International Journal of Aviation, Aeronautics, and Aerospace

Volume 9 | Issue 3 Article 8

2022

Accuracy of Commercially-Available Speech Recognition Systems in Identifying PIREP Terminology

Deborah Carstens, Ph.D.
Florida Institute of Technology, carstens@fit.edu
Michael S. Harwin, J.D., M.S.
Florida Institute of Technology, MikeHarwin@icloud.com
Tianhua Li, Ph.D.
Florida Institute of Technology, tli2017@my.fit.edu
Michael Splitt, M.S.
Florida Institute of Technology, msplitt@fit.edu
Ridwan Olabanji, M.S.
Florida Institute of Technology, rolabanji2015@my.fit.edu

To help avoid aviation accidents and injuries related to weather (e.g., poor visibility), the NTSB (2017) made several recommendations to enhance the current PIREP submission system. These recommendations included: (a) to simplify procedures to reduce the amount of time flight service station specialists take to record PIREP information from pilots, (b) to provide air traffic controllers with automated data collection tools, and (c) to electronically submit PIREPS directly from pilots. To facilitate electronically submitting PIREPs, one approach is to use an SRS to transcribe, code, and automatically submit PIREPs into the National Airspace System (NAS). To encourage pilots to use an SRS, the SRS must transcribe the PIREP accurately and consistently to avoid having to make corrections to the transcription by hand. The purpose of the research was two-fold: (1) to analyze the performance of COTS SRSs to identify and transcribe PIREPs, and (2) to determine if the transcribed PIREPs are accurate enough to allow FSS to enter them into the PIREP system and increase PIREP volume.

Research Questions and Hypotheses

The research questions (R.Q.s) were:

1. What is the difference in performance across the five levels of SRSs (i.e., Braina, Dragon Home, Google, LilySpeech, and Transcribe)
2. What is the difference in performance rates between two levels of gender (i.e., male and female)
3. What is the interaction between SRS and gender
4. Are the transcribed PREIPs accurate enough to allow FSS to enter them into the PIREP system?

The research hypotheses were:

1. There will be a difference in the transcription performance for the short, average, and long PIREPs
2. There will not be a difference in transcription performance between gender for the short, average, and long PIREPs
3. There will not be an interaction effect in performance between SRS and gender for the short, average, and long PIREPs.

Literature Review

Measures of performance of SRSs are a function of error rate. Measuring the performance of SRSs is a function of error rate (Errattahi et al., 2018). For the current research, the term SRS will be used when not specifically addressing Errattahi et al. automatic speech recognition (ASR) specific literature. By calculating an error rate, researchers can compare the performance between different software platforms. To calculate an error rate, the researchers must identify three types of errors in the context script (i.e., PIREP script, etc.): substitutions, deletions, and insertions. A substitution occurs when the SRS

transcribes the spoken word into a different word. A deletion occurs when the SRS platform does not transcribe a word. A deletion is often called a miss. An insertion occurs when the SRS transcribes more words than were outlined in the referenced script. A popular way to evaluate errors is using the WER, which is determined using the following formula:

$$WER = \frac{S + D + I}{N_1} = \frac{S + D + I}{H + S + D}$$

where S = total substitutions, D = total deletions, I = number of insertions, N_1 = total input words, and H= total hits (matched words) (Errattahi et al., 2018).

Although WER is the most used method to measure ASR software platforms, this method has several shortcomings. First, WER is not an accurate percentage because there is no upper limit. The WER could exceed 100% in noisy conditions because, as the formula shows, it gives more weight to insertions than to deletions (Errattahi et al., 2018). Therefore, the method does not identify how good a system is, but only that one ASR software platform is better than another ASR software platform. Because the study's purpose is to compare SRSs to one another, a WER was calculated for each of the five SRSs.

A second method that alleviates the problems using WER is calculating the Relative Information Lost (RIL) (Errattahi et al., 2018). However, a third method, WIL, can be used in place of RIL because WIL approximates the RIL information. WIL is based on hits (matched words), substitutions, deletions, and insertion counts. The formula is as follows and was also used to calculate the error rate in this study.

$$WIL = 1 - \frac{H^2}{(H + S + D)(H + S + I)}$$

where S = total substitutions, D = total deletions, I = number of insertions, and H = total hits (matched words).

