Noise in Speech-to-Text Voice: Analysis of Errors and Feasibility of Phonetic Similarity for Their Correction

Hanna Suominen
NICTA / Locked Bag 8001,
Canberra ACT 2601, Australia
The Australian National University
University of Canberra
hanna.suominen@nicta.com.au

Gabriela Ferraro
NICTA / Locked Bag 8001,
Canberra ACT 2601, Australia
The Australian National University
gabriela.ferraro@nicta.com.au

Abstract

In Australian healthcare, failures in information flow cause over one-tenth of preventable adverse events and are tangible in clinical handover. Regardless of a good verbal handover, anything from two-thirds to all of this information is lost after 3–5 shifts if notes are taken by hand or not taken. Speech to text (SST) and information extraction (IE) have been proposed for taking the notes and filling in a handover form with extrapolated evaluations from related studies promising over 90 per cent correctness for both STT and IE. However, this cascading evokes a fruitful methodological challenge: the severe implications that errors may have in clinical decision-making call for superiority in STT; the correctness percentage measured in a peaceful laboratory is decreased to 77 by noise in clinical practise; and the STT errors multiply when cascaded with IE. We provide an analysis of STT errors and discuss the feasibility of phonetic similarity for their correction in this paper. Our data consists of one hundred simulated handover records in Australian English with STT recognising 73 per cent of the $7,277$ words (1 h 8 min 5 s) correctly. In text relevant to the form, 836 unique error types are present. The most common errors include inserting *and*, *in*, *are*, *arm*, *is*, *a*, *the*, or *am* ($5 \leq n \leq 94$), deleting *is* ($n = 17$), and substituting *and*, *obs are*, *2*, *he with in*, *also*, *to*, or *and she* ($7 \leq n \leq 11$), respectively. Eighteen per cent of word substitutions sound exactly the same as the correct word and 26 per cent have a similarity percentage above 75. This encourages using phonetic similarity to improve STT.

1 Introduction

Fluent *information flow* is important in any information-intensive area of decision making, but critical in healthcare. Clinicians are responsible for making decisions with even life-and-death impact on their patients' lives. The flow is defined as links, channels, contact, or communication to a pertinent person or people in the organisation (Glaser et al., 1987). In Australian healthcare, failures in this flow are associated with over one-tenth of preventable adverse events (ACS, 2008; ACS, 2012). Failures in the flow are tangible in *clinical handover*, that is, when a clinician is transferring professional responsibility and accountability, for example, at shift change (AMA, 2006). Regardless of verbal handover being accurate and comprehensive, anything from two-thirds to all of this information is lost after three to five shifts if no notes are taken or they are taken by hand (Pothier et al., 2005; Matic et al., 2011).

There is a proposal to use a semi-automated approach of *speech to text* (STT) and *information extraction* (IE) for taking the handover notes (Suominen et al., 2013). First, a STT (a.k.a. speech recognition) engine converts verbal information into written, free-form text. Then, an IE system fills out a handover form by automatically identifying relevant text-snippets for each slot of the form. Finally, this pre-filled form is given to a clinician to proof and sign off.

The semi-automated approach evokes an STT challenge. First, the correctness of STT is challenged by background noise, other people's voices, and other characteristics of clinical practise that are far from a typical setting in a peaceful office. Second, the STT errors multiply when cascaded with IE. Third, correctness in cascaded STT and IE needs to be carefully evaluated as excellent, because of the severe implications that errors may have in clinical decision-making. In

Hanna Suominen and Gabriela Ferraro. 2013. Noise in Speech-to-Text Voice: Analysis of Errors and Feasibility of Phonetic Similarity for Their Correction. In *Proceedings of Australasian Language Technology Association Workshop*, pages 34−42.

summary, the original voice (i.e., information) in the big noise from clinical setting and STT errors needs to be heard.

