An Exploratory Analysis of Air Traffic Controller Speech Intelligibility Using Voice Data from a Simulation Experiment

Yancy Vance Paredes and Nancy J. Cooke
Arizona State University
Mesa, AZ

Air Traffic Controllers (ATCs) communicate with pilots through radio communication. Speech intelligibility is vital in ensuring that the message is conveyed accurately. Factors such as speech rate affect this. Additionally, workload and stress have been shown to affect how people communicate significantly. In this paper, we attempt to analyze the voice data of ATCs who participated in a simulated experiment in the context of these non-verbal aspects of communication, particularly transmission length and speech rate. To better understand, we analyzed our data at two levels: aggregate and individual. Moreover, we focused on a single participant to see how such non-verbal characteristics evolve. Understanding these intricacies would contribute to building automated detectors in real-time voice transmissions that would leverage technology to avert any incidents brought about by stress and workload.

INTRODUCTION

Air Traffic Controller (ATC) is a demanding job as they play a crucial role in ensuring the safety of passengers and crews in an aircraft and the airspace in general. Costa (1996) identified sources of stress for ATCs, among which is task load, which is inherent to the demand of the work. They must be at their optimal performance (i.e., lower stress level) to be effective. Workload was identified as one of the determining factors in causing human error (Kantowitz & Sorkin, 1983). Humans tend to be reliable when the workload is moderate and does not change suddenly (Kantowitz & Casper, 1988). Workload is a multifaceted structure that cannot be studied directly but can be inferred from several quantifiable variables (Averty, Collet, Dittmar, Athènes, & Vernet-Maury, 2004). In the aviation literature, most studies that attempt to measure ATC workload typically rely on self-report subjective ratings (during or after an experiment) (Athènes, Averty, Puechmorel, Delahaye, & Collet, 2002). Such approaches are known to be prone to errors. Devising objective measures to help avert these incidents or alerts that forecast based on current working conditions could improve ATC safety. Some studies have attempted to devise measures for workload based on various task parameters to generate a workload index. A common approach is to count the number of aircraft being managed simultaneously. Another is to incorporate other information such as duration (i.e., length) or content of the radio exchanges and judgments by experienced observers to provide an objective workload measure (Athènes et al., 2002).

ATCs communicate with pilots through radio, exchanging spoken or verbal messages. Speech is primarily a verbal activity. Aside from the message, the receiver also receives extralinguistic information from the speaker (Prinzo, Lieberman, & Pickett, 1998). According to the Modulation Theory of Speech, information transmitted in speech can be classified into four qualities: linguistic, expressive, organic, and perspectival (Traunmüller, 1994). Speech is supported by nonverbal aspects or non-linguistic (or paralinguistic) characteristics such as intonation, voice quality, prosody, rhythm, jitter, pausing, and speech rate, which carry information on the psychological and physiological state of the speaker (Rothkrantz, Wiggers, Van Wees, & Van Vark, 2004; Traunmüller, 1994).

Most accidents in aviation could be attributed to human error, most of which are associated with communication (Billings & Cheaney, 1981). Miscommunication is affected by several factors, including the pilot's workload, audio signal quality, speech accents of ATC and pilot, English proficiency, and use of standard phraseology (Molesworth & Estival, 2015). Communication is susceptible to distortion of systematic, environmental, and internal factors that affect its comprehensibility or speech intelligibility (SI). There is growing interest in research on devising objective and non-intrusive SI metrics in fields such as speech processing (Feng & Chen, 2022; Sørensen, Boldt, & Christensen, 2019). SI is measured as the average percentage of words or phonetic units (e.g., syllables) that the listener can recognize. In the context of ATC, many factors contribute to low intelligibility, such as the quality of spoken utterance, background noise, speech accent, or a fast speaking rate. One of the major causes of aircraft accidents has been attributed to low SI of the ATC caused by a rapid rate of speech (O'hare, Wiggins, Batt, & Morrison, 1994). Speaking rate affects the intelligibility of speech and its comprehension (Kopparapu, 2015). It is reported that in the ATC industry, the average speaking rate is four syllables per second (Said, 2011, as cited in Hou, Tian, Chng, Ma, & Li, 2018). Hou et al. (2018) discussed the effects of a high speaking rate on vowels (stressed and distressed) and consonants which led them to propose an approach to stretch a given utterance to improve SI automatically. Such verbal miscommunications have also been identified as a causal factor in operational errors and pilot deviations (Prinzo & Britton, 1993). Goldman-Eisler (1961) found that an increase in speed of articulation (i.e., time per second) is associated with increased use of prepared and well-learned sequences. A similar result was found by Prinzo et al. (1998) where ATC participants reverted to an automatic response leading to routine communication (characterized by canned and repetitive nature) in dealing with high demanding scenarios as opposed to a cognitive response when dealing with a less demanding scenario. In another study, a voice stress analysis was conducted to estimate mental state or