Transcription error rates can be affected by a person's voice acoustics (Mendoza et al., 1996). Female voices have a higher frequency pitch than male voices. Because male and female voices have acoustical differences, we compared the WER and WIL between male and female voices for any statistical differences in transcription accuracy.

Transcription accuracy can also be affected by a person's accent and dialect. Accents are not synonymous with dialects. An accent refers to how words are pronounced, while a dialect relates to grammar, syntax, and vocabulary. People who speak the exact text but pronounce the textual words differently are considered to have an accent. For this reason, a participant's accent was documented. There are 24 different dialects throughout the US (Delaney, 2017). Labov et al. (2006) also wrote a reference manual, The Atlas of North American English: Phonetics, Phonology, and Sound Change, that provides similar dialect information. Although there are 24 different dialects throughout the US, the dialect was not a factor in this

study because the participants were reading from a script. Hence, each participant used the same grammar, syntax, and vocabulary.

Methodology and Design

An experimental method was employed with a 2 x 5 mixed factorial design. The within-subject factor was SRS with five levels (Braina, Dragon Home, Google, LilySpeech, and Transcribe). The between-subject factor was gender, with two levels (male and female). The statistical analysis used a repeated-measures marginal model with an unstructured covariance structure, and an α-level of .05 was used to determine if there were main effects of SRS, main effects of gender, and interaction effects between SRS and gender.

Population and Sample

The target population consisted of male and female pilots 18 years of age or older, who were American English native-speakers, who answered a survey question about mostly having lived in the United States. The accessible population was male and female students, staff, and faculty 18 years of age or older at a university in Florida.

The sampling strategy was non-probability convenience sampling. Students, faculty, and staff on campus were recruited to read three PIREP scripts consisting of short, average, and long lengths. The PIREPs were recorded on an iPad. The sample size was 86. One participant's data were excluded because the participant did not document their gender. One participant listed other for their gender; due to there being only one data point for this gender category, this gender level was removed because an inferential statistical analysis could not be performed for this category. Therefore, the final sample size was 84.

The study began by reviewing characteristics of 12 commercial off-the-shelf (COTS) SRSs. The characteristics included cost, performance data (if available on the SRS website), the interface (i.e., user-friendly), platform (i.e., phone application, PC, etc.), a summary of pros, a summary of cons, and SRS features. User reviews on multiple e-Commerce platforms were reviewed. Based on the features such as (a) cost, (b) performance data published on SRS websites, (c) interface (i.e., user-friendly), (d) platform (i.e., phone application, PC, etc.), (e) summary of pros and cons, and (f) reviews and SR features, five SRSs were included in the study: (a) Brain Artificial (Braina), (b) Dragon Home/Dragon® Home v15 speech recognition, (c) Google Dictation (Voice Notepad - Speech to Text with Google), (d) LilySpeech, and (e) Transcribe by Wreally.

Participants were provided with a survey consisting of six demographical questions, including (a) gender, (b) age category, (c) birth city and state, (d) city and state mostly lived while growing up, (e) accent, and (f) pilot certificates held. Participants read three different PIREP lengths, short (40 words), average (53 words), and long (74 words), into an iPad Voice Memos application (App). These

PIREP lengths were chosen based on a review of sample PIREPs from the literature and PIREPs submitted on the Aviation Weather Website. A room was chosen in a location that minimized ambient noise. A sound meter was used at the beginning of each participant session to document the A-weighted decibel (dBA) to confirm there was no excessive ambient noise. Each PIREP was played into each of the five SRSs. To calculate the WER and WIL error rates, each transcribed PIREP was compared to the reference script word-by-word to determine the errors. All the words were analyzed using lower case, and numbers remained in Arabic (e.g., one changed to 1). Symbols were changed to text (e.g., -4 was changed to minus 4). The comparison was conducted automatically using a code based on Python developed by the researchers specifically for this and then checked manually by an experimenter. The numbers of hits, substitutions, insertions, and deletions were determined and used to calculate WER and WIL.

