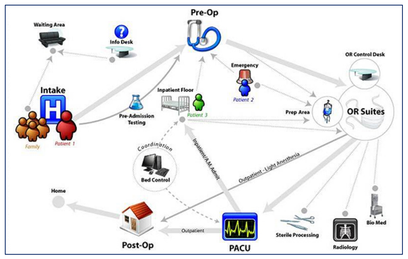
Evaluation of Google's Voice Recognition and Sentence Classification for Health Care Applications

Majbah Uddin, Nathan Huynh, and Jose Vidal
University of South Carolina

Kevin Taaffe, Lawrence Fredendall, and Joel Greenstein
Clemson University

Research Objectives

1. To examine the use of voice recognition technology in perioperative services (Periop) to enable Periop staff to record workflow milestones using mobile technology.
2. To allow the Periop staff to provide care without being interrupted with data entry and querying tasks.
3. To investigate the effectiveness of different post-process algorithms to improve the performance of Google's speech recognizer.

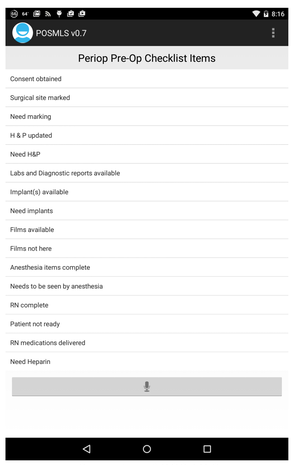


Perioperative process flows [1]

Smart-App Development

A smart-app named Perioperative Services Mobile Learning System (POS-MLS) was developed by the research team using Android API (Level 19). The test devices included Nexus 4, 7, and 10. The voice recognition was enabled by the Android platform with its built-in speech recognizer.

Tested mobile devices



Screenshot of POS-MLS

Experimental Set-Up

The following three post-processing classifiers were tested in this study. For bag-of-sentences, a many-to-few mapping was created between phrases returned by the speech recognizer and phrases needed to recognize. SVM and MAXENT algorithms were implemented using RTextTools [2].

1. Bag-of-sentences
2. Support vector machine (SVM) [3]
3. Maximum entropy (MAXENT) [4]

We conducted 16 experiments that were designed to test the ability of the app to recognize the Pre-op checklist items correctly using voice.

- Every phrase was spoken five times for all three levels (i.e., Google-only, Train-5, and Train-10).
- We have a total of 80 observations for each phrase at all three levels.

Phrases	Training Repetitions	Testing Repetitions	Post-Processing Methods: Google-only	Bag-of-sentences	Support Vector Machine	Maximum Entropy
As-is	0	5	✓			
	5	5		✓	✓	✓
	10	5		✓	✓	✓
Reduced	0	5	✓			
	5	5		✓	✓	✓
	10	5		✓	✓	✓
Personalized	0	5	✓			
	5	5		✓	✓	✓
	10	5		✓	✓	✓

Summary of experiments

Correctness by level

- Statistically significant differences in recognition correctness between training levels were identified for 11 of the 16 phrases.
- Seven of the reduced phrases were identified correctly less often than the corresponding as-is phrases.
- For every phrase, when the Google-only approach did not recognize an as-is or reduced phrase at least half the time, both training levels (Train-5 and Train-10) improved recognition correctness.

As-is Phrase	% Correct Classification (Number of Correct Classification)			*p*-Value
	Google-only	Train-5	Train-10	
Consent obtained	66.3 (53)	73.8 (59)	75.0 (60)	0.414
Surgical site marked	28.8 (23)	53.8 (43)	57.5 (46)	<0.001
Need marking	31.3 (25)	65.0 (52)	63.8 (51)	<0.001
H&P updated	40.0 (32)	62.5 (50)	70.0 (56)	<0.001
Need H&P	11.3 (9)	50.0 (40)	53.8 (43)	<0.001
Labs and diagnostic reports available	18.8 (15)	41.3 (33)	43.8 (35)	0.001
Implant(s) available	65.0 (52)	65.0 (52)	75.0 (60)	0.292
Need implants	75.0 (60)	75.0 (60)	76.3 (61)	0.978
Films available	57.5 (46)	66.3 (53)	70.0 (56)	0.237
Films not here	40.0 (32)	61.3 (49)	73.8 (59)	0.005
Anesthesia items complete	28.8 (23)	37.5 (30)	53.8 (43)	0.001
Need to be seen by anesthesia	37.5 (30)	62.5 (50)	65.0 (52)	<0.001
RN complete	8.8 (7)	81.3 (65)	76.3 (61)	<0.001
Patient not ready	86.3 (69)	91.3 (73)	78.8 (63)	0.079
RN medications delivered	3.8 (3)	50.0 (40)	67.5 (54)	<0.001
Need heparin	5.0 (4)	46.3 (37)	60.0 (48)	<0.001