Motivated by this challenge, we provide an analysis of STT errors and discuss the feasibility of *phonetic similarity* for their correction in this paper. Phonetic similarity (PS, a.k.a phonetic distance) addresses perceptual confusion between speech sounds and is used to improve STT (Mermelstein, 1976). To illustrate **phonetically similar words**, *PS measures can be seen as the* ***rites*** *of* ***righting writing****, that is* ***right****.*

The rest of the paper is organised as follows: In Section 2, we provide background for clinical STT and IE. In Section 3, we describe our simulated handover data, STT methods, PS measures, and analysis methods. In Section 4, we present the results of the error analysis and discuss the feasibility of phonetic similarity for error correction. In Section 5, final conclusions and directions for future work are given.

2 Background

In clinical STT, different engines give comparable results and can reach over 90 per cent of the words being correct. A comparison on the same dataset shows the mean correctness percentages of 85–86; 85–87; and 90–93 for Dragon Medical 3.0; L&H Voice Xpress for Medicine 1.2, General Medicine; and IBM ViaVoice 98, General Medicine, respectively (Devine et al., 2000). The dataset consists of four medical report entries (two progress notes, one assessment summary, and one discharge summary) and twelve US English male physicians.

Only 30–60 min tailoring to a given voice improves the correctness percentage up to 99 but in a preliminary evaluation of STT with minimal tailoring, Australian English, six simulated handover cases (over $1,200$ words of continuous free-form text), and Dragon Medical 11.0, the percentage is 79, 64, and 54 for a native male physician, native female nursing scientist, and Spanish-accented female nurse, respectively (Suominen et al., 2013). The percentages for tailored STT originate from experiments on the aforementioned four medical report entries and twelve US English male physicians; 47 emergency-department charts and two US English physicians (Zick and Olsen, 2001); and 206 surgical pathology reports, seven Canadian English pathologists, a researcher with an accent (Al-Aynati and Chorneyko, 2003).

However, these correctness percentages, measured in peaceful laboratory settings, are challenged by noise in clinical practise. On eight voices, a total of about 3,600 typical short anaesthesia comments in Danish, and with noise being present, only 77 per cent of words are correct (Alapetite, 2008).

The review (Meystre et al., 2008) discusses 174 studies from 1995 to 2008 on clinical IE. It concludes that the quality of these systems has gradually improved, exceeding the F1-measure (i.e., the harmonic mean of the proportion of slots that the system filled correctly and the proportion of snippets that the system extracted from those it should have extracted) of 90 per cent in several cases. These systems mostly focus on chest and other types of radiography reports, echocardiogram reports, discharge summaries, and pathology reports. Their typical tasks include extracting codes; enriching or structuring the content and utility of the electronic health record, especially to support computerised decision-making; surveillance; supporting research; de-identification of clinical text; and terminology management.

3 Materials and Methods

3.1 Materials

The dataset of 100 simulated handover records used in this study was created as follows.

First, a senior researcher in clinical language processing (i.e., HS) imagined an *Australian medical ward*. With an aim for balance in patient types, she created simulated profiles of 25 *cardiac*, 25 *neurological*, 25 *renal*, and 25 *respiratory patients* of the ward. Each imaginary *profile* included a photo from a free-to-use gallery, name, age, admission story, in-patient time, and the familiarity of this patient to the nurses giving and receiving the handover (Fig. 1).

Second, a registered nurse with over twelve years experience from clinical nursing was hired to create nursing-handover records for the hundred profiles as *written, free-form text records*, *structured forms*, and *spoken free-form text records* (Fig. 1, Table 1). She spoke Australian English as a second language and was originally from Philippines. In the creative writing task, HS guided her to write realistic reports in the role of the nurse giving the handover. In the structuring task, HS guided her to use these written, free-text records to identify text snippets relevant to the slots of the handover form by using *Knowtator* (Ogren, 2006). The handover form was developed in collaboration with HS and nurse. It was based on international standards and practical experiences. The identification task was multi-class classification, that is, each word belonged to precisely one or none of the slots. In the speaking task, HS guided the nurse to read the written, free-text records out loud in the role of the nurse giving the handover. The digital recorder and microphone were *Olympus WS-760M* (200 AUD) and *Olympus ME52W* (lapel, noise cancelling, 15 AUD), previously shortlisted as producing a superior percentage of correct words in STT (i.e., up to 79) (Suominen et al., 2013) .