Results

Descriptive Analysis

Demographic data included 49 males and 35 females. One participant selected other, and one participant's response was missing. There were 72 participants aged 18-29, one aged 30-39, three aged 40-49, three aged 50-59, and five aged 60 or above. The participants were born in 28 different states, and two were born in U.S. territories. Participants had lived in 24 of the 50 states. The results indicate that most participants were born in Florida (i.e., 20) and mostly lived in Florida (i.e., 38). Of the 84 participants, there were 66 participants that self-reported that they do not have an accent. The researchers did not hear any participants with pronounced accents. The participants' pilot certification level is summarized in Table 1. The mean, standard deviation, and range of the WIL and WER of each SRS for the short, average, and long PIREP transcriptions are displayed in Table 2.

Table 1
Aviation Certificates Held by the Participants

Certificates	Frequency
Student Certificate	13
Private Certificate	3
Sport Certificate	0
Recreational Certificate	0
Commercial Certificate	5
Airline Transport Pilot Certificate	1
Certified Flight Instructor Certificate	4
Other Certificate	6
No Certificate	65

Note. $N = 84$. Participants could choose more than one certificate category.

Table 2
WER and WIL Descriptive Statistics, Including the Mean, Standard Deviation, and Range

	Categories		Min	Max	Mean	SD
WER	**Short**	**Braina**	2.50%	52.50%	20.92%	12.31%
		Dragon	7.50%	57.50%	20.18%	11.25%
		Google	7.50%	52.50%	20.60%	9.25%
		Transcribe	5.00%	42.50%	18.81%	9.31%
		LilySpeech	5.00%	47.50%	20.12%	8.06%
	Average	**Braina**	2.50%	30.00%	8.57%	6.09%
		Dragon	5.66%	83.02%	20.89%	13.64%
		Google	7.55%	84.91%	23.02%	13.21%
		Transcribe	3.77%	47.17%	17.57%	9.83%
		LilySpeech	1.89%	35.85%	14.96%	7.09%
	Long	**Braina**	1.89%	39.62%	15.57%	7.60%
		Dragon	1.89%	22.64%	7.48%	4.40%
		Google	8.11%	85.14%	31.74%	17.47%
		Transcribe	8.11%	64.86%	25.26%	12.24%
		LilySpeech	4.05%	60.81%	26.40%	10.54%
WIL	**Short**	**Braina**	6.76%	67.57%	23.21%	9.43%
		Dragon	8.11%	47.30%	24.87%	9.09%
		Google	1.35%	21.62%	7.66%	4.06%
		Transcribe	4.94%	60.00%	29.91%	15.15%
		LilySpeech	14.44%	77.42%	29.87%	12.68%
	Average	**Braina**	14.44%	62.19%	30.86%	11.45%
		Dragon	9.75%	61.10%	28.76%	12.40%
		Google	9.75%	65.43%	30.14%	10.68%
		Transcribe	4.94%	44.00%	14.14%	9.08%
		LilySpeech	9.29%	89.08%	30.52%	14.73%
	Long	**Braina**	14.52%	86.58%	33.26%	13.84%
		Dragon	5.62%	63.92%	25.23%	12.79%
		Google	3.74%	51.85%	23.39%	9.99%
		Transcribe	3.74%	51.70%	23.27%	9.99%
		LilySpeech	3.74%	36.57%	12.02%	6.31%

Note. The table excludes missing or removed data. $N = 84$.