psychological stress using several speech characteristics (Brenner & Shipp, 1988). They found that the average fundamental frequency, amplitude, and speech rate increased when task difficulty increased.

Several factors heavily affect the speaking rate (e.g., native or non-native speakers). For English, the average is between 130 and 200 words per minute (O'Sullivan, 2009 as cited in Kopparapu, 2015). There is an interplay between the complexity of the content and the rate (Kopparapu, 2015). For example, 130-145 for complex, 145-175 for average, and 175 for simple. Speaking rate (words per minute) can be computed if the number of syllables per second (sps) is known using the following formula: wpm = sps / γ * 60, where γ is a language-dependent constant that captures the average number of syllables per word. For English, it is suggested that γ = 1.5 which have been estimated from adult speech samples (Yaruss, 2000). Cardosi, Falzarano, and Han (1998) provided some recommendations to reduce communication errors between pilots and controllers. First, controllers are encouraged to speak slowly and distinctly. At a normal rate of 156 words per minute, 5% of the controller's instructions resulted in a readback error or request for a repeat. Increasing the rate to 210 words per minute increased the ratio to 12%. Another recommendation is to keep transmissions short with a maximum of four instructions per transmission.

Miscommunication could also be caused by the length of the message (Barshi, 1997; Cardosi et al., 1998). This is attributed to exceeding the capacity of short-term memory. Messages with three or more aviation topics led to a substantial increase in misunderstanding. Although Barshi and Farris (2013) found that speech rate by itself may not be a determining factor in misunderstandings in ATC communication, it is worth noting that in their experiments, their participants were undergraduate students and had different contexts (e.g., did not use aviation phraseology or a radar screen). Speech rate and message length play a role in comprehending spoken messages.

This paper looks at the various non-verbal aspects of radio communication and their relationship to workload in a simulated working environment. Specifically, we focus on the transmission length and the speech rate.

THE SIMULATION EXPERIMENT

A mid-fidelity radar simulation program was used to simulate the Terminal Radar Approach Control (TRACON) in Phoenix Sky Harbor (KPHX) Quartz West and South East arrivals. In this experiment, ATC participants interacted with three pseudo-pilots who were part of the research team. They were trained on phraseology, proper communication procedures, and simulation controls. They act as pilots of multiple aircraft assigned to their routes: Standard Terminal Arrival Route (STAR); ARLIN4, BLYTHE5, SUNNS8, and HYRDRR1. Alongside the ATC participant is a ghost controller who is also part of the research team and controls the tower. The simulation program logs pertinent information about a scenario, such as distance information of aircraft. Three 25-minute scenarios were designed with varying workloads by manipulating the traffic density. The first scenario was the baseline in which 4-5 aircraft are visible simultaneously. The second scenario was high-workload, in which 10-12 aircraft were under the control of the ATC. The third scenario was similar to the second except for the addition of predefined and fixed off-nominal events such as moderate turbulence, pilot deviation NORDO (No radio aircraft), runway switch, and minimal fuel advisory. The following constraints were imposed for all scenarios: (a) You must accept all handoffs from the center approach. The center will not hold. (b) You will only hand off to the final approach/KPHX tower. (c) No route modifications that result in aircraft leaving your control. (d) You will not request/issue commands to land at an airport other than the field destination. No alternate airports. You may only hand off to the final approach. (e) Keep aircraft in your airspace. No handoffs (except to 120.9 sector) and no point outs. (f) You must not declare emergencies.