3.2 STT Methods

Dragon Medical 11.0 was used to convert the audio files to written, free-form text records. Audio files were converted from stereo to mono tracks and from WMA to WAV files on *Audacity 2.0.3*. Dragon was initialised for the *Age* of *22-54 years* and *Accent* of *Australian English*, and tailored to the nurse's voice by her reading the document of *The Final Odyssey* using the aforementioned recorder and microphone (3, 893 words, 29 min 22 s). Tailoring was left minimal since it could limit comparability with other studies and might not be feasible for every clinician in practise.

Dragon *vocabularies* of *general*, *medical*, *nursing*, *cardiology*, *neurology*, and *pulmonary disease* were compared. The *SCLITE scoring tool of the Speech Recognition Scoring Toolkit 2.4.0* was used to analyse correctly recognised, substituted, inserted, and deleted words. The reference standard in all comparisons consisted of the original written reports (i.e., not transliterations by hand) where punctuation was removed and capitalisation was not considered as a distinguishing feature.

The vocabulary resulting in the best correctness (i.e., *nursing* with both highest mean (73%) and lowest standard deviation (SD, 7%) of correct words, Fig. 2) was chosen for the error analysis. In 74 out of 100 cases, this vocabulary gave the largest number of correct words. With 25 cardiac (neurological) [respiratory] patients, the matching vocabulary (i.e., cardiology, (neurology), and [pulmonary disease]) gave more correct words than any other vocabulary only 3 (4) [0] times. The matching vocabulary gave more correct words than the nursing vocabulary only 4 (3) [6] times.

Name: Leila Sonya Da Silva
Age: 34 years
Admission story: Leila has difficulties to control her diabetes. Her blood sugars tend to climb up too high but in the morning, the values are too low. Her diabetes was diagnosed when she was 6. She suffers from hypertension too but has had no medication to it yet.
At medical ward: She has been at the ward for a day. She is new both to you and the next nurse.

Leila sonya Da silva, bed 5, 34 under Dr Liu, came in for management of her diabetes. With history of type 1 DM since childhood and HPN. She is still for referral to the diabetic educator and she is self caring with her own BGLS and insulin. Her insulin is on a sliding scale insulin and on variable dose so just ask the doctor for the next dose depending on her blood sugar. Her BGL trend used to be high during the AM.so still need the team to review for that.Her BP is not so bad and of a high normal range and still for review.otherwise she is pretty much self caring and ambulant and there are no other problems noted.

Heading	*Slots*
Introduction	Room, Bed, Dr, Name, Age, Gender, Allergy, Admission reason/diagnosis, Chronic condition, Problem history
My shift	Status, Contraption, Activities of daily living, Input/diet, Output/diuresis/ bowel movement, Wounds/skin, Risk management, Other observation
Medication	Medicine, Dosage, Status
Appointment	Description, Place, Time, Status, Clinician
Future	Goal/task to be completed/ expected outcome, Alert/warning/abnormality, Care/discharge/transfer plan

Figure 1: A profile, report, and form structure

Table 1: Descriptive statistics of the records, words (w), and inside words (i)

