International Journal of Medical Informatics (2004) **73**, 719—730

International Journal of
Medical
Informatics

www.intl.elsevierhealth.com/journals/ijmi

A simple error classification system for understanding sources of error in automatic speech recognition and human transcription

Atif Zafar[a,b,*], Burke Mamlin[a,b], Susan Perkins[b,c], Anne M. Belsito[b], J. Marc Overhage[a,b], Clement J. McDonald[a,b]

[a] *School of Medicine, Regenstrief Institute, Indiana University, 1001 West 10th Street, RG5 Indianapolis, IN 46202, USA*
[b] *Regenstrief Institute for Health Care, Indianapolis, IN, USA*
[c] *Division of Biostatistics, Indiana University, Indianapolis, IN, USA*

Received 9 September 2003 ; received in revised form 12 May 2004; accepted 18 May 2004

KEYWORDS
Speech recognition;
Electronic medical records;
Error rates;
Medical errors;
Medical notes;
Documentation;
Information systems;
Transcription

Summary *Objectives:* To (1) discover the types of errors most commonly found in clinical notes that are generated either using automatic speech recognition (ASR) or via human transcription and (2) to develop efficient rules for classifying these errors based on the categories found in (1). The purpose of classifying errors into categories is to understand the underlying processes that generate these errors, so that measures can be taken to improve these processes. *Methods:* We integrated the Dragon NaturallySpeaking^{TM} v4.0 speech recognition engine into the Regenstrief Medical Record System. We captured the text output of the speech engine prior to error correction by the speaker. We also acquired a set of human transcribed but uncorrected notes for comparison. We then attempted to error correct these notes based on looking at the context alone. Initially, three domain experts *independently* examined 104 ASR notes (containing 29,144 words) generated by a single speaker and 44 human transcribed notes (containing 14,199 words) generated by multiple speakers for errors. Collaborative group sessions were subsequently held where error categorizes were determined and rules developed and incrementally refined for systematically examining the notes and classifying errors. *Results:* We found that the errors could be classified into nine categories: (1) annunciation errors occurring due to speaker mispronunciation, (2) dictionary errors resulting from missing terms, (3) suffix errors caused by misrecognition of appropriate tenses of a word, (4) added words, (5) deleted words, (6) homonym errors resulting from substitution of a phonetically identical word, (7) spelling errors, (8) nonsense errors, words/phrases whose meaning could not be appreciated by examining just the context, and (9) critical errors, words/phrases where a reader of a note could potentially misunderstand the concept that was related by the speaker. *Conclusions:* A simple method is presented for examining errors in transcribed documents and classifying these errors into meaningful

[*] Corresponding author. Tel.: +1 317 554 0000x2067; fax: +1 317 554 0114.
E-mail address: azafar@iupui.edu (A. Zafar).

doi:10.1016/j.ijmedinf.2004.05.008

and useful categories. Such a classification can potentially help pinpoint sources of such errors so that measures (such as better training of the speaker and improved dictionary and language modeling) can be taken to optimize the error rates.

1. Introduction

Accurate, legible and detailed documentation of clinical encounters, created as efficiently as possible, is now a primary workflow goal in most clinical settings. Electronic documentation strategies can potentially provide such productivity improvements. Among the numerous available modes of electronic data entry (typing, electronic templates, optical character recognition, dictation with transcription), speech input is popular because it is the most efficient. Automatic speech recognition (ASR) by computer and dictation by the clinician/transcription by a human being (D/T) are two competing modes of speech input. Computer-based ASR has attracted much attention because of its promise to deliver timely dictations at minimum cost. D/T is generally more expensive and slower (8—15 cents/line and 2—5 days of turnaround time) [1,2].

Our objective was to understand the sources of error within these two competing modes of electronic data entry (ASR and D/T). We accomplished this by developing a simple system for classifying the dictation errors generated when using these modalities. We emphasized those error types which pin-pointed fixable causes of errors and developed classification rules which could be efficiently employed, when attempting to proofread the notes for dictation errors.

2. Background

Although vastly superior to yesterday's discrete speech technology, today's continuous speech engines still produce a substantial number of recognition errors. Conversion of speech into text, whether by a computer or a human transcriber, invariably results in some errors. The causes of these errors are manifold, and include problems with either system design (accuracy of speech recording or completeness of the vocabulary) or process (the way the user speaks or the computer or human transcriber hears and understands) [3] (Fig. 1a and b).

Although error correction is an important step in the processing of dictated notes, an error that is missed during proofreading may not cause problems with interpretability of the intended concept of the speaker. This is because readers can often guess a plausible conceptual meaning of a garbled phrase from the context of an unedited transcribed note. However, if they are unable to guess what was said or if they guess incorrectly then care may potentially change and the error may have clinical significance.

It is important to study uncorrected notes for interpretability because in many instances (esp. with radiology notes), final (proofread) notes may not be available when the information in the notes is needed for clinical care. Holman et. al. [8] looked at unedited radiology reports (dictated through a commercial transcription company) and found that 2% of these preliminary (uncorrected) reports could have led to additional (unnecessary) testing or treatment and that 0.4% of these reports would place the patient at moderate risk for substantial morbidity.

A validated error classification system serves as an effective instrument for benchmarking the accuracy of the various modes of data entry. Prior studies, both within [3—6] and outside of medicine [11—18], have addressed the topic of error rates in computer transcribed notes. Much has been learned about the syntactic (punctuation, grammer, spelling etc.) and semantic (out-of-vocabulary words, sound-alikes, tense mismatches etc.) nature of errors in the last twenty years [13—18]. However, many of the studies in medicine have not properly distinguished between these syntactic and semantic criteria as they described the errors. They also employed techniques where users either simply read in existing documents (ignoring the errors that occurred when notes are created de-novo) or used a batch recognition model, where speech files are captured and transcribed off-line (ignoring the ''learning'' which occurs when users get real-time feedback and dynamically adjust their speaking styles to compensate for the error). Furthermore, few of these studies compared speech recognition with the competing modes of data entry (dictation/transcription or direct keyboard entry).