Inferential Statistics

Preliminary Analysis

A preliminary analysis was conducted because the Dragon Home software required the user to read a paragraph to check the microphone. It was discovered that the WER and WIL for females were much higher than for males when using Dragon, which suggested a potential extraneous variable (i.e., initiation). To determine whether the gender of voice used for initiation can affect the WIL and WER, the male and female error rates were computed using a male-initiated Dragon Home and then computed using a female-initiated Dragon Home. A two-by-two mixed ANOVA was used to analyze the error rate of gender interacted by the initiation gender. The within-subjects factor was the initiation and had two levels

(male initiation and female initiation). The between-subjects factor was gender and had two levels (male voices and female voices). For all three PIREP lengths (short, average, and long), males had significantly lower WER and WIL than females when Dragon Home was initiated with a male voice. When Dragon Home was initiated with a female voice, females had significantly lower WER and WIL than males. It demonstrated that the initiation does not simply check the microphone but also affects transcribing, which is not consistent with the manual.

Dragon Home was not initiated for each participant before they read their PIREPs. This would have required 84 different calibrations, thus making a participant-specific calibration an unreasonable use of the participant's time. Therefore, we combined the male transcriptions that were recorded using a male calibration with the female transcriptions that were recorded using a female calibration and designated this sample as Dragon Home.

Primary Analysis

The primary statistical analysis was a repeated-measures 2 x 5 marginal model because it calculates residuals more accurately than multivariate or univariate methods. The between-subjects factor was gender and had two levels (i.e., male and female). The within-subjects factor was SRS type and had five levels (i.e., Braina, Dragon Home, Google, LilySpeech, and Transcribe). The main effect between gender (male and female) and the main effect between the five SRSs for the short, average, and long PIREPs were analyzed. Any interactions between gender and the five SRSs were identified for the short, average, and long PIREPs.

Assumptions. Because the Shapiro-Wilk method of determining normality is known to be unreliable, we analyzed the Q-Q plots. The Q-Q plots showed the short, average, and long PIREPs were not normally distributed. However, because the sample size was relatively large, no transformation was conducted. For the homogeneity of variances, a scatterplot with the residuals and predicted values was evaluated. A constant variance was identified for the five off-the-shelf SRSs. Therefore, the assumption was satisfied.

Short. When analyzing the WER for the short PIREPs, there was no significant main effect between the five SRS, $F(4, 82) = 1.426$, $p = .233$. There was no significant main effect for gender, $F(1, 82) = .255$, $p = .615$. There were no significant ordinal and disordinal interactions between the SRS and gender, $F(4, 82) = .302$, $p = .876$.

When analyzing the WIL for the short PIREPs, there was no significant main effect between the five SRS, $F(4, 82) = 1.307$, $p = .274$. There was no main effect between gender, $F(1, 82) = .149$, $p = .700$. There no were significant ordinal and disordinal interactions between the SRS and gender, $F(4, 82) = .593$, $p = .669$.

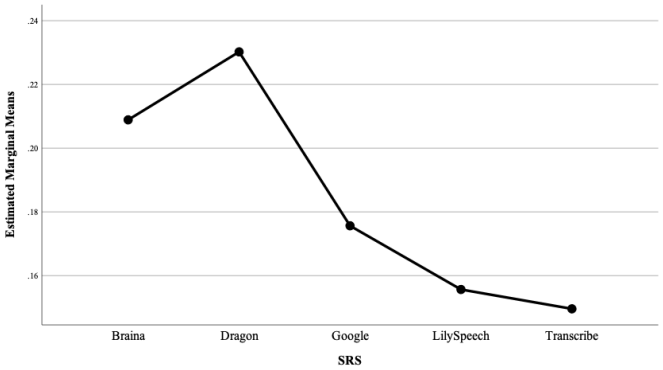
Average. When analyzing the WER for the average length PIREPs, there was a significant main effect between the five SRS, $F(4, 82) = 12.826$, $p < .001$.

Differences between the SRSs were identified with a pairwise post-hoc Sidak analysis (see Table 3). A graphical representation of the SRS means is displayed in Figure 1. There was no main effect between gender, $F(1, 82) = .119$, $p = .731$. There were no significant ordinal and disordinal interactions between the SRS and gender, $F(4, 82) = .062$, $p = .993$.