Patient type (n)	cardiac (25)	neurological (25)	renal (25)	respiratory (25)	*All* (100)
Record length					
Min–Max [w]	19–162	26 – 106	29–149	31–209	19–209
Mean (SD) [w]	70 (37)	60 (22)	71 (33)	83 (39)	71 (34)
Min–Max [s]	17–89	16 – 60	16–77	20–97	16–97
Mean (SD) [s]	44 (19)	38 (13)	36 (17)	46 (18)	41 (17)
n of w's (uniq.)	1,795 (556)	1,545 (500)	1,818 (496)	2,119 (604)	7,277 (1,304)
Top 1 (*n*)	and (95)	and (64)	and (88)	and (100)	and (347)
Top 2 (*n*)	he (59)	is (60)	is (72)	is (69)	is (256)
Top 3 (*n*)	for (58)	he (54)	he (69)	on (63)	he (243)
Top 4 (*n*)	is (55)	she (38)	in, she (46)	he (61)	in (170)
Top 5 (*n*)	the, with (43)	in (35)		with (51)	for (163)
Top 6 (*n*)		with (34)	the (38)	in (49)	with (162)
Top 7 (*n*)	in (40)	on (33)	with (34)	for (43)	she (151)
Top 8 (*n*)	to (32)	for (31)	came (32)	she (42)	on (141)
Top 9 (*n*)	of (30)	to (29)	for (31)	the (37)	the (138)
Top 10 (*n*)	came (27)	came (24)	to (30)	to (33)	to (124)
n of i's (uniq.)	1,140 (447)	1,006 (397)	1,086 (408)	1,305 (483)	4,547 (1,106)
Top 1 (*n*)	he (57)	he (52)	he (63)	and (51)	he (220)
Top 2 (*n*)	for (47)	she (35)	she (39)	he (48)	she (139)
Top 3 (*n*)	and (26)	for (25)	and (34)	she(40)	and (131)
Top 4 (*n*)	bed, she (25)	dr (22)	bed, is (24)	for (27)	for (118)
Top 5 (*n*)		and, old (20)		dr (25)	dr (88)
Top 6 (*n*)	dr (23)		to (23)	is, on, to (20)	to (84)
Top 7 (*n*)	to (22)	bed, to (19)	old, yrs (21)		bed (80)
Top 8 (*n*)	the (21)				is (76)
Top 9 (*n*)	her, old (18)	yrs (17)	all (20)	room (18)	old (72)
Top 10 (*n*)		her, is (16)	for (19)	of (16)	all, her (61)