Our pilot study, on the other hand, was performed in a real clinical setting where the user was burdened with time pressures to create and proofread his notes, had real-time recognition feedback, allowing him to adjust his speaking style and make corrections on the fly, and had the task of creating the notes from scratch, adding ''thinking time'' and ''interruptions'' to the ASR process. This ensured that all types of errors generated in uncontrolled conditions could be accounted for. Furthermore, our classification system was based on an understanding of the entire transcription process — from thought to action — so that the root causes of syntactical errors could be determined and fixed.

3. Methods

3.1. Study site and workflows

The primary objective was to develop a categorization of errors found in ASR and D/T notes and we thus needed to generate a corpus of notes created by using ASR and by D/T. For gathering the ASR data, twenty practicing internists within the Indiana University Medical Group (IUMG) outpatient internal medicine practice were invited to participate. However, a single person, who was enthusiastic about the technology, agreed to participate. We also captured uncorrected but D/T notes from five other internists at other outlying community health centers within the IUMG system.

Physicians in the IUMG clinics use a computerized provider order entry (CPOE) system, called the Medical Gopher^{TM}, to perform all clinic discharge activities. They have to type in all test requisitions, consult requests, medication refills, billing levels, follow-up appointment planning and print patient

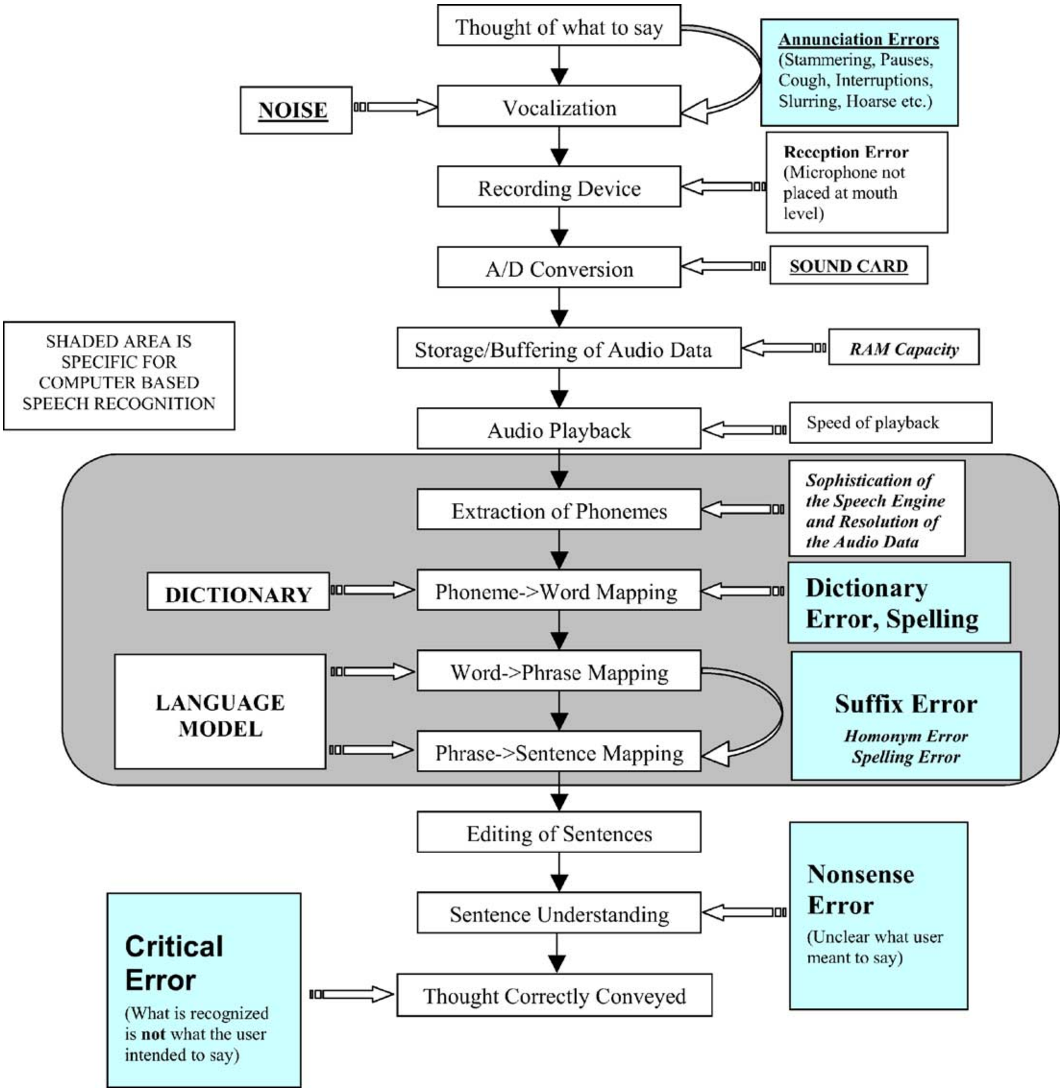

Thought of what to say
Annunciation Errors
(Stammering, Pauses, Cough, Interruptions, Slurring, Hoarse etc.)
NOISE
Vocalization
Recording Device
Reception Error
(Microphone not placed at mouth level)
A/D Conversion
SOUND CARD
SHADED AREA IS SPECIFIC FOR COMPUTER BASED SPEECH RECOGNITION
Storage/Buffering of Audio Data
RAM Capacity
Audio Playback
Speed of playback
Extraction of Phonemes
Sophistication of the Speech Engine and Resolution of the Audio Data
DICTIONARY
Phoneme->Word Mapping
Dictionary Error, Spelling
LANGUAGE MODEL
Word->Phrase Mapping
Phrase->Sentence Mapping
Suffix Error
Homonym Error
Spelling Error
Editing of Sentences
Nonsense Error
(Unclear what user meant to say)
Sentence Understanding
Critical Error
(What is recognized is not what the user intended to say)
Thought Correctly Conveyed

Fig. 1 (A) Computer error classification in the speech recognition pipeline. There can be multiple points in the speech recognition pipeline where errors may occur. These range from the speech input to the editing and review part. The speech pipeline is described fully below. (B) Human transcriptionist error classification in the speech recognition pipeline. There can be multiple points in the speech recognition pipeline where human transcribers can create errors in the final document. These range from the speech input to the editing and review part.

