-degradable transmission. The digital voice file is generated by sampling the sound wave at predetermined intervals, and then representing each sample with bits or bytes of data. When the voice needs to be heard, a digital-to-analog conversion is performed in order to reconstitute the actual sound of the voice. The resulting sound quality will be related to:

8.6.1.1 The accuracy of the sampling algorithm.

8.6.1.2 The amount of data originally collected (the number of samples per unit of time multiplied by the number of bits per sample), and

8.6.1.3 The impact of any data compression.

8.6.1.4 It will also be affected by the hardware.

NOTE 1—Compression can be "lossless" or "lossy." A lossless compression reduces the amount of data required to represent the original digital voice file, but has absolutely no impact on sound quality. The original file can be replicated—precisely—at any time. Lossy compression actually loses some information, resulting in degradation of the sound quality inherent in the original voice file and an inability to precisely regenerate that original file. However, "compression ratios" are typically greater with this type of compression. Suitable lossy compression algorithms are able to reduce data storage and transmission requirements while maintaining a sufficient voice quality for the sound file's intended use.

8.6.2 *Voice File Formats:*

8.6.2.1 RIFF is a tagged-file specification used to define formats for multimedia files (See Resource Interchange File Format (RIFF) Standard). Tagged-file structure helps prevent compatibility problems that often occur when file-format definitions change over time. Because each piece of data in the file is identified by a header, an application that does not recognize a given data element can skip over the unknown information.

8.6.2.2 WAV file uses the RIFF format and has a .WAV filename extension. RIFF format voice files are self-descriptive; that is, the voice file format is defined within the file. This standard supports the WAV voice file formats.

(1) DSS is a compressed WAV file format with a .DSS filename extension. With DSS, the original WAV file is compressed to a ratio of approximately 12:1 with lossless results.

8.7 *Potential Features that may be Included in SRT Systems with Complementary Software Applications :*

8.7.1 Document completion may include XML tags. See Specification E2184.

8.7.2 Assignments of codes using natural language processing.

8.7.2.1 Generates bill for services rendered.

8.7.3 Medication alerts, reminders, and prompts.

8.7.4 Monitoring/reporting.

8.7.4.1 Usage.

8.7.4.2 Productivity.

8.7.5 System backup.

8.7.6 Manage and test accuracy with test scripts.

8.7.7 Create a flag/bookmark/electronic “sticky note.”

8.7.8 Security features. See Specification E1902.

8.7.8.1 Ensure authorized access only (See Guide E1985).

8.7.8.2 Audit trail.

9. Keywords

9.1 automatic speech recognition; context; digital voice files; language model; medical dictation; medical documentation; medical transcription; phoneme; RecO; speech recognition; synchronization; transcriptionist-editor