Automated Assessment of Aphasic Speech Using Discrete Speech Recognition Systems.

Abstract

The need for computer-based aphasia treatment programs (CBATP) is increasing as viable options are sought for optimizing the quality, quantity and accessibility of therapy while also reducing its cost. A challenge in the development of CBATPs has been automated assessment of spoken responses. This feature would allow for provision of feedback to the patient and/or performance data to the speech-language pathologist. In this poster, we present data comparing the automated judgments of response accuracy of several systems with that of experienced clinicians. Patient data for this project was downloaded from the *AphasiaBank* database (McWhinney, Fromm, Forbes, & Holland, 2011).

Detailed Description

The need for computer-based aphasia treatment programs (CBATP) is increasing as viable options are sought for optimizing the quality, quantity and accessibility of therapy while also reducing its cost. A challenge in the development of CBATPs has been automated assessment of spoken responses. This feature would allow for provision of feedback to the patient and/or performance data to the speech-language pathologist.

Research Question:

- How accurately can a speech recognition (SR) system with three different grammar configurations classify responses produced by people with aphasia in a visual confrontation naming task as correct or incorrect?

Background:

Previously, researchers examined the use of SR software by people with aphasia for dictation, as an input mode to communication devices, and as an input mode for therapy software (Wade, Petheram, & Cain, 2001). More recent studies used speech recognition systems with large vocabulary, speaker dependent (requiring training) software systems designed for dictation tasks (Estes & Bloom, 2011; Dahl, Linebarger, & Berndt, 2008; Linebarger, Schwartz & Kohn, 2001; Wade, et al., 2001). Results from these studies were mixed. Overall, performance accuracy and the effort required to train the system were perceived as significant barriers.

The focus of this study is to determine the accuracy that can be achieved when using SR systems that **do not require training** (speaker independent) and are specifically designed for use with **non-continuous small-vocabulary speech.** "Non-continuous" means that words are spoken discretely with space between them, as would be the case in a visual confrontation naming task. "Small vocabulary" means that the system uses a predefined grammar that includes only tens to hundreds of words rather than the 100,000 plus words supported in large vocabulary systems. Prior research has indicated that non-continuous speech recognition is "more accurate and appropriate for speakers with breathing or speech difficulties, including those with dysarthria" (Wade et al., 2001). Small vocabulary systems are advantageous because of the ease of creating the grammar. For a comprehensive overview of speech recognition within the context of communication disorders, see Venkatagiri (2002).

Method/Approach
Speech samples from 62 persons with aphasia were downloaded from the *AphasiaBank* testing corpus (McWhinney, Fromm, Forbes, & Holland, 2011). The speech samples consisted of single-word responses to the short form of the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 2001). The *AphasiaBank* corpus includes data for 37 people classified as having Broca's aphasia however, we excluded data from 6 people: 4 produced very few productions and for 2 the BNT short form data was not included in the corpus. Data for all 11 people in the corpus who were classified with Wernicke's aphasia were also included. To make the number of fluent and non-fluent samples equal, data for the first 20 people classified with anomic aphasia were also included. .

The sound track from each patient's video recording was trimmed so that each sample contained the patient's best attempt for each test item. For the 31 non-fluent PwA analyzed so far, there were a total of 403 samples. Sound samples were played on an Apple MacBook Pro using Sound Studio software. Sound was passed into the Xoom input jack from the Apple headset jack.

The AT&T Speech Mashup speech recognition system was used (see Figure 1). This system is publicly available grammar-based recognition system. Three different grammars were tested:
1) 15 words from BNT short form
2) Any speech/sound + 15 words from BNT short form
3) 15 words from BNT short form + 45 additional words from the BNT long form.

Each of the 403 sounds was played into one of the three speech recognition grammars and the system's response was recorded by the PI as accepted or rejected. That is, if the target word was "house" and the system judged the word to be "house" then it was "accepted." If the system judged the word to be "comb" then the word was "rejected."

The 403 samples were then judged for correctness (correct/incorrect) by two experienced clinicians who were blinded to the speech recognition results.