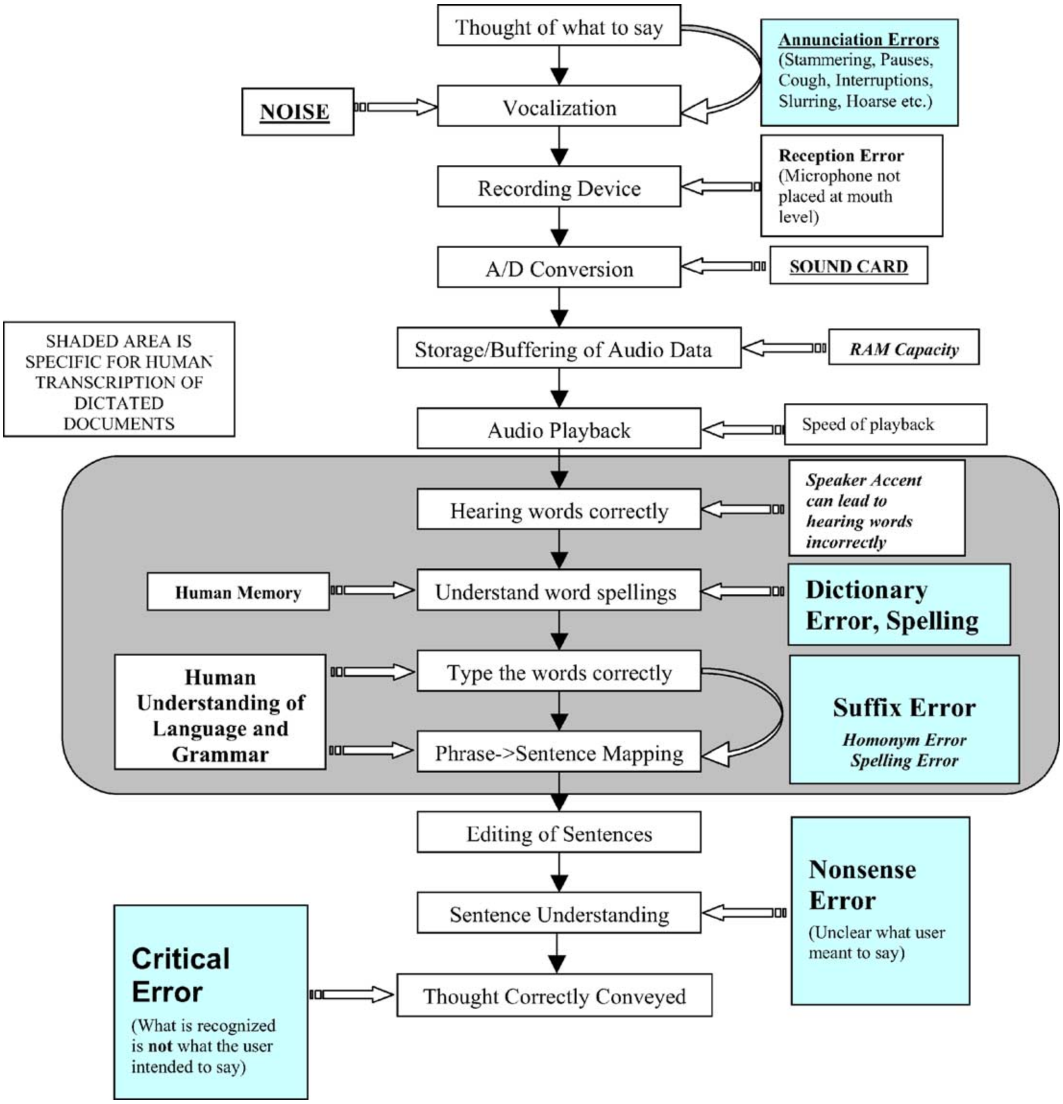

Thought of what to say
Annunciation Errors
(Stammering, Pauses, Cough, Interruptions, Slurring, Hoarse etc.)
NOISE
Vocalization
Recording Device
Reception Error
(Microphone not placed at mouth level)
A/D Conversion
SOUND CARD
SHADED AREA IS SPECIFIC FOR HUMAN TRANSCRIPTION OF DICTATED DOCUMENTS
Storage/Buffering of Audio Data
RAM Capacity
Audio Playback
Speed of playback
Hearing words correctly
Speaker Accent can lead to hearing words incorrectly
Human Memory
Understand word spellings
Dictionary Error, Spelling
Human Understanding of Language and Grammar
Type the words correctly
Phrase->Sentence Mapping
Suffix Error
Homonym Error
Spelling Error
Editing of Sentences
Nonsense Error
(Unclear what user meant to say)
Sentence Understanding
Critical Error
(What is recognized is not what the user intended to say)
Thought Correctly Conveyed

Fig. 1 (*Continued*).

education materials. Most physicians use the keyboard to enter these items, however, mouse input is also available if desired. Visit documentation is handled in one of three ways: typing into the Medical Gopher™ within specific free-text note fields, handwriting the note in the chart or dictating into a telephone for later D/T processing. Such a clinic discharge planning session using the Medical Gopher™ is called an ''order session'' [9].

3.2. Capturing ASR data

In order to enable ASR to be used with little impediment to workflow, we integrated our test engine (Dragon Systems' Naturally Speaking v4.0™) into the Medical Gopher™, the CPOE system at IUMG clinics [8]. We used the Korteam Intl. SpeakingClinically™ Medical Dictionary. The workstations were Compaq desktop computers with Intel Celeron processors running at 300 MHz, with 160 MB of RAM, a speech-quality (16-bit) sound card (SoundBlaster PCI 128) and a noise-canceling microphone (VXI Parrot). The total cost of each installation was about $2000 (in 1998 dollars), the same system costing about $1200 today. It took 2 months or programming time to perform the integration.

The integration was seamless so that a user did not have to learn a new system in order to use the speech software. The system worked by ''voice-enabling'' the note fields into which a user would normally type. The transcribed note was then presented to the user for error correction using familiar keyboard commands. We captured both the unedited output of the speech engine and the user corrections and stored them in separate files so that they could be compared for research purposes. The system also records timing data for completion of an entire order session as well as the time spent within each data-entry field. This capability allows us to benchmark the efficiency of electronic data entry.