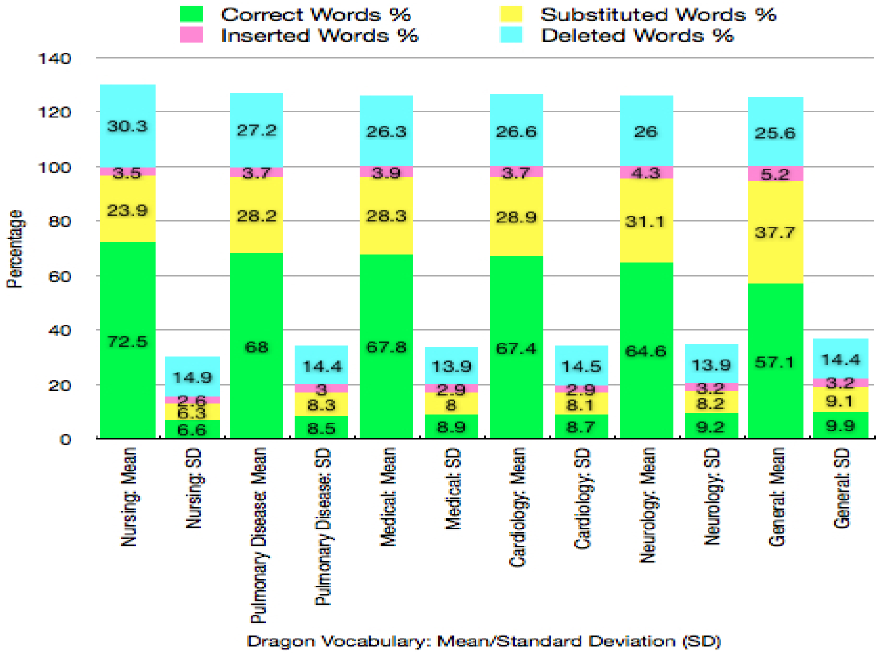

Figure 2: STT with different vocabularies: mean and SD over the 100 records

3.3 PS Measures

Measuring PS is relevant in speech processing, spelling correction, dialectometry, historical distance between sounds, and many other contexts. PS measures quantify the similarity between speech forms (e.g., words) on the basis of their sounds. Usually they consist of two steps (Kondrak, 2002): First, *words are transcribed into their phonetic representation*. Second, a *weighted* or *unweighted edit-distance* is applied to calculate the similarity between the transcriptions. Recent approaches weight the edit distance by hand on the basis of linguistic knowledge (Kondrak, 2000) or automatically using learning algorithms (Mann and Yarowsky, 2001; Kondrak, 2002; Mackay and Kondrak, 2005).

We calculated *PS of substitutions errors* from a STT engine. Similarly to other studies (Kaki et al., 1998; Jeong, 2004; Pucher et al., 2007), our hyphotesis was that substitutions that sound similar to the reference standard can be solved by applying a correction metric that combines a generator of sound-alike words with principles for distributional semantics. In other words, a good correction candidate was a word that sounds similar to the reference standard and fullfills its usage context. As a first step towards the creation of such correction metric, we implemented a procedure for selecting *(quasi-)homonym substitutions* (i.e., sound (almost) the same but have different meaning) based on phonetic distance.

We built a simple PS measure, which *combines a sound-alike algorithm with edit distance*. To transcribe the words into a phonetic representation, we used the *Double Metaphone* (DMetaphone) phonetic encoding algorithm (Philips, 2000) which is part of the *Methaphone* family (Philips, 1990). We chose DMetaphone, because it approximates accented English from Slavic, Germanic, French, Spanish, among others languages. DMetaphone returned for each word an aproximation of its sound instead of a sequence of phonemes. It translated each consonant into a limited set of characters where similar sounds are represented by the same character (e.g., *b* and *p* both sound like *p*). To calculate the similarity between the encoded words, we applied the unweighted edit-distance. This computed the minimum number of edit operations (i.e., substitutions, insertions, and deletions) required to transform an encoded word into another.

3.4 Analysis Methods

We used *content analysis* (Stemler, 2001) to analyse STT errors quantitatively and qualitatively. The correct, substituted, inserted and deleted words were defined by the SCLITE scoring tool.

For the PS discussion, we performed two experiments. First, we computed PS for *single-word substitutions* (e.g., *four–for*), in which the first word is the STT word and the second word is from the reference standard. Each word was encoded into its DMetaphone value using the *Apache Commons Metaphone* utility. The edit distance between the encoded words was calculated using the open source *Simmetric* library from Sheffield University. Second, we computed PS for *multi-word substitutions* (e.g., *doctors signed–dr san*). Because DMetaphone is designed to encode a single word at a time, each word in a multi-word concept was individually encoded into its metaphone value, encoded words were combined as sequences, and the edit distance was used to calculate the similarity between the sequences.

In all analyses and experiments, we used the entire dataset and the subset that affects the IE system (i.e., *inside* refers to text identified as relevant to the slots of the handover form).

4 Results and Discussion

Fifteen per cent (18%) of all unique substitutions (unique inside substitutions) sound exactly the same as in the reference standard and 23 per cent (26%) have a similarity score above 75 per cent (Tables 2&3). Consequently, substitutions with a high PS value can be considered as candidates for error correction.

In text relevant to the handover form, 836 unique error types are present (Table 2). The most common of them include inserting *and*, *in*, *are*, *arm*, *is*, *a*, *the*, or *am* ($5 \leq n \leq 94$), deleting *is* ($n = 17$), and substituting *and*, *obs are*, *2*, *he with in*, *also*, *to*, or *and she* ($7 \leq n \leq 11$), respectively.

Five types of substitution errors are present:

1. proper names;
2. singular vs. plural forms;
3. use of abbreviations in the reference standard and complete forms in STT;
4. systematic differences between the reference standard and STT (e.g., Australian (reference) vs. US (STT) spelling and writing

Table 2: Correct, substituted, inserted, and deleted single-words
These descriptive statistics also include cases where STT deleted (inserted) a word (i.e., white space is computed as a word). In the top substitutions, the first word is the STT word and the second from the reference standard.

