AUTOMATED SPEECH RECOGNITION TECHNOLOGY TO SUPPORT IN FLIGHT WEATHER-RELATED COMMUNICATION FOR GA PILOTS

Gaojian Huang
*N*HanCE Research Lab, Purdue University
West Lafayette, IN
Brandon J. Pitts, Ph.D.
*N*HanCE Research Lab, Purdue University
West Lafayette, IN

Weather information latency during flight in general aviation (GA) has resulted in numerous incidents. Hands-free automated speech recognition (ASR) systems have the potential to help overcome this challenge and facilitate rapid weather-related information exchange. However, it is unclear to what extent ASR systems can support pilot communication in such noisy environments. The goals of this study were to (1) evaluate the performance of 7 commercially-available ASR systems to recognize weather phrases during GA operations and (2) determine whether speech-to-noise (S/N) ratio, flight phase, and accent type modulate system performance. Overall, the highest accuracy percentage achieved by any system was 72%, when the S/N ratio was at least 3/2. This research can help to inform the selection and development of next-generation technologies to be used in safety-critical, information-rich domains.

For more than two decades, adverse weather conditions has been cited as one of the most frequent causes of fatal accidents among general aviation (GA) pilots (e.g., Duke & George, 2016; Federal Aviation Administration, 2010) To help improve safety, GA pilots need to be aware of the weather conditions along their flight path. Traditionally, pilots are provided with weather briefings prior to flight and may receive updated weather information from Flight Service while flying (Ahlstrom, Ohneiser, & Caddigan, 2016). However, to date, weather information latency during flight, i.e., the time delay between flight environment weather conditions and the presentation of this information on cockpit displays, still represents a major problem in GA and limits the decision making abilities of pilots.

The emergence of NextGen technologies may offer pilots tools to improve their situational awareness and result in better real-time strategizing. For example, mobile devices and tablets are increasingly able to support aviation software that can inform pilots as new weather information becomes available. In addition, not all NextGen advancements require manual interactions. Automated speech recognition (ASR) technology is one particular development that can assist with activating commands and quickly obtaining critical information. These systems translate natural spoken language/words into text (Këpuska, 2017), which can then be used to execute specified functions. The benefit of hands-free interactions is especially important in the context of extreme weather conditions during flight, when pilots' cognitive and manual workload are already high. Recently, commercial ASR systems, such as Google Cloud Speech API and Microsoft Bing Speech API, have been developed and used in applications, such as portable devices, smart homes, and autonomous vehicles (Këpuska, 2017; Kimura, Nose, Hirooka, Chiba, & Ito, 2019). Significant progress in the development of artificial intelligence and machine and deep learning technologies has resulted in these systems achieving detection rates as high as 90% (Yu & Deng, 2016).

In aviation, previous work has investigated the use of speech recognition systems in flight (e.g., Arthur III, Shelton, Prinzel III, & Bailey, 2016) in both real-world and laboratory environments, but not with respect to specific weather-related communication. One other open question regarding ASR systems in flight is the extent to which they can perform in noisy environments (Hansen, 1996). The goal of this study was, therefore, to determine the effectiveness of commercially-available speech recognition systems to support weather-related communication in GA.

Method

Participants

Thirty participants from Purdue University and a multidisciplinary research project team volunteered to take part in this study. All participants were required to be fluent in English. The 30 participants were divided into 6 accent/dialect groups based on their geographical origins (East Asia, India, Latin America, Northern and Southern U.S, and UK/Australia/South Africa). This study was approved by the Purdue University Institutional Review Board (IRB Protocol ID: 1804020515).

Apparatus and Test Stimuli

Speech recognition system selection. During an initial market analysis phase, 50 potential commercially-available systems were identified based on accessibility (e.g., downloadable), capability (e.g., performance/accuracy), interface design, and cost. The final selection of systems was focused on: speaker-independent, customizable vocabulary database, platform type, and performance in noisy environments. In total, the following 7 systems were chosen for evaluation: Braina Pro; Dragon NaturallySpeaking (with and without speech training component); Google Cloud Speech API; Microsoft Bing Speech API; Houndify; Lily Speech.