We were unable to capture the actual voice recordings to make comparisons because of a limitation in the speech engine programming interface. As a result, our error system depended heavily on making ''educated guesses'' about the possible conceptual interpretations of an error from the context of the dictation. Although this is not the optimal way to characterize errors it does permit us to get an upper bound for the ''interpretability'' of the clinical concepts as represented in an uncorrected transcribed note, which is important for rating the clinical significance of an error.

The study subject initially spent 45 min training the ASR system to recognize his voice. He then spent 1 month dictating random notes (one half-day per week) in noisy environments getting ''used to'' the system. During the ASR arm of the study, he dictated outpatient clinic notes over 6 months. After the completion of this trial we conducted a short interview asking the subject about his experience with the technology, advice he may give to others and what he learned by using the system.

3.3. Capturing D/T data

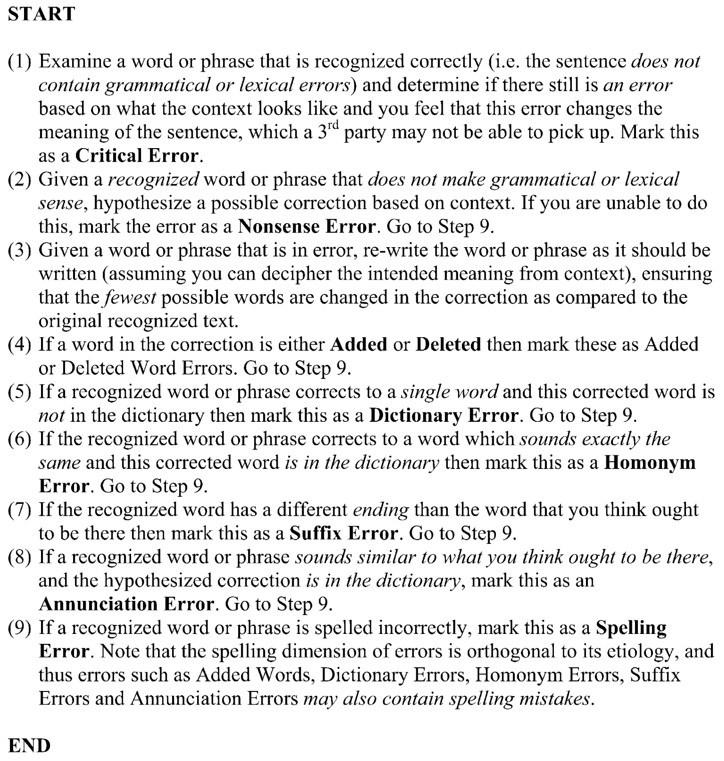
We also captured random human-transcribed outpatient notes dictated by five practicing physicians from our IU Medical Group Primary Care practices, which had not been reviewed and corrected by the clinicians. All of these notes (ASR and D/T) were generated from actual clinic patient encounters and the user did not simply ''read-in'' existing documents. For the ASR notes, we examined both the uncorrected output from the speech engine and the corrected notes.

3.4. Developing the error classification scheme

We have previously reported on the mechanics of how speech recognition pipelines work [1]. We examined this pipeline in detail (Fig. 1a and b), as speech is converted into text, enumerating points within this data flow where errors could potentially occur. This comprised a first-pass categorization. We then performed a literature review of prior medical [3—7] and non-medical [11—18] (see esp. reference [13]) studies looking for other error categorizations. We combined the first-pass classification with error categories as reported in the literature to come up with an empiric classification scheme. We subsequently performed a validation of this empiric classification in two stages: (1) independent note review by three practicing physicians and (2) group collaborative review with the development of simple classification rules.

(1) Initially, the three reviewers independently examined the notes (direct output of the ASR engine and the speaker corrected notes as well as the human transcribed notes) to create a preliminary estimate of error rates and to categorize the errors based on the empiric scheme. The reviewers initially identified all syntactical errors in a document (those resulting in grammatically or technically incorrect phrases) and categorized them as either interpretable from context (whose intended conceptual meaning was clear by looking at the whole sentence, paragraph or dictation) or un-interpretable. In a second pass, the errors were examined in more detail and possible explanations hypothesized: whether the error was due to a word missing from the dictionary, whether it was a tense mismatch, a nonsensically added or deleted word, a homonym mismatch or an ''other'' category. The other category was labeled as a catch-all category of ''annunciation errors''. Finally, in a third pass, by looking at a phrase, if there were multiple possible hypotheses for what could have been spoken, some contradicting others, then the error was labeled as a ''critical error''.
(2) In the second stage, the three physician reviewers held group collaborative sessions where the errors were compared, error categories defined and simple classification rules devised. As we rated errors together, we found that the initial definitions of errors were imprecise and different reviewers would classify the same error in different, but equally valid, ways (see the discussion section for examples of this). As a consequence of this imprecision we created some simple rules for detecting and classifying errors (see Fig. 2) which removed the uncertainty about how each error ought to be classified. We required 10 rounds of iterative review before we achieved consensus about the error class definitions and classification rules. In the last round of review, we held a group collaborative session where we rated a note together in order to verify the consensus error categories and classification rules.

All of the reviewers used the context of the utterance to decide how to classify each error and to hypothesize what the speaker may have potentially



START

(1) Examine a word or phrase that is recognized correctly (i.e. the sentence *does not contain grammatical or lexical errors*) and determine if there still is *an error* based on what the context looks like and you feel that this error changes the meaning of the sentence, which a 3rd party may not be able to pick up. Mark this as a **Critical Error**.

(2) Given a *recognized* word or phrase that *does not make grammatical or lexical sense*, hypothesize a possible correction based on context. If you are unable to do this, mark the error as a **Nonsense Error**. Go to Step 9.

(3) Given a word or phrase that is in error, re-write the word or phrase as it should be written (assuming you can decipher the intended meaning from context), ensuring that the *fewest* possible words are changed in the correction as compared to the original recognized text.

(4) If a word in the correction is either **Added** or **Deleted** then mark these as Added or Deleted Word Errors. Go to Step 9.