Table 3
Pairwise Comparisons of SRS Average Length PIREPs' WER Error Rate

Comparisons	Difference	Significance
Dragon - Braina	2.2%	> .999
Dragon - Google	5.5%	.003
Dragon - Transcribe	8.1%	< .001
Dragon - LilySpeech	7.5%	< .001
Braina - Google	3.3%	.321
Braina - Transcribe	5.9%	< .001
Braina - LilySpeech	5.3%	.007
Google - Transcribe	2.6%	.005
Google - LilySpeech	2.0%	.293
Transcribe - LilySpeech	-0.6%	> .999

Figure 1
Means Plot of WER SRS Average Length PIREPs

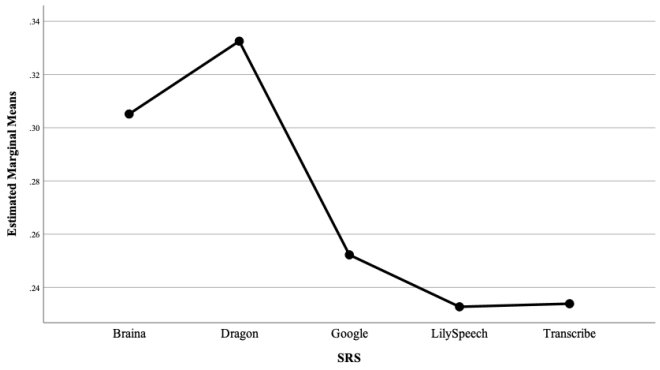


When analyzing the WIL for the average length PIREPs, there was a significant main effect between the five SRS, $F(4, 82) = 13.903$, $p < .001$. Differences between the SRSs were identified with a pairwise post-hoc Sidak analysis (see Table 4). A graphical representation of the SRS means is displayed in Figure 2. There was no main effect between gender, $F(1, 82) = .477$, $p = .492$. There were no significant ordinal and disordinal interactions between the SRS and gender, $F(4, 82) = .096$, $p = .984$.

Table 4
Pairwise Comparisons of SRS Average Length PIREPs' WIL Error Rate

Comparisons	Difference	Significance
Dragon - Braina	2.8%	> .999
Dragon - Google	8.1%	< .001
Dragon - Transcribe	9.9%	< .001
Dragon - LilySpeech	10.1%	< .001
Braina - Google	5.3%	.024
Braina - Transcribe	7.1%	< .001
Braina - LilySpeech	7.3%	< .001
Google - Transcribe	1.8%	.554
Google - LilySpeech	2.0%	.749
Transcribe - LilySpeech	0.2%	> .999

Figure 2
Means plot of WIL SRS average length PIREPs



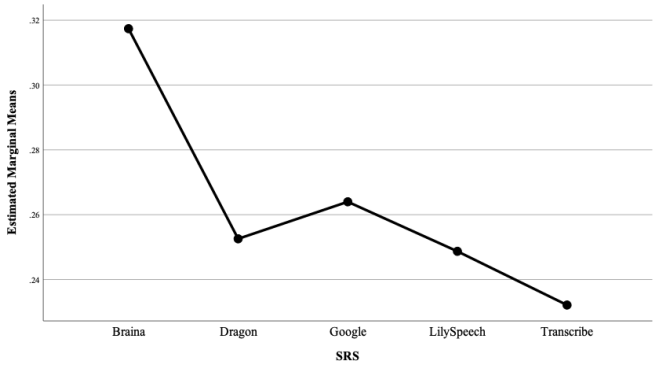
Long. When analyzing the WER for the long-length PIREPs, there was a significant main effect between the five SRS, $F(4,82) = 7.624$, $p < .001$. Differences between the SRSs were identified with a pairwise post-hoc Sidak analysis (see Table 5). A graphical representation of the SRS means is displayed in Figure 3. There was no main effect between gender, $F(1, 82) = .206$, $p = .651$. There were no significant ordinal and disordinal interactions between the SRS and gender, $F(4, 82) = 2.296$, $p = .066$.