Speech & Aircraft nose file generation. A human-subject experiment was conducted to create samples of spoken weather-related phrases. In particular, the 30 participants were recorded reciting 35 separate weather-related phrases commonly used by GA pilots (e.g., ‘show PIREPs’, ‘show convective SIGMET’, etc.). An aviation quality headset (i.e., ASA AirClassics HS-1A) was used to make these recordings in a quiet laboratory environment. At a different time, background aircraft cockpit noise samples were also recorded during the taxi, cruise, and takeoff flight phases of a test flight carried out at The Ohio State University airport (Don Scott Field). The intensity range of these samples was 95-124 dB. The device used to create these recordings was a Sony ICD-PX333 Digital Voice Recorder.

Test stimuli. The recorded speech files and aircraft cockpit noise samples were digitally combined using, Audacity 2.2.2, to create the “test stimuli.” The goal was to evaluate conditions in which: a) the background noise was louder than the speech (S < N), b) the noise and speech volumes were the same (S = N), and c) the speech was louder than the noise (S > N). To this end, the combined speech and noise file was adjusted to different speech-to-noise (S/N) ratios in each of the three categories. Specifically, 9 S/N intensity ratios (1/2, 5/8, 3/4, 7/8, 1/1, 5/4, 3/2, 7/4, 2/1) were initially selected based on psychophysical research involving the differentiation between two concurrent stimuli (Biberger & Ewert, 2015; Bradley, Reich, & Norcross, 1999). Also, a baseline condition with only speech (no background noise) was generated.

Factors selection. All files were processed internally within the 7 speech recognition software packages and recognition accuracy rate was calculated. After preliminary investigation, 8 S/N ratios (1/2, 5/8, 3/4, 1/1, 5/4, 3/2, 2/1, and the baseline condition) and 2 types of flight phases (taxi and cruise) were selected, because no statistically significant differences were found between adjacent S/N ratio and flight phases and those that were excluded.

Experimental Design

Overall, the experiment employed a 2 (flight phase) × 8 (S/N ratio) × 7 (system) × 6 (accent) full factorial design. Flight phases, S/N ratios, and systems were within-subject variables, and accent was a between-subject variable. The 6 accent/dialect groups were determined based on self-reported information provided by participants prior to the experiment. Sixteen auditory files were created for each participant (2 flight phases and 8 S/N ratios). This resulted in a total of 480 files and 3,360 total runs.

Procedures

Each participant first signed a consent form. Next, they familiarized themselves with the 35 phrases (i.e., pronunciation and sequence). Once participants indicated that they were ready to record, the experimenter left the room and the participant started and stopped the recordings as instructed. All phrases were read using participants' normal speaking volume (~60 dB).

Data Analysis

The dependent variable was phrase accuracy rate (PAR), i.e., the percentage of phrases correctly recognized by the software out of the total number of phrases. This measurement was inspired by previous work which used Word Error Rate (WER) as the ASR performance measure (e.g., Vipperla, 2011). A 4-way analysis of variance (ANOVA) was used to identify main and interaction effects. Results were considered significant at $\alpha = 0.05$. Since none of the 7 systems recognized speech when the S/N ratio was less than 1 (i.e., PAR = 0%), a perfect separation assumption was used and only data in cases where S/N ratio ≥ 1 were included in analysis.

Results

There was a significant main effect of system on PAR, $F\ (6, 1994) = 796.067, p < .001,\ \eta_p^2 =\ .705$. In particular, post-hoc analysis revealed that the Dragon NaturallySpeaking (with speech training component) (mean PAR = 72.3%, standard error of mean (SEM) = .018) has a significantly higher PAR compared to all other systems, see Figure 1.