(5) If a recognized word or phrase corrects to a *single word* and this corrected word is *not* in the dictionary then mark this as a **Dictionary Error**. Go to Step 9.

(6) If the recognized word or phrase corrects to a word which *sounds exactly the same* and this corrected word *is in the dictionary* then mark this as a **Homonym Error**. Go to Step 9.

(7) If the recognized word has a different *ending* than the word that you think ought to be there then mark this as a **Suffix Error**. Go to Step 9.

(8) If a recognized word or phrase *sounds similar to what you think ought to be there*, and the hypothesized correction *is in the dictionary*, mark this as an **Annunciation Error**. Go to Step 9.

(9) If a recognized word or phrase is spelled incorrectly, mark this as a **Spelling Error**. Note that the spelling dimension of errors is orthogonal to its etiology, and thus errors such as Added Words, Dictionary Errors, Homonym Errors, Suffix Errors and Annunciation Errors *may also contain spelling mistakes*.

END

Fig. 2 Rules for classifying errors using context alone (use the sequence of steps as outlined below in order).

spoken. This ''guessing task'' involved identifying the most probable and clinically correct conceptual interpretation of an error, which may not correspond exactly with what the speaker may have said. However, we corrected the notes in a way we felt that a competent reader with domain knowledge of internal medicine may interpret them in order to make the most clinical sense. Such corrections did not require the voice recordings.

4. Results

The single speaker in the ASR arm generated 104 outpatient clinic notes, comprising 29144 words, over six months (12/98 to 5/99). A group of 44 random outpatient clinic notes, generated using D/T from five other speakers, were also acquired for analysis. The distribution of the types of clinical encounters (initial versus follow-up visits) was comparable among the corpus of notes (92% follow-up visits and 8% initial visits in the ASR arm and 88% follow-up visits and 12% initial visits in the D/T arm). No consultation visits were documented. We evaluated the ASR and D/T notes for error types.

Our consensus error classification system is shown in Table 1. After a detailed review of the corpus of notes, nine classes of errors were identified. We started out with one extra class in the preliminary scheme (stop word errors) and this category was collapsed into either the added or deleted category because they did not change the underlying meaning and resulted in extra words being added or deleted. If an error had different and potentially conflicting plausible interpretations, it was labeled as a critical error. Since we used context to classify each error, we came across cases where the recognized phrases were so garbled that it was impossible to use context to guess what was spoken. We labeled these errors as nonsense errors (or errors whose true meaning could not be determined from context).

During the first few rounds of group collaborative review, only 30—50% of the errors were assigned to the same category by all of the reviewers. Similarly, the initial rate at which all three reviewers identified the same word/phrase as an error was low (40—55%). Then, by a process of repetitive refinement of the classification system, we arrived at a consensus for how these errors should be classified (Fig. 2). It was very interesting to see that even with three reviewers carefully looking for errors, each one initially missed about 20—30% of the existing errors. Furthermore, the subject himself found that longer dictations took too much time to proofread and he had to ''rush through'' this process. As a result he missed about 66% of the errors. We conducted a post-trial interview (Table 2) and found

Table 1 Examples of errors rated by the classification

Error Type	*Example of the Error*
Annunciation errors	*Sheila wise* (She is otherwise) *Four egos* (Before he goes) Uses Albuterol for *opera superinfection* (upper respiratory infection)
Stop words (Add/del/substituted)	She *is* (has) tried several therapies Denies heart attack *the* (or) paraplegia If he goes through *if* (with) it
Suffix errors	We have seen *markedly* (marked) improvements He has *eat* (eaten) his meal She has hematochezia but denies *melanotic* (melena)
Homonym errors	She has a bone spur at the *see 3* (C3) vertebral level The patient is a twenty three year old *mail* (male) We will *ashore* (assure) him that his symptoms are benign
Words added	*The* she stopped her Zoloft on her own With *also* good relief of pain She received *of* radioiodine ablation
Words deleted	She's been out of her (—) for two months 3 weeks ago she (*had a*) sore throat We will (*be*) teaching him to check his blood sugars
Dictionary errors	*I/O to form* (iodoform) *Tom all coding* (Tylenol with codeine) *Perry ocular* (periocular) *Nasal pearl* (Benazepril) *Laredo Pam* (Lorazepam) Thyroid palpable and mild *except almost* (exophthalmos) *The salas* (Dorsalis) pedis pulses are palpable *Buddhist opponent* (But his Troponin)
Spelling errors	Patter (pattern) Dietitian (dietician) Her (here)
Critical errors	Blood sugars are in the *8290* (80 to 90) range She has occasional end *inspiratory* (expiratory) wheezes But has *noted* (no) symptoms If SGOT is elevated will *continue* (discontinue) Zocor BP is __22/95. (Could be 122 or 222 etc.)
Nonsense errors	The *last ventral what is to worry 18* She describes *no matter on June* He *use of friends and now* has had relief of his dyspnea

Please note that italicized phrases represent the errors and the (phrases in parenthesis) represent an attempt to guess the correct intended meaning of the erroneous phrase.

that ASR generally was faster than his typing speed but he had to be careful in how he spoke and when he paused. He also found that lack of terms continued to be problematic and certain words just could not be recognized well. He also found it difficult to error correct long documents in a short period of time in the clinic.

We will now define the types of errors found, provide a hypothesis for why they occur and detail some examples:

4.1. Classification system for errors

In the following discussion, we hypothesize the causes of errors and highlight them in ***italicized bold*** text.