	Correct words	Substituted words	Inserted words	Deleted words
All (Inside)	5,270 (3,237)	1,685 (1,132)	2,111 (1,541)	322 (178)
	Unique correct	Unique substitutions	Unique insertions	Unique deletions
All (Inside)	839 (710)	1,187 (827)	449 (371)	154 (93)
Inside	Top correct (n)	Top substitutions (n)	Top insertions (n)	Top deletions (n)
1	he (178)	years yrs (48)	and (210)	is (20)
2	she (134)	in and (22)	is (136)	are (13)
3	for (112)	one 1 (17)	in (106)	and (11)
4	dr (87)	also obs (12)	she (71)	s (8)
5	and (80)	to 2 (12)	are (58)	obs (6)
6	old (71)	and he (1)	all (45)	of (5)
7	to (70)	he his (9)	arm (44)	bed (4)
8	bed (64)	also are (7)	for (43)	her (4)
9	all (56)	ambien ambulant (6)	the (37)	4 (3)
10	the (55)	ambulating ambulant (6)	he (35)	all (3)
11	stable (54)	antibiotics abs (6)	that (34)	fbc (3)
12	is (52)	desilva de (5)	a (27)	for (3)
13	her (50)	for 4 (5)	her (19)	got (3)
14	of (44)	hypertension hpn (5)	eats (15)	he (3)
15	on (38)	in nil (5)	on (15)	silva (3)
16	pain (33)	she he (5)	also (14)	the (3)
17	with (31)	tomorrow tom (5)	am (12)	to (3)
18	his (27)	ultrasound us (5)	does (11)	a (2)
19	self (27)	and nil (4)	bed (10)	hdx (2)
20	caring (26)	george jorge, his he, is are, is obs, is s, iv ivabs, lee li, p prn, x xray (4)	s (10) to (10)	normal (2) review(2)

numbers as digits (reference) vs. letters (STT)); and

5. misspelling/typos in the reference standard (e.g., *feeling* vs. *feelling* or *arrhythmia* vs. *arrythmia*).

Substitutions and insertions are the most common error types, both in all data and within text identified as relevant to the slots of the handover form. The majority of the top insertions and deletions corresponds to functional words, (lexemes with little semantic meaning such as determiners, prepositions, auxiliary verbs and pronouns).

According to our PS measure, for the set of all word substitutions, 23 per cent have a similarity percentage above 75 and in 15 per cent of these highly similar cases the STT and reference words sound exactly the same (Tables 3&4). When experimenting with the set of inside substitutions, 26 per cent have a similarity percentage above 75 and in 18 per cent of these highly similar cases the STT and reference words sound exactly the same. Thus, around a fourth of the substitution errors can be considered as candidates for their correction.

A proper name is included in 24 per cent of substitutions and 3 per cent of them sound exactly the same (e.g., *Lane* vs. *Laine* or *Lee* vs. *Li*). Correcting this is critical in the healthcare context. Different spellings of the same word are not uncommon (e.g., *Johnson* vs. *Johnsson* or *organised* vs. *organized*). Aproximately 2 per cent of the substitutions that sound the same are due the difference between singular and plural forms of the same lemma (e.g., *investigation* vs. *investigations* or *fibrosis* vs. *fibroses*).