Table 5
Pairwise Comparisons of WER SRS Long-Length PIREPs

Comparisons	Difference	Significance
Dragon - Braina	-5.4%	.242
Dragon - Google	-0.8%	> .999
Dragon - Transcribe	2.3%	.652
Dragon - LilySpeech	0.7%	> .999
Braina - Google	4.6%	.167
Braina - Transcribe	7.7%	.001
Braina - LilySpeech	6.0%	.018
Google - Transcribe	3.1%	< .001
Google - LilySpeech	1.4%	.626
Transcribe - LilySpeech	-1.7%	.258

Figure 3

Means plot of WER SRS for Long-Length PIREPs



When analyzing the WIL for the long-length PIREPs, there was a significant main effect between the five SRS, $F(4, 82) = 5.342$, $p < .001$. Differences between the SRSs were identified with a pairwise post-hoc Sidak analysis (see Table 6). A graphical representation of the SRS means is displayed in Figure 4. There was no main effect between gender, $F(1, 82) = .781$, $p = .379$. There were significant ordinal and disordinal interactions between the SRS and gender, $F(4, 82) = 2.492$, $p = .049$ (see Table 7). A graphical representation of the interactions is displayed in Figure 7.

Table 6
Pairwise Comparisons of WIL SRS Long Length PIREP Error Rate

Comparisons	Difference	Significance
Dragon - Braina	-6.0%	.116
Dragon - Google	-0.9%	> .999
Dragon - Transcribe	1.6%	.968
Dragon - LilySpeech	0.3%	1.00
Braina - Google	5.1%	.061
Braina - Transcribe	7.6%	< .001
Braina - LilySpeech	6.3%	.007
Google - Transcribe	2.5%	.046
Google - LilySpeech	1.1%	.887
Transcribe - LilySpeech	-1.3%	.577

Figure 5
Means Plot of WIL SRS for Long Length PIREPs

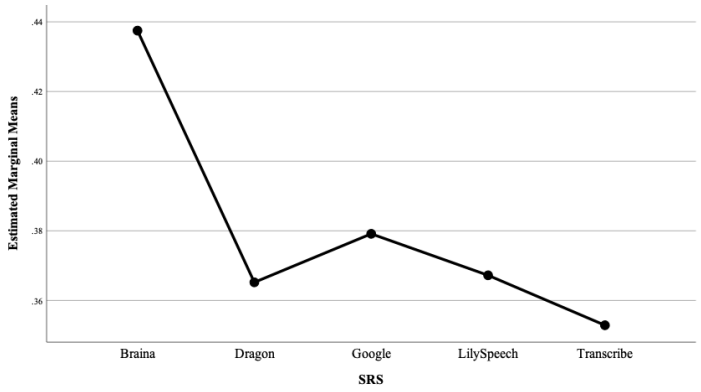
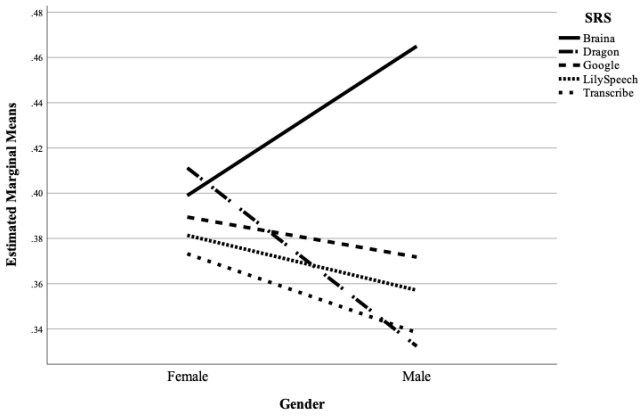


Table 7
Interaction Effects of WIL of Long-Length PIREPs

Interaction	Braina	Dragon	Google	Transcribe	LilySpeech
Braina					
Dragon	Disordinal*				
Google	Ordinal*	Disordinal			
Transcribe	Ordinal*	Disordinal*	Ordinal*		
LilySpeech	Ordinal*	Disordinal	Ordinal	Ordinal	

Note. The disordinal interaction between Dragon and Google is marginally significant with p =.051. * $p < .05$.