(1) **Added** and (2) **deleted** word errors resulted from either a word being added or a word being deleted. These errors were easy to identify in the note because the knowledgeable

Table 2 Results of the interview with the speaker

Question	Response
How would you compare ASR to typing your notes?	ASR is faster than my typing but I had to speak in a manner differently than I am used to. If I hesitated in the middle of a sentence it made an error. I learned to pronounce some words in a very particular way or it just didn't understand them. There was generally a delay between when I started dictating and when the words appeared on screen. This was distracting and often made me say the same phrase twice.
How did you rate the error correction ability of the speech recognition system?	I generally corrected errors at the end of the dictation. It was possible to correct errors by re-dictating them but sometimes it just didn't understand and I had to type in the correction.
How efficient is the process of error correction? How much time does it take?	It can be fairly time consuming to correct a long document and I know I must have missed some errors because I did not have the time in clinic to go through the document carefully.
What kinds of errors did you observe occurring?	If I paused in the middle of a sentence, spoke too fast or if there was some noise in the background, small words such as ''and'' or ''if'' seemed to get added to the note. Many times, if I slurred a syllable it would replace it with something that sounded similar or would replace the intended word with the wrong tense. Sometimes it would make silly mistakes such as replace Lasix with Lay 6. Sometimes it just didn't understand anything and the whole phrase was garbled. If I did not pause before dictating the first line it made an error.
Did you feel that the process of speech recognition impeded your workflow?	Yes and No. I had to learn to use a new system which was distracting but once I got used to it I could dictate and correct fairly quickly. The system is still imperfect and takes time and patience to learn to use properly.

readers could ''fill-in-the-blanks''. ***This often occurred (as we learned in a post-trial interview) when the user spoke too fast or slurred his speech***. ***Ambient noise also sometimes resulted in these errors being introduced into the note***. We also categorized stop words (words such as ''and'', ''if'' or ''the'' which do not add content but make the sentence grammatically correct) as either added or deleted words.

(3) **Dictionary** errors are phrases that are recognized in lieu of a ***word missing from the dictionary***. The resulting translation sounded very similar to but not always exactly like the intended word. For example, since the word Benazepril was missing from the dictionary, the ASR engine chose ''nasal pearl''. We used context to decide what was potentially spoken. Another example is: ''we used I/O to form solution to prep the skin'' (should be Iodoform) and ''He had good relief of his arthritis pain with solo bricks'' (should be Celebrex). If we couldn't decide from context what was meant then we classified the word or phrase as a nonsense error.

(4) **Homonym** errors were misrecognized phrases which sounded exactly like the intended word. This happened when, for example, the phrase ''he is a 24-year-old male'' was recognized as ''he is a 24 year old mail''. Other examples included: ''She had osteophytes at the see 3 vertebral level'' and ''We will order a serum tighter of rheumatoid factor''. These errors were ***always due to a suboptimal language model*** (a statistical list of possible word adjacencies) [13] in the ASR software. Note that if the specific words (C3 and titer in the above examples) were absent from the dictionary, we would have classified these errors as dictionary errors and not homonym errors. Thus, homonym errors require that the intended word be present in the dictionary and that it is still misrecognized

(5) A **Suffix** error was due to the incorrect ending substituted for the intended one. This ***resulted when the user incompletely pronounced all the terminal syllables of a word***, leaving the computer to ''guess'' what was said. These errors ***could also have occurred if there was a language model problem***. Note that if the

specific tense is absent from the dictionary, our classification system would classify this as a dictionary error. These error types were separated into a stand-alone category (rather than calling them Annunciation Errors) because they occurred so often and we felt that separating them out would teach something specific about pronunciation to the speaker. An example is the word ''melanotic'' being recognized instead of ''melena''. Another example is ''She did not tolerating the procedure''. We inferred that the intended word was ''tolerate''. If ''tolerate'' is present in the dictionary, then the error is a pronunciation or language modeling problem and is correctly classified as a suffix error. If the term is not present in the dictionary, then it is really a dictionary error and not a suffix error.

(6) **Spelling** errors occurred only in human-transcribed notes and not with speech-recognized notes. They could also potentially occur with typed notes. This is because the dictionary in the speech software contains only correctly spelled words, and the ASR system is designed to recognize whole words and not individual letters.

(7) **Annunciation** errors were due to ***slips in how the speaker pronounced words***. If the speaker paused in the middle of a sentence, spoke too fast, slurred his words, used disfluencies [7] (um's, ah's) or used an accent (hoarse voice for example) that was not recognizable then these errors occurred. This was a ***''catch-all'' category*** and we classified all other errors for which the reviewers could guess a plausible intended meaning and which did not fit into the other categories as Annunciation errors. There can be many causes of these errors. For example, if a user has a hoarse voice or there are problems with ***prosody*** (intonation — see http://www.eptotd.btinternet.co.uk/pow/powin.htm for a web-based tutorial on prosody) then annunciation errors may occur. For example, consider the sentence: ''There was a bulge in his opal teal fossa when he flexed his knee''. We could readily understand from the context that the intended word was popliteal. We also knew that the word was indeed in the dictionary so we concluded that the error probably resulted from mispronunciation. Another example was the sentence ''There was an access on his left shoulder that was training a black cheesy substance.'' We could easily infer that ''access'' should have been ''abscess'' and ''training'' should have been ''draining''.

4.2. Critical errors

An error was deemed **critical** if the recognized word/phrase could lead to a change in meaning of the intended utterance. Most of the critical errors resulted ***when a positive was changed to a negative (or vice versa)or a number was misrecognized***. Several examples are as follows: ''We will check his liver function tests and if elevated will continue Zocor'' (should have been discontinue) and ''Her blood sugars are in the 8290 range'' (should have read 80 to 90). In one case, the speech engine output read ''he has a subcutaneous nodule on his right elbow with extension only to 106°''. This was bothersome because, without a goniometer, it would have been very hard to estimate the degree of flexion/extension, accurate to ± 1°, and the speaker did not have a goniometer accessible when he did his physical exam in the clinic. This could have certainly changed care because 106° is a valid extension angle for the elbow and this would be an indication for physical therapy or imaging.

4.3. Nonsense errors

Finally, if the reviewers ***could not guess a plausible intended meaning from context***, then the word/phrase was categorized as a **nonsense** error. Nonsense errors made the sentence completely un-interpretable; even astute readers were unable to infer the intended word from context. Had we recorded the dictations, these errors would probably be re-classified into either the annunciation error category or the dictionary error category. Since our objective was to look at the worst-case scenario (where none of the errors are corrected and the user has no voice recording with which to compare the text), we developed this category. An example of such an error is ''suspect me this 0 FIRDA done for removal to but will refer to dermatology''. We cannot infer the intended word or phrase for suspect me this 0 FIRDA so the sentence is nonsensical.