As expected, the number of substitutions that sound similar is quite low (Table 4). Only 14 per cent of single-word substitutions are minimally 75 per cent similar. For multi-word substitutions, the

Table 3: Top errors within inside words
- refers to a single white space and the total number of incorrect multi-words is 1,204 and 836 of them are unique

STT	*reference*	*n*	*STT*	*reference*	*n*
and	-	94	in the	nil -	4
years	yrs	48	iv antibiotics	ivabs -	4
in	-	25	lee	li	4
are	-	21	x ray	xray -	4
-	is	17	-	fbc	3
arm	-	11	and	he	3
in	and	11	and there is	- - -	3
is	-	11	antibiotics	abs	3
a	-	9	arm she	- -	3
also -	obs are	8	ii	2	3
to	2	8	is	s3	
and she	he -	7	is a	- -	3
the	-	7	kinsey	kenzie	3
am	-	5	lane and	laine -	3
hypertension	hpn	5	our	-	3
-	of	4	she	he	3
ambulating -	ambulant and	4	tomorrow	tom	3
and she is	- - -	4	ultrasound	us	3

top similarity percentage is 72. After removing instances that contained an empty column, 651 unique substitution types are present in the data.

For PS from 0.74 to 0.50, the single-word substitutions are still phonetically close to the reference (e.g., *cause* vs. *course*, *weeks* vs. *weak*, and *from* vs. *for*) which suggest that they might be also considered as secondary correction candidates in future experiments. When PS is below 0.5, errors are heterogeneous, meaning that some of them still sound a bit similar (e.g., *bed* vs. *the*) but others sounds completely different (e.g., *energies* vs. *physiotherapist*), and should not be taken into account for their correction based on this PS approach. Fifty per cent of the substitution errors occur with words shorter than 4 characters. These short words are obviously more difficult for STT than longer words.

Not all substitutions that sound similar to the reference should be considered as potential candidates for error correction. For example, errors due to abbreviations, typos, and spelling variations represent 9 per cent of the errors, and are not strictly speaking STT errors. This is because the original written records, and not careful transliterations by hand, were used as a reference standard. The use of abbreviations in the writing environment and the use of the complete form in STT seems natural for people but creates an inconsistency in the error analysis. For example, the nurse is always using *yrs* when writing instead of *year*, *obs* instead of *observations*, *his K* instead of *his potassium*, among others.

5 Conclusion and Future Work

A detailed error analysis is a crucial step in the development of pipeline applications (i.e., applications that cascade methods) similar to the one described in this paper. We have found that a substantial amount of STT errors occurs with words that are phonetically similar to each other. Consequently, using an error correction method based on PS seems appropriate in reducing the error rate.

As the first step towards the correction method, we have assessed a PS measure that calculates the similarity between words. This component will be used in the future as a post-processing method to select errors for their correction. Single-word substitutions are more suitable for this post-processing than insertion and deletion errors. However, we address the inserted and deleted words indirectly via the multi-word substitution and white-space analyses (Tables 2–4).

Based on the presented analysis, the correction method will take into account the following four characteristics:

Table 4: Examples of the sound-alike substitutions

Analysis	*STT*	*reference*	*phoneSim*
Single-word	Gaylor	Gayler	1.0
	dialyses	dialysis	1.0
	results	result	1.0
	harrowed	Harrod	1.0
	cord	GORD	1.0
	ambulance	ambulant	1.0
	arrhythmia	arrythmia	1.0
	Lane	Laine	1.0
	doctors	doctor	1.0
	ambulating	ambulant	1.0
	wheelie	wheely	1.0
	years	yrs	1.0
	and/ even	endone/ eventhough	0.75
	heart/ relater	heartburn/ later	0.75
	every/ state	everytime/ stent	0.75
	menders/ arrive	Mendez/ arrived	0.75
Multi-word	george desilva s	jorge de silva	0.72
	in ampulla	and ambulant	0.72
	aspergilloses are she	aspergillosis he	0.71
	blanford	plan for	0.71
	can assume	cannot seem	0.71
	coronae idd sees	coronary artery disease	0.71
	you ve am	if all	0.71
	flexing	clexane	0.71
	one keay	wound care	0.71
	do explained	explain	0.70
	endo p r n	endone prn	0.70
	haemodialyses am	heamodialysis	0.70
	racquel saw iris date dino	raquel soares caetano	0.68
	this orders	disorders	0.66
	cystic fibroses and	cyctic fibrosis	0.66

1. detection and correction of errors in proper names;
2. difference between single-word and multi-word errors;
3. spelling correction strategies; and
4. grammar checking to ensure correctness.