Figure 6
Interaction plot of WIL Long Length PIREPs



Discussion and Conclusion

Males had lower WER and WIL than females when the Dragon Home system was calibrated with a male voice. Furthermore, females had lower WER and WIL than males when the system was calibrated with a female voice. However, we conclude that Dragon Home needed to be calibrated by individuals to obtain the best performance, but this study design was not practical. Therefore, we combined the male-initiated transcriptions with the female-initiated transcriptions and designated this sample as Dragon Home.

Short PIREPs

Because there were no significant differences in the main effects of the SRS, gender, or the interaction effects for the WER or WIL, we conclude the following: (1) neither of the SRS levels were better in transcription accuracy and would perform with the same accuracy, (2) because gender did not affect the accuracy of the SRS performance, it does not matter whether a male or female voice was used to read the short PIREPs, and (3) male and female operate the same for all SRSs (i.e., Braina, Dragon Home, Google, LilySpeech, and Transcribe). The research hypotheses that there would be differences in the SRS main effects, gender, and interactions were not supported, but the research hypothesis that there would be no main effect in gender or interactions was supported.

There was no statistical difference between the SRSs. Although the WER and WIL cannot be used to accurately indicate the average number of corrections a pilot would need to make to a transcribed PIREP, the research suggests from our review of the PIREP transcription scripts for each of the SRSs, that these SRSs would facilitate pilots submitting short-length PIREPs because enough information is readable for FSS to process a PIREP. However, performance could be different based on a differently worded short-length PIREP.

Average PIREPs

There was a significant main effect between SRS of the WER and WIL. The research hypothesis that there would be a main effect of SRS was supported. We conclude that Google, LilySpeech, and Transcribe had significantly better performance than Dragon and Braina. For the WIL, there was no significant difference among Google, LilySpeech, and Transcribe. For the WER, there was no significant difference between LilySpeech and Transcribe, but Google was significantly different from Transcribe but not significantly different than LilySpeech. However, we found the difference was not practically significant as it was less than 3%, which accounted for one or two errors. Thus, Google, Transcribe, and LilySpeech could be considered superior to Braina and Dragon's performance. As previously mentioned with the short-length PIREP, that these SRSs would facilitate pilots submitting short-length PIREPs because enough information is readable for FSS to process a PIREP. However, performance could be different based on a differently worded short-length PIREP.

Because there were no significant differences in the main effects of the WER or WIL for gender, we conclude that gender did not affect the accuracy of the SRS. The research hypothesis that there would not be a main effect of gender was supported. Thus, it does not matter whether a male or female voice was used to record the average PIREPs. Nor was there any significant interaction between the SRS factor and the gender factor. The research hypothesis that there would not be an interaction was supported. The gender levels (male and female) operated the same on all levels of the SRS factor (Braina, Dragon Home, Google Dictation, LilySpeech, and Transcribe).

Long PIREPs

There was a significant main effect between SRS of the WER and WIL. Thus, the research hypothesis was supported. For the long PIREP, Dragon's error rate was significantly reduced compared to short and average PIREPs. Because of the reduced error, there was no significant difference between Dragon's WER and WIL than any of the other SRSs. However, Braina was significantly different from Transcribe and Lily. Although Google was significantly different from Transcribe, the significance was not practically significant. We conclude that Dragon, Google, LilySpeech, and Transcribe had the best performance, and these SRSs would facilitate pilots submitting long-length PIREPs because enough information is readable for FSS to enter weather information into the PIREP system.

There were no significant differences in the WER and WIL for the main effects of gender for the long PIREPs. Thus, the research hypothesis that gender would not have a main effect was supported. We conclude that gender did not affect the accuracy of the SRS, and it did not matter whether a male or female voice was used to record the long-length PIREPs.