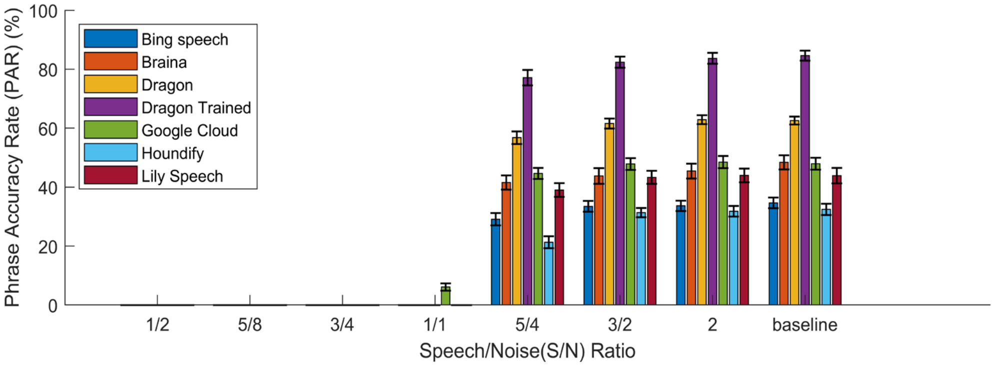

Figure 1. Phrase accuracy rate (PAR) as a function of speech/noise ratio for 7 speech recognition systems during cruise flight phase (error bars represent standard error of mean)

There was also a significant main effect of Speech/Noise (S/N) ratio on PAR, $F\ (4, 1994) = 1392.741, p < .001,\ \eta_p^2 = .736$. In particular, all systems performed better when the S/N ratio was at least 3/2 (mean PAR = 51.2%, SEM = 0.010) compared to when the S/N was 5/4 (mean PAR = 47.6%, SEM = 0.010) or 1/1 (mean PAR = 11.0 %, SEM = 0.008). The baseline condition (mean PAR = 52.3%, SEM = 0.010) and an S/N ratio of 2/1 (mean PAR = 51.8%, SEM = 0.010) did not differ from an S/N of 3/2.

PAR was affected by accent type, $F\ (5, 1994) = 111.568, p < .001,\ \eta_p^2 = .219$, (note here the relatively small effect; Watson, Lenz, Schmit, & Schmit, 2016). Specifically, the Northern American (mean = 49.8%, SEM = 0.014) and Southern American (mean = 49.6%, SEM = 0.013) accents were slightly more recognizable than those from any other region (East Asia mean = 37.6%, SEM = 0.012; Latin America mean = 41.1%, SEM = 0.014; India mean = 38.3%, SEM = 0.013; and UK/Australia/South Africa mean = 41.2%; SEM = 0.014).

Discussion

This study evaluated the extent to which commercially-available speech recognition systems could recognize weather-related terminology in a GA environment. The highest phrase accuracy rate (PAR) achieved by any system was 72% (which included a training component). Also, all systems performed best when the speech-to-noise (S/N) ratio was at least 3/2. Finally, U.S. accents were slightly more recognizable than those from any other world regions.

None of the ASR systems used in this study achieved a PAR of 100%. Typically, default speech recognition vocabulary databases do not include aviation-related phraseology. Dragon NaturallySpeaking (with speech training component), however, achieved the highest accuracy rate. This result is consistent with previous work which found that Dragon NaturallySpeaking was significantly more accurate compared to other common speech systems (Rami, Svitlana, Lyashenko, & Belova, 2017). Specifically, in this study, the performance of Dragon NaturallySpeaking increased from 54% to 72%, without and with training, respectively. This suggests that training systems how to pronounce particular words can significantly increase detection accuracy. It is critical that training be conducted using a well-crafted aviation-specific vocabulary training set and default references for terms likely to be confused. For example, if the system perceives “Sig Mat,” it should default to SIGMET. Relatedly, in this study, we focused on the accuracy of complete phrases (as opposed to words) as an implication for the execution of weather-related commands. However, accuracy rates would have been much greater if calculations were done based on words (as used in Këpuska, 2017).

In terms of S/N ratio, even though many ASR systems are marketed to perform in noisy environments, the best detection rates recorded for all systems evaluated in this study was when the S/N ratio was 3/2 or greater. This indicates that minimal background noise may not interfere with pilot communication to speech systems. However, if an environment produces a considerable amount of noise, then a high S/N ratio may be achieved through the selection of the proper headset equipment (e.g., those with microphones close to the speaker's mouth) or the use of a microphone that recognizes speech using throat vibration signals. Also, noise absorption material may be installed in the cockpit to reduce ambient noise sources.