Participants and Procedure

A total of six retired ATCs participated in this study. They all had prior experience in civilian TRACON (M=30 years, SD=10.97 years). Each participant was trained to use the software prior to the actual scenarios. They all had to perform the three scenarios with a break in between. To counterbalance the ordering effects, the order in which scenario they were going to perform was predetermined. Participants were debriefed and interviewed after the experiment. They were compensated for their participation in the experiment.

Multiple sensors were used throughout the experiment (e.g., eye-tracking, ECG, affect). However, we focus only on the voice data recorded in this analysis. This includes transmission exchanges between the ATC participant, the pseudo-pilots, and the ghost controller. These voice recordings were transcribed using an automated speech recognition (ASR) service. The ASR service automatically added timestamps to verbal messages. However, due to the domain-specific nature of the content of the transmissions, the transcripts had to be manually inspected and corrected.

EXPLORATORY DATA ANALYSIS

This study is part of a larger project that is fusing data in the National Airspace System to predict risk. The data reported here represent only a subset of the data we collect from the human component. The goal is to develop a model that combines the data for the best risk prediction. In this analysis, we focused on two non-verbal features of transmission, particularly the **length** and the **speech rate**. A transmission's length is the duration in seconds calculated using the timestamp information from the transcript. A script was written to automatically count the number of words in a transmission and obtain the word count. Using this information, we divided it by the transmission length to estimate the speaker's speech rate as words per minute. Additionally, Syllables, a Python library, was used to benchmark this estimate, which can estimate the number of syllables based on pattern matching. In this analysis, we looked into two speech rate types: words per second (wps) and syllables per second (sps). Most of the linguistic and speech literature used sps. We found a significant positive correlation between the two units ($r = 0.85, p < .001$). Therefore, in succeeding analyses, we decided to focus on sps to estimate speech rate.

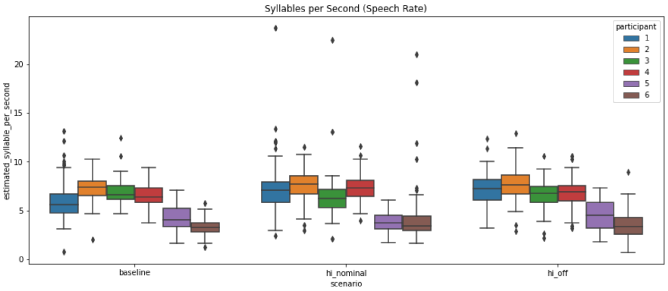

Figure 1. Speech rate of participants grouped by scenarios.

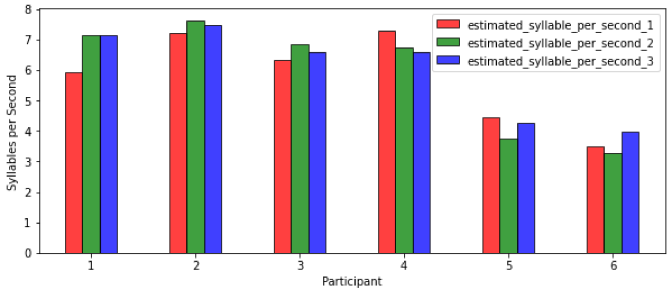

Figure 2. Speech rate in scenarios arranged based on assigned pre-defined order.