Results (based on 31 non-fluent PwA; data for remaining 31 patients will be included in the poster)
Clinician inter-rater reliability was 84%. Clinician A was generally more accepting than Clinician B. Overall agreement of the 3 SR grammars with the two clinicians ranged from 66%-74% (See Table 1). The two 15-word grammars performed best for PwAs for whom clinicians judged as 30% to 70% accurate. For the PwAs with <30% accuracy the system tended to falsely accept utterances as correct. And for those with >70% accuracy the system tended to falsely reject correct utterances (see Figure 2).

The results from the 60-word grammar are misleading because it rejected nearly all utterances whether correct or incorrect. When this bias is combined with the fact that the samples consist of disordered speech, the system was "right" most of the time (See Figure 3).

We explored whether severity of aphasia could be used as a predictor of accuracy. As Table 2 shows, accuracy was roughly equivalent across low (> 60), medium (50-59) and high (<50) severity groups.

We also analyzed whether apraxia of speech and dysarthria impacted system accuracy. The data suggest that the presence or absence of motor speech disorders did not influence SR performance (See Table 3).

Conclusions:

- The speech recognition implementations tested here did not achieve the same level of accuracy as trained clinicians.
- The performance of the different grammars depended on the level of impairment.
- Accuracy was negatively impacted when increasing the grammar size from 15 to 60 words.

Discussion

- In this poster, we compared the automated judgments of response accuracy for several speech recognition systems with that of experienced clinicians. We conclude that the systems do not approach the accuracy of the clinicians but the performance of the systems is above chance and with further refinement could be useful to integrate into a CBATP.

- The quality of the *AphasiaBank* sound samples vary and there is considerable ambient and/or electronic noise. We believe that recognition performance could improve to more acceptable levels if the utterances were live (rather than recorded) and if the speech was input directly into the system.

Next Steps

- Complete these analyses with individuals with other types of aphasia.
- Collect and analyze samples using live (rather than recorded) samples.
- Analyze the impact of error types (e.g. literal paraphasias) on SR accuracy.
- Pursue other speech recognition implementations to determine if system accuracy can be improved.

References

AT&T Speech Mashup: https://service.research.att.com/smm

Estes, C. and Bloom, R. L. (2011) Using voice recognition software to treat dysgraphia in a patient with conduction aphasia, Aphasiology, 25: 3, 366-385.

Dahl, D. A., Linebarger, M. C. & Berndt, R. S. (2008). Improving automatic speech recognition of aphasic speech through the use of a processing prosthesis. Technology and Disability 20 (2008) 283-294.

Kaplan E, Goodglass H, Weintraub S. (2001). Boston Naming Test. Philadelphia (PA): Lippincott Williams & Wilkins.

Linebarger, M. C., Schwartz, M. F. & Kohn, S. E. (2001). Computer-based training of language production: An exploratory study. Neuropsychological Rehabilitation, 11 (1), 57-96.

MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). *AphasiaBank*: Methods for

studying discourse. *Aphasiology, 25(11). 1286-1307*.
Wade, J., Petheram, B. & Cain, R. (2001). Voice recognition and aphasia: can computers understand aphasic speech? Disability and Rehabilitation, 23 (14), 604-613.
Venkatagiri, H. S. (2002). Speech Recognition Technology Applications in Communication Disorders. American Journal of Speech-Language Pathology, 11, 323-332.

Table 1: Percent agreement for judgments of speech production correctness.

			Speech Recognition Grammars		
	Clinician A	Clinician B	Any speech/sound + 15 words	15 words	Any speech/sound + 60 words
Clinician A		84%	74%	72%	66%
Clinician B	84%		71%	71%	74%

Table 2: Accuracy of speech recognition judgments as a function of WAB score and Clinician.

WAB Score	Clinician	Any speech/sound + 15 words	15 words	Any speech/sound + 60 words
<50	A	.75	.76	.68
	B	.67	.69	.86
50-59.9	A	.69	.68	.65
	B	.69	.69	.69
>60	A	.75	.70	.60
	B	.76	.72	.67

Table 3: Accuracy of speech recognition judgments with and without co-occurring apraxia of speech.

Apraxia of Speech	Clinician	Any speech/sound + 15 words	15 words	Any speech/sound + 60 words
Y	A	.77	.75	.70
	B	.74	.72	.78
N	A	.50	.45	.32
	B	.58	.59	.47
		.65	**.63**	**.57**

Figure 1: Network diagram showing connectivity to the AT&T Speech Mashup-based speech recognition system.