Accent type was found to have an effect on PAR. Native Northern and Southern U.S. participants' speech was more detectable (i.e., detection accuracy ~ 50%) than participants from East Asia, Latin America, India, or UK/Australia/South Africa. One possible explanation for this finding is that the systems evaluated in this experiment were developed using (American) English speakers. This interpretation infers that in order to increase recognition accuracy, corpuses used to create and train ASR systems should include a wide range of demographic factors, such as accents/dialects, speech rates, and age groups. Finally, it is no surprise that PAR was not affected by flight phase. Although the background noise frequencies between the two conditions may have slightly differed, their overall loudness and rhythm were perceived comparably by the ASR system, especially given that the sounds did not resemble human speech.

In summary, while the outcome of this work will be useful in field research and to the GA community, more research is needed to determine, for example, minimum requirements prior to adoption into practice. Still, this research may help to guide decisions regarding the selection and use of smart devices and applications in complex domains.

Acknowledgements

This work was supported by the Federal Aviation Administration (FAA), as part of the Partnership to Enhance General Aviation Safety, Accessibility and Sustainability (PEGASAS) Project 4: Weather Technology in the Cockpit (WTIC). The authors would like to thank Han Wang, Yuyuan Liu, and Matthew Scott for developing the test environment and stimuli, and for assisting with data analysis. We also thank the statistics professionals from the Purdue University Department of Statistics for their statistical consulting service.

References

Ahlstrom, U., Ohneiser, O., & Caddigan, E. (2016). Portable Weather Applications for General Aviation Pilots. *Human Factors*, *58*(6), 864–885.

Arthur III, J. J., Shelton, K. J., Prinzel III, L. J., & Bailey, R. E. (2016). *Performance Evaluation of Speech Recognition Systems as a Next-Generation Pilot-Vehicle Interface Technology*.

Biberger, T., & Ewert, S. D. (2015). Envelope-power based prediction of auditory masking and speech intelligibility. *The Journal of the Acoustical Society of America*.

Bradley, J. S., Reich, R., & Norcross, S. G. (1999). A just noticeable difference in C50 for speech. *Applied Acoustics*, *58*(2), 99–108.

Capobianco, G., & Lee, M. D. (2001). The Role of Weather in General Aviation Accidents: An Analysis of Causes, Contributing Factors and ISSUES. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(2), 190–194.

Duke, R., & George, T. (2016). *2016 Pilot Report Survey*.

Federal Aviation Administration. (2010). Weather-related aviation accident study 2003–2007. Retrieved January 24, 2019, from https://www.asias.faa.gov/i/studies/2003-2007weatherrelatedaviationaccidentstudy.pdf

Hansen, J. H. L. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, *20*(1–2), 151–173.

Këpuska, V. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). *International Journal of Engineering Research and Applications*, *07*(03), 20–24.

Kimura, T., Nose, T., Hirooka, S., Chiba, Y., & Ito, A. (2019). Comparison of Speech Recognition Performance Between Kaldi and Google Cloud Speech API (pp. 109–115). Springer, Cham.

McShefferty, D., Whitmer, W. M., & Akeroyd, M. A. (2015). The Just-Noticeable Difference in Speech-to-Noise Ratio. *Trends in Hearing*, *19*, 233121651557231.

Rami, M., Svitlana, M., Lyashenko, V., & Belova, N. (2017). Speech Recognition Systems : A Comparative Review. *IOSR Journal of Computer Engineering*, *19*(5), 71–79.

Stern, M. K., & Johnson, J. H. (2010). Just Noticeable Difference. In *The Corsini Encyclopedia of Psychology* (pp. 1–2). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Vipperla, R. (2011). Automatic Speech Recognition for ageing voices Ravichander Vipperla.

Watson, J. C., Lenz, A. S., Schmit, M. K., & Schmit, E. L. (2016). Calculating and Reporting Estimates of Effect Size in Counseling Outcome Research. *Counseling Outcome Research and Evaluation*, *7*(2), 111–123.

Yu, D., & Deng, L. (2016). *Automatic Speech Recognition*. https://doi.org/10.1007/978-1-4471-5779-3