Descriptive Statistics

We begin by focusing only on the transmissions that originated from the ATC participant. Participants were exposed to three conditions with varying workloads. Were there any differences between the participants' transmission length and speech rate between scenarios? To answer this question, a Multivariate Analysis of Variance (MANOVA) was performed to determine how these differ across participants and scenarios. Figure 1 illustrates the distribution of the speech rate of each participant for all the transmissions in every scenario. Results showed a significant difference in transmission length and speech rate between participants and scenarios. We carried out individual ANOVA tests, performed post hoc analyses for the two factors, and tested against a Bonferroni-adjusted alpha level of .025. Interestingly, we found that only the baseline (M=4.13, SD=2.02) and high-workload (M=3.79, SD=1.93) had significant differences in transmission length. In terms of speech rate, both high-workload (M=6.13, SD=2.42) and high-workload off-nominal (M=6.08, SD=2.04) had no significant difference, while both high-workload conditions had significant differences from the baseline (M=5.68, SD=1.99). This finding is contrary to that of Prinzo et al. (1998) in which they did not find any significant differences in the speech rate of the participants between conditions (light and heavy scenarios). In our case, we did find some significant differences. It could be that our baseline and high workload conditions were more extreme than those of Prinzo and colleagues.

Individual Differences

As earlier mentioned, participants were exposed to the different scenarios at different orders. We wanted to identify how the order of the scenarios may have played a role in their features. Participants 1 and 2 had baseline as their first scenario; 3 and 4 had high-workload; 5 and 6 had high-workload off-nominal. As illustrated in Figure 2, there was no significant difference within participants across three conditions. It raises the question of whether workload affected the participants' physiological aspects as exhibited in speech. It is noteworthy that the speech rate of the participants in their first scenario has been the same for the other two scenarios regardless of workload. Further, participants 5 and 6 had lower speech rates and both were exposed to the high-workload off-nominal conditions at the beginning.

Scenario Analysis

So far, we have only looked at the summary statistics: the trend in terms of the length of their transmission and their speech rate (syllables per second). We would like to know how the participants respond to the varying workload as operationalized by traffic density on the radar. The following uses the five-second interval window in which missing values were interpolated based on the pattern. To account for individual differences since such non-linguistic features are very personal to the participants (Goldman-Eisler, 1961), we obtained the percentage of change to allow for comparison between participants and across conditions.

We zoomed into one participant as illustrated in Figure 3. We visualized the difficulty of a scenario by showing the progression of traffic density (red dotted line). The blue line represents the percentage of change in the speech rate of the participant. This represents the change from the previous to the current time interval (i.e., T and T-5 seconds). The figure also highlights the sections in gray those instances when a loss of separation (LOS) was detected. LOS refers to the event where the minimum safe separation between multiple aircraft is not maintained (i.e., three miles spacing horizontally within 40 miles of a major airport as defined by the Federal Aviation Administration), potentially leading to unsafe operational conditions. For this participant, it is interesting to note how the workload has not significantly affected the speech rate. Could it be that the participant managed the task at hand? We attempted to perform a time-series correlation between the differences in the workload and the differences between the speech rate of the participants. However, we did not find any significant findings. Further analysis is warranted especially considering the temporal component of the data.

LIMITATIONS AND FUTURE WORK

In this paper, we explored and focused on two non-content communication features, namely: transmission length and speech rate. In the future, we intend to explore various measures that allow for better quantification of workload and task complexity. For example, Averty et al. (2004) proposed a metric called Traffic Load Index (TLI), which accounts for the gravity or seriousness of the potential conflict and the urgency related to time pressure.

The current approach focuses only on summary statistics after the experiment. More value can be seen if predictive systems can be developed to forecast possible values in the fu-