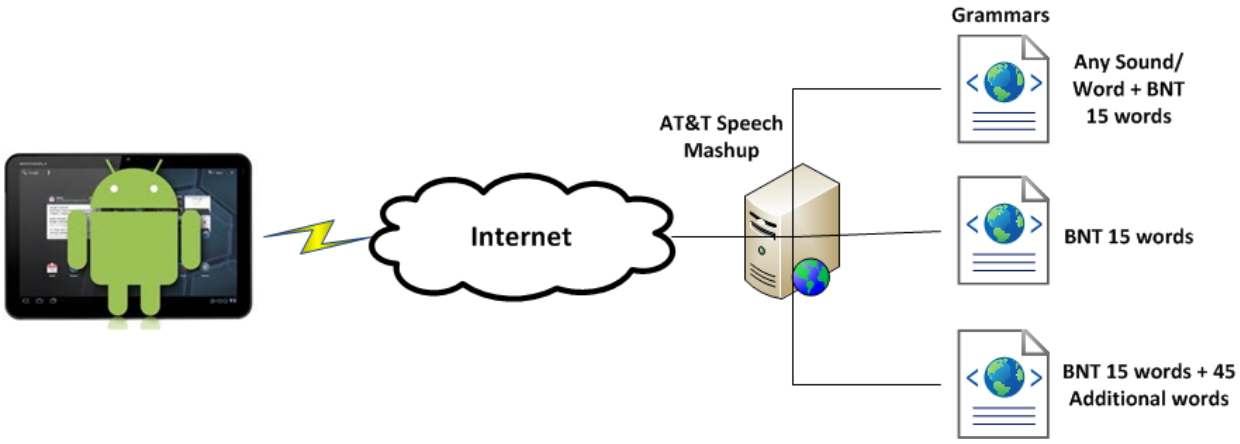

Figure 2: Percent Accuracy Judgment for each AphasiaBank participant across clinicians and 15-word grammar

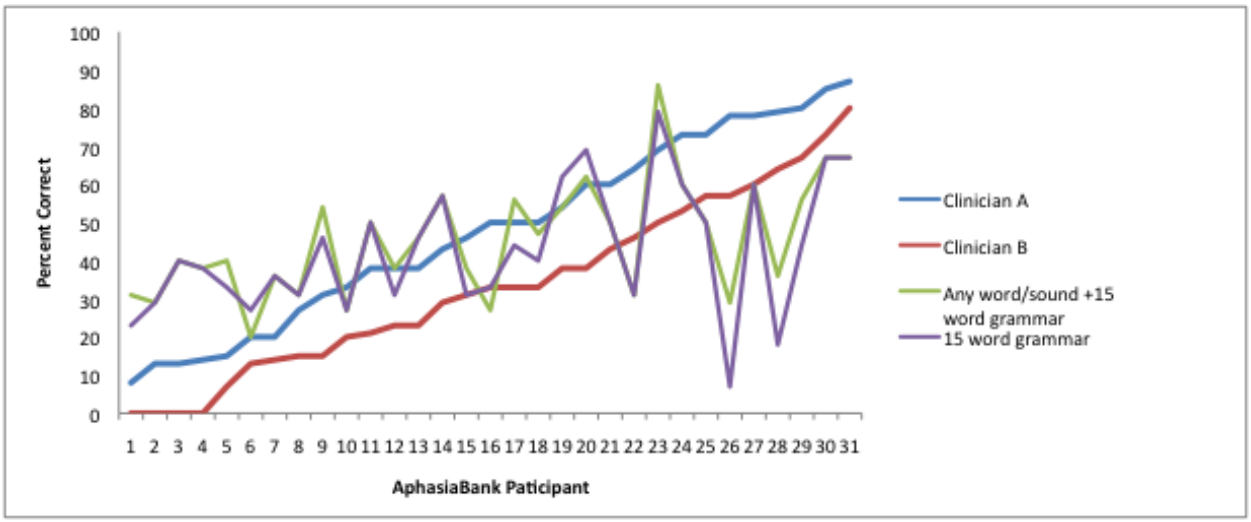

Figure 3: Percent Accuracy Judgment for each AphasiaBank participant across clinicians and the 60-word grammar.