We then applied this scheme to the small corpus of notes we examined for errors. We found that all types of errors except spelling errors could be found in the ASR notes. We discovered that half the errors with ASR were annunciation related and could be improved with better speaking technique. In comparison, the human transcribed notes contained few annunciation errors, suffix errors, spelling errors and deleted word errors and lacked dictionary errors, homonym errors, added word errors, and nonsense errors. We found that there were significantly more errors detected in the ASR documents than in the human transcribed notes (1184 errors in ASR vs.

18 errors in human transcribed documents). However, we found that 88.9% of the errors in the ASR notes were such that a plausible clinical interpretation could be made from the context alone. On the contrary, with ASR, an alarming 9.4% (111) of the errors were nonsense errors, and 1.6% (20) of errors were critical errors. With human transcribed notes, only one critical error was detected.

5. Discussion

While developing the error classification system, there were initially many discrepancies in how raters classified errors. Initially, the reviewers disagreed about the classification of an error. This was due in part to the fact that the same error could be classified in potentially different ways. For example, consider the ASR transcribed phrase ‘‘patient presented followed for chronic medical problems’’. There were two possible ways to correct this phrase, with no consequential change in meaning. These two corrections as specified by raters A and B were:

(A) ‘‘Patient presents for follow-up of chronic medical problems’’

(B) ‘‘Patient presented for follow-up for chronic medical problems’’

Rater A:	Presented is a suffix error. Should be Presents* For is a deleted word error (should be in the corrected version) Followed is a dictionary error and should be follow-up (since the term ‘‘follow-up’’ is not in the dictionary) For (after the followed in the original phrase) is an annunciation error and should be of*
Rater B:	Presented is not an error at all and belongs in the sentence as is* For is deleted from the phrase after the word presented. Followed is a dictionary error and should be follow-up For (after the followed in the original phrase) is not an error at all*

*Words where there is a disagreement.

Both interpretations were correct. But given that we wanted to break up a phrase into the smallest possible ‘‘chunk’’ that was correctable, Rater B’s approach was the right one by our system (two versus four errors).

We also found that many errors were missed during the proofreading process. Overlooking errors are not a flaw in the error classification but a problem inherent in the correction process that is likely to be even worse when the dictating physician is reviewing notes in a busy environment where they may be distracted. Our findings can be considered a lower-bound for how many errors may be missed during the correction process. One solution to this would be to have the speech engine highlight those words or phrases that are recognized with a less than optimal accuracy so that users can quickly spot errors in a long note and correct them.

It is also imperative to speak clearly at a natural speed. Speaking too fast or to slowly may introduce errors. Slurring, hesitation, disfluent speech (um’s and ah’s) can all introduce errors. We found that, with ASR, about half of the errors were rated as annunciation errors that did not fit one of the other categories. The user in this trial was a native English speaker who had practice in speaking into the voice system prior to enrolling in the study. Thus, users have to be especially careful in how they pronounce words. With continued use, the system actually ‘‘trains the user’’ in how to pronounce words correctly (from trial and error). Also, the same user has a tendency to utter the same word or phrase slightly differently from time to time (prosodic problem), which can change the accuracy. Speakers who are used to dictation services tend to have more variability in their speech patterns and may find it very hard to change their speaking habits to attain good accuracy. This requirement may thus put off a lot of potential users [10].

Note that the annunciation error category is actually a much larger category of errors, which can be further dissected by more careful studies. We bundle many possible causes of these errors into one category in this study. For example, differences in user accents, in the speaking rate, in the emotional state of the reader (prosody), in the clarity of the speech (whether they slur words), in the amount of user training time etc. are all captured in this ‘‘catch-all’’ category. Words or phrases are classified as annunciation errors only when they do not fit one of the other immediately identifiable and ‘‘fixable’’ causes (dictionary or language model problems).

We found that terms missing from the dictionary still continued to be problematic, despite using an extensive dictionary. In part, this reflects institutional biases towards the usage of certain terms (e.g., we use the term Biox to mean pulse oximetry). Furthermore, new drug and test names are often missing from the dictionary, requiring local ‘‘maintenance’’ of this dictionary. Since many of these missing terms are longer medical words, which have very few ''phonetic analogs'' (similar sounding words that the speech engine can confuse with), simply adding them to the dictionary (without re-training them) should suffice. This can be done by the computer support staff for the practice without the users having to enter them manually. Also, certain abbreviations need to be avoided because they are absent from the dictionary or poorly recognized, regardless of the amount of training done. There also needs to be a short pause before dictating the first paragraph, allowing time for the user speech profile to load and initialize properly.

Homonym errors and substitution of improper suffixes and tenses also continued to pose problems. This can either reflect words missing from the dictionary or a language model that has not been tuned well. Suffix related errors may also result from a user incompletely pronouncing the end-syllables in a word so that a slurred utterance of a word such as ''radiating'' may actually result in ''radiate'' being recognized (even though the word ''radiating'' is in the dictionary). Language models are created by examining a large corpus of existing notes for the occurrence of words, generating a statistical model of word adjacency. Thus, one can improve these error rates by feeding existing dictated and transcribed notes into the speech engine using the vendor supplied language modeling tools.

Our classification system addressed the consequences of a mistake in terms of two parameters, note interpretability and error criticality (whether in the opinion of a reviewer it has the potential to change care). We found that a unique and clinically sound interpretation of the intended meaning could be estimated for 88.9% of the errors, by looking at the context alone. Thus, for the short period after dictation when the data in a note may be needed for clinical use, physicians may still find these notes acceptable, regardless of the high number of errors with ASR, and care may not necessarily be affected. However, uncorrected documents are obviously unacceptable for coding, billing, consultation reporting, legal and other purposes and a mechanism for offline correction is necessary for ultimate acceptability. This is especially true for critical errors which have the potential to alter care.

Thus, given these potential problems, it is advisable that users pilot test an ASR system before deploying it in a real clinical setting. Users can quickly get a sense for the error rate, given their speaking style. Knowledge of the types of errors can then help direct efforts towards removing these sources of error. Such efforts could be carried out in several ways: Users can add words to the dictionary or provide additional documents to the training module to improve the language model. Training users to speak more clearly and continuously can help to improve annunciation error rates. Highlighting potentially erroneous words/phrases by using the recognized probability values can help a user with limited time pinpoint errors in a lengthy document. They can also take corrective action ensure that the dictation environment is as noise free as possible, and that they are relatively protected from interruptions while dictating.