Comparison of percent correct and number of correct classifications at different training levels for as-is Phrases

Correctness by Phrase Type and Classifier

Correctness by Phrase Type

- All differences in recognition correctness as a function of training were significant ($p < 0.05$), with the exception of the difference between Train-5 and Train-10 for the as-is phrase ($p = 0.129$).
- The average recognition correctness for the as-is phrase was 61% when the app was trained with at least five repetitions.
- The correctness percentages for the reduced phrase, for all three levels, was always greater than that of the as-is phrases (38% vs. 47%, 61% vs. 63%, etc.).
- Personalized phrases were identified correctly more frequently than as-is and reduced phrases for pre-op checklist items within a voice recognition application, suggesting that personalized phrases may be more suitable.

	Google-only		Train-5		Train-10				
	Average	Std. dev.	Average	Std. dev.	Average	Std. dev.	*p*-Value[a]	*p*-Value[b]	*p*-Value[c]
As-is	37.7	11.2	61.4	17.9	66.3	18.7	<0.001	<0.001	0.129
Reduced	46.5	22.3	62.7	14.5	70.2	15.9	0.003	<0.001	<0.001
Personalized	53.8	22.7	72.3	16.2	78.7	12.7	<0.001	<0.001	0.002

[a] Test between Google-only and Train-5; [b] test between Google-only and Train-10; [c] test between Train-5 and Train-10

Test variable	*p*-Value: Google-only	Train-5	Train-10
As-is and Reduced	0.025	0.382	0.127
As-is and Personalized	<0.001	0.007	0.006
Reduced and Personalized	0.022	<0.001	0.003

Correctness by Classifier

- Classification using SVM and MAXENT algorithms improved classification correctness significantly more than the bag-of-sentences approach in most cases (5 out of 6).
- Train-5 with as-is phrases yielded the maximum average correctness for SVM of 82% and for MAXENT of 84%.
- Unlike the bag-of-sentences approach, increasing training repetitions did not lead to further correctness of classification.
- The MAXENT algorithm outperformed SVM for three different cases (as-is, using both Train-5 and Train-10, and personalized using Train-5 only).

		SVM		MAXENT				
		Average	Std. dev.	Average	Std. dev.	*p*-Value[a]	*p*-Value[b]	*p*-Value[c]
As-is	Train-5	81.9	11.8	84.0	9.4	<0.001	<0.001	0.018
	Train-10	80.9	8.7	83.8	7.7	<0.001	<0.001	0.022
Reduced	Train-5	78.6	14.1	80.2	9.9	<0.001	<0.001	0.166
	Train-10	77.4	15.5	79.1	13.1	0.004	<0.001	0.114
Personalized	Train-5	79.0	13.0	81.3	13.5	0.001	<0.001	0.015
	Train-10	76.7	14.5	80.6	11.6	0.292	0.222	0.052

[a] Test between Bag-of-sentences and SVM; [b] test between Bag-of-sentences and MAXENT; [c] test between SVM and MAXENT

Conclusions

- This study sought to identify a suitable algorithm to classify phrases in order to improve the performance of Google's speech recognizer to allow hands-free use of mobile technology.
- The as-is phrases and the Google-only speech recognizer used without any classifier had the lowest phrase recognition correctness in their respective settings.
- The use of reduced phrases or personalized phrases improved recognition correctness compared to using the as-is phrase.
- The use of two different post-process learning algorithms enhanced speech recognition correctness, compared to the post-process bag-of-sentences approach.
- Training (i.e., repetitions of phrases) significantly increased speech recognition correctness for all levels of post-processing.

References

1. http://www.pcts.com.
2. Jurka et al. (2013). RTextTools: A supervised learning package for text classification. *R Journal*, *5*(1), 6–12.
3. Cortes, & Vapnik (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.
4. Nigam et al. (1999). Using maximum entropy for text classification. presented at the *IJCAI-99* Workshop, Stockholm.

Acknowledgment

National Science Foundation WHERE DISCOVERIES BEGIN
GREENVILLE HEALTH SYSTEM
PALMETTO HEALTH
MUSC MEDICAL UNIVERSITY of SOUTH CAROLINA Changing What's Possible