There were no significant differences in the interaction effects of the SRS and WER, but there was significance for the WIL for the long-length PIREPs, although the *p*-value for the WIL was .049 while the *p*-value for the WER was .066. *A* Sidak post-hoc analysis identified pairs with significant differences in the WIL (see Table 5 and Figure 3). The only disordinal interactions were between Dragon and Transcribe and between Dragon and Braina (see Figure 3). Dragon and Google were marginally significant, with a *p*-value of .051. For the WIL, Dragon was able to transcribe male voices more accurately than Transcribe, and Transcribe could transcribe female voices more accurately than Dragon. Dragon was able to transcribe male voices more accurately than Braina, and Braina could transcribe female voices more accurately than Dragon. Despite the interaction effect, we conclude, as we did above, that the PIREPs are readable enough for FSS to process them. However, performance could be different based on a differently worded long-length PIREP. While the research hypothesis was supported for the WER because there was no interaction, it was not supported for the WIL because of the significant interactions.

Summary

Except for the significant disordinal interaction between Dragon and Transcribe and the significant disordinal interaction between Dragon and Braina, it appeared that Google, LilySpeech, and Transcribe had the best performance transcribing the PIREPs regardless of gender. Dragon Home would still have to be calibrated with the user's voice for the best performance and only had performance similar to Google, LilySpeech, and Transcribe for the long-length PIREPs.

Whether the SRS was a paid-for service or a free service, it did not affect the WER or WIL. Although Transcribe is a paid-for service, and LilySpeech is free, there was no evidence one had better performance than the other for all three length PIREPs. Similarly, Dragon, a paid-for service, had higher WER and WIL in transcribing the average length PIREP than LilySpeech. Dragon also had higher WER in transcribing the long-length PIREP than LilySpeech.

Because pilots are trained to use standard language to submit PIREPs, their grammar, syntax, and vocabulary would be the same. Thus, one's dialect would not affect the results. Although some participants reported they had accents, we did not hear participants using a dominant accent. We conclude the results could be generalized to the English-speaking pilot population in the United States who speak without an appreciable accent. However, none of the SRSs had the transcription accuracy to allow pilots to use these systems to facilitate submitting PIREPs into the NAS. Pilots having to make too many corrections to their transcription would discourage their use to submit PIREPs.

Moreover, the recommendation about the length was provided for reference. It did not demonstrate that the SRSs performed differently on certain lengths

because the phraseology included in each PIREP for a certain length was various. It was possible that the words used in short and long PIREPs were harder to be transcribed accurately by SRSs.

Future Research

Because all five COTS SRSs were not able to transcribe the PIREPs without some information loss, constructing an SRSs for aviation use that contains an exclusive aviation vocabulary should be considered. An SRS that is programmed with its own aviation library of terms would have a substantially reduced vocabulary, thereby allowing the SRSs to match words more accurately than COTS SRSs. For example, the word *haze* was consistently transcribed as *hayes*. Because *hayes* would not be included in an aviation-specific library of terms, the word *haze* would probably be transcribed accurately.

Because this study was limited to participants without accents, although participants reported they had accents, a larger study with participants who spoke with accents or different acoustical voices could reveal differences in the WER and WIL. Additional studies would allow the researchers to understand the challenges that SRSs could pose to pilots with different accents. Understanding these challenges would then provide an opportunity for current and future SRS technology companies to identify solutions to these challenges so that transcription error rates would be minimized should pilots use these technologies to generate PIREPs. SRS technologies that minimize transcription errors could contribute to more PIREPs being submitted by pilots, which aligns with the NTSB's (2017) recommendations for increasing the effectiveness and distribution of PIREPs.

References

Delaney, R. (April 15, 2013). *Dialect map of American English*. http://robertspage.com/dialects.html

Errattahi, R., Hannani, A, E., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *ScienceDirect*, *128*, 32-37.

Labov, W., Ash, S., Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change.* Mouton de Gruyter

Mendoza, E., Nieves, V., Munoz, J., & Trujillo, H. (1996). Differences in vice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of Voice*, *10*(1), 59-66.

National Transportation Safety Board. (2017). NTSB/SIR-17/02: *Improving pilot weather report submission and dissemination to benefit safety in the National Airspace System*. https://www.ntsb.gov/safety/safety-studies/Documents/SIR1702.pdf