There are certain limitations to this study. We could recruit only one user to do the trials. However, we felt that we could reliably capture all the types of errors predicted by our empiric classification system, given enough notes or words dictated by this one person. We hypothesize that if there are differences in the error types among users then they will all be bundled into the annunciation error category. Future studies may tease them out more carefully. This also points to the difficulty in recruiting busy clinicians for studies, especially if they have to alter their practice routines or learn new technologies. We also could not capture the actual voice recordings from the speaker because of limitations in the programming interface supplied by the speech engine vendor. However, we looked at the technology behind the speech engine and predicted based on the processing stages which types of errors may be found. This approach corroborated well with published literature where speech recordings were used for error category determination. Furthermore, our estimate of error rates was based only on the examination of the context of an error, which may underestimate the true error rate. However, in this pilot study we were only interested in developing a classification scheme, which can subsequently be applied in order to study error rates more rigorously in the future. Finally, since the time of the study, the software has gone through three revisions (v7.0) and the accuracy may be improved over the older version. In any case, the classification scheme still applies today, regardless of the improvement in accuracy, although some errors may or may not occur (such as suffix errors).

6. Conclusion

We have described a simple classification scheme for errors in ASR and human transcribed notes. A set of rules for classifying errors is also presented which can quickly provide a sense of error etiology in clinical notes. Knowledge of the types of errors can then help direct efforts towards removing these sources of error. We found that despite a higher error rate with ASR as compared with human transcribed notes, many of the errors in ASR notes could be corrected by examining the context alone, assuming domain knowledge of medicine. We also discovered that a disturbing percentage of errors were left uncorrected and couldn't be understood from the context alone, with a small percentage having the potential to change care. We recommend that pilot testing and system optimization should be done before an ASR system is deployed in a real clinical setting.

Acknowledgements

This work was supported by a grant from the National Library of Medicine (Grant No. T15LM07117). I would also like to acknowledge the work of co-author Anne Belsito for developing the implementation of the software interface between the Gopher™ and Dragon's NaturallySpeaking™ software. I would also like to acknowledge Dr. Susan Perkins for her assistance with computing the concordance rates for errors among the three raters. We also wish to acknowledge Dragon Systems Inc. and Korteam International (now Breveon Inc.) for their generous technical support during this study.

References

[1] A. Zafar, M. Overhage, C. McDonald, Continuous speech recognition for clinicians, JAMIA 6 (3) 195.

[2] S.M. Borowitz, Computer-based speech recognition as an alternative to medical transcription, JAMIA 8 (1) 101.

[3] W.A. Lea, Problems in predicting performances of speech recognizers, D. Pallett (Ed.), in: Proceedings of the Workshop on Standardization for Speech I/O Technology, National Bureau of Standards, 1983, pp. 15—24.

[4] E.G. Devine, S.A. Gaehde, A.C. Curtis, Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports, JAMIA 7 (5) 462.

[5] S.J. Herman, Accuracy of a voice-to-text personal dictation system in the generation of radiology reports, Am. J. Radiol. 165 (1995) 177—180.

[6] A.H. Robbins, D.M. Horowitz, M.K. Srinivasan, M.E. Vincent, K. Shaffer, N.L. Sadowsky, M. Sonnenfeld, Speech-controlled generation of radiology reports, Radiology 164 (1987) 569—573.

[7] R.M. Ramaswamy, G. Chalijub, O. Esch, D.D. Fanning, E. vanSonnenberg, Continuous speech recognition in MR imaging reporting, Am. J. Radiol. 174 (2000) 617—622.

[8] B.L. Holman, P. Alibadi, S.G. Silverman, B.N. Weissman, L.E. Rudolph, E.F. Fener, Medical impact of unedited preliminary radiology reports, Radiology 191 (1994) 519—521.

[9] C.J. McDonald, J.M. Overhage, W.M. Tierney, P.R. Dexter, D.K. Martin, J.G. Suico, A. Zafar, et al., The regenstrief medical record system: a quarter century experience, Int. J. Med. Inf. 54 (1999) 225—253.

[10] D.N. Mohr, D.W. Turner, G.R. Pond, J.S. Kamath, C.B. De Vos, P.C. Carpenter, Speech recognition as a transcription aid: a randomized trial with standard transcription, JAMIA 10 (1) (Jan/Feb. 2003) .

Non-medical and DARPA references (see http://www.nist.gov/speech/publications/index.htm)

[11] P. Zhan, S. Wegmann, S. Lowe, Dragon systems' 1997 Mandarin broadcast news system, 1998 DARPA Broadcast News and Understanding Workshop, Lansdowne, Virginia, February 1998.

[12] J. Fiscus, A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER), in: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 1997.

[13] L. Chase, ''Error Responsive Feedback Mechanisms for Speech Recognizers'', PhD Thesis, CMU 1997, (http://www.ri.cmu.edu/pubs/pub_444.html). Also see the CMU Error Analysis Toolkit (http://www-2.cs.cmu.edu/afs/cs/user/lindaq/ERA/) and the CMU Statistical Language Modeling Toolkit (http://www.speech.cs.cmu.edu/SLM_info.html).

[14] Schaaf, T, Kemp, T. Confidence measures for spontaneous speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1997 (http://www.cs.auckland.ac.nz/∼pat/760_2001/seminars/Daniel_DY_NGU.htm).

[15] D. Vergyri, Use of word level side information to improve speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000.

[16] E. Eide, H. Gish, P. Jeanrenaud, A. Mielke, Understanding and improving speech recognition performance through the use of diagnostic tools, in: Proceedings of ICASSP95, Detroit, Michigan, May 1995, vol 1, p. 221ff.

[17] Sheryl Young, Detecting misrecognitions and out-of-vocabulary words, in: Proceedings of ICASSP94, Adelaide, Australia, April 1994. p. II21ff.

[18] T. Kemp, A. Jusek, Modelling unknown words in spontaneous speech, in: Proceedings of ICASSP96, Atlanta, Mai 1996, p. 530ff.