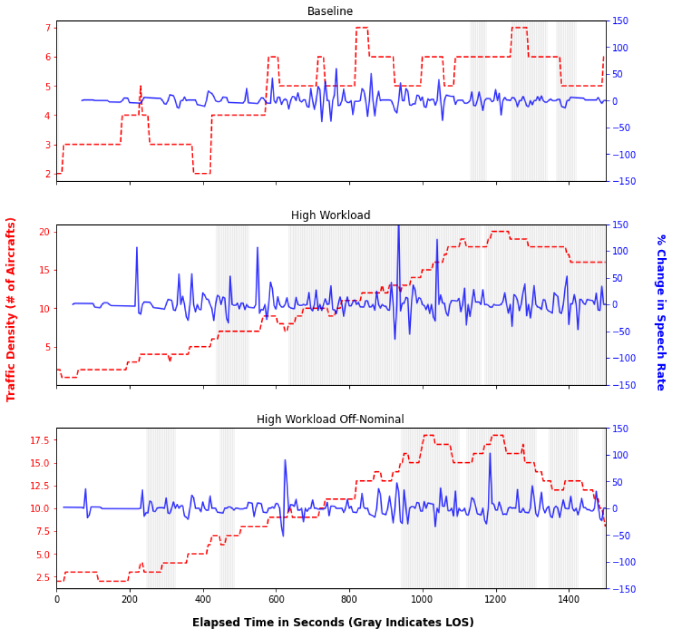

Figure 3. Overview of the three scenarios for a single participant.

ture and prevent potential accidents. Succeeding analyses must take into account the temporal component of the data. Because we are handling multiple time-series data, we can do a vector autoregression to account for the multiple factors and examine how one factor affects the other. With the growing popularity of machine learning, another direction would be to explore the performance of recurrent neural networks, such as LSTM (long short term memory), to account for the variations of the values that could potentially detect anomalous behavior and could indicate or alert potential dangers and prevent them from happening. Interest in the seq2seq approach (Sutskever, Vinyals, & Le, 2014) has started to rise especially using aviation datasets for forecasting. However, this would entail the collection of more data from more participants. Lastly, most of the data in this analysis were obtained from transcripts after running through an automated speech recognition system. In reality, such a process could add unnecessary delay and may sometimes be unreliable, especially for domain-specific corpus (phraseology of ATC). It requires a specific model that has been trained. We had to manually run and correct them before proceeding with the analysis. With the nature of radio communication, some research or potential direction has been looking at the voice signals themselves, or acoustic analysis (Prinzo et al., 1998). Voice signals could uncover speech rate, emotions, or stress indicators. These features can be extracted and analyzed in real-time, provided the existence of models. Finally, as previously noted, speech rate or any non-linguistic properties are personality-dependent (Goldman-Eisler, 1961; Prinzo et al., 1998). Therefore, in building future models, these idiosyncrasies have to be considered.

ACKNOWLEDGEMENT

The research reported in this paper was supported by funds from NASA University Leadership Initiative program (Contract No. NNX17AJ86A, Project Officer: Dr. Anupa Bajwa, Principal Investigator: Dr. Yongming Liu). The support is gratefully acknowledged.

REFERENCES

Athènes, S., Averty, P., Puechmorel, S., Delahaye, D., & Collet, C. (2002). ATC complexity and controller workload: Trying to bridge the gap. In *Proceedings of the International Conference on HCI in Aeronautics* (pp. 56–60).

Averty, P., Collet, C., Dittmar, A., Athènes, S., & Vernet-Maury, E. (2004). Mental workload in air traffic control: An index constructed from field tests. *Aviation, Space, and Environmental Medicine*, *75*(4), 333–341. doi: 10.1044/2018_JSLHR-H-17-0423

Barshi, I. (1997). *Effects of linguistic properties and message length on misunderstandings in aviation communication* (Unpublished doctoral dissertation). University of Colorado at Boulder.

Barshi, I., & Farris, C. (2013). *Misunderstandings in ATC communication: Language, cognition, and experimental methodology*. Routledge.

Billings, C. E., & Cheaney, E. S. (1981). *Information transfer problems in the aviation system* (Tech. Rep.).

Brenner, M., & Shipp, T. (1988). *Voice stress analysis* (Tech. Rep.).

Cardosi, K., Falzarano, P., & Han, S. (1998). *Pilot-controller communication errors: An analysis of aviation safety reporting system (ASRS) reports* (Tech. Rep.).

Costa, G. (1996). *Occupational stress and stress prevention in air traffic control*. International