Even though only one-fourth of all substitution errors could be considered as correction candidates, every corrected word is one less potential error in clinical decision-making.

Acknowledgments

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. NICTA is also funded and supported by the Australian Capital Territory, the New South Wales, Queensland and Victorian Governments, the Australian National University, the University of New South Wales, the University of Melbourne, the University of Queensland, the University of Sydney, Griffith University, Queensland University of Technology, Monash University and other university partners. We acknowledge the contribution of Maricel Angel, RN in creating our dataset as a NICTA employee.

References

ACSQHC, Australian Commission on Safety and Quality in Health Care, 2008. *Windows into Safety and Quality in Health Care,* `goo.gl/wB0XZ1`.

ACSQHC, Australian Commission on Safety and Quality in Health Care, 2012. *The OSSIE Guide to Clinical Handover Improvement,* `goo.gl/IvS7dc`.

M Al-Aynati and K Chorneyko. 2003. Comparison of voice-automated transcription and human transcription in generating pathology reports. *Archives of Pathology and Laboratory Medicine*, 127(6):721–5.

A Alapetite. 2008. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics*, 77(1):68–77.

AMA, Australian Medical Association, 2006. *Safe Handover: Safe Patients,* `goo.gl/9U8wjm`.

E Devine, S Gaehde, and A Curtis. 2000. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. *Journal of the American Medical Informatics Association: JAMIA*, 7(5):462–8.

S Glaser, S Zamanou, and K Hacker. 1987. Measuring and interpreting organizationalculture. *Management Communication Quarterly*, 1(2):173–98.

Minwoo Jeong. 2004. Using higher-level linguistic knowledge for speech recognition error correction in a spoken q/a dialog. In *Proceedings of the HLT-NAACL special workshop on Higher-Level Linguistic Information for Speech Processing*, pages 48–55.

S Kaki, E Sumita, and H Iida. 1998. A method for correcting errors in speech recognition using the statistical features of character co-occurrence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 653–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

G Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 288–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

G Kondrak. 2002. Determining recurrent sound correspondences by inducing translation models. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

W Mackay and G Kondrak. 2005. Computing word similarity and identifying cognates with pair hidden Markov models. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 40–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

G Mann and D Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

J Matic, P Davidson, and Y Salamonson. 2011. Review: bringing patient safety to the forefront through structured computerisation during clinical handover. *Journal of Clinical Nursing*, 20(1–2):184–9.

P Mermelstein. 1976. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 116:374–88.

S Meystre, G Savova, K Kipper-Schuler, and J Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. review. *Yearbook of Medical Informatics*, pages 128–44.

P Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–5, Morristown, NJ, USA. Association for Computational Linguistics.

L Philips. 1990. Hanging on the Metaphone. *Computer Language*, 7(12 (December)).

L Philips. 2000. The Double Metaphone search algorithm. *C/C++ Users Journal*, 18(5), June.

D Pothier, P Monteiro, M Mooktiar, and A Shaw. 2005. Pilot study to show the loss of important data in nursing handover. *British Journal of Nursing*, 14(20):1090–3.

M Pucher, A Türk, J Ajmera, and N Fecher. 2007. Phonetic distance measures for speech recognition vocabulary and grammar optimization. In *Proc. 3rd congress of the Alps Adria Acoustics Association, Graz.*

S Stemler. 2001. An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17).

H Suominen, J Basilakis, M Johnson, P Sanchez, L Dawson, L Hanlen, and B Kelly. 2013. Preliminary evaluation of speech recognition for capturing patient information at nursing shift changes: accuracy in speech to text and user preferences for recorders. In *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis*, Sydney, NSW, Australia.

R Zick and J Olsen. 2001. Voice recognition software versus a traditional transcription service for physician charting in the ed. *The American Journal of Emergency Medicine*, 19(4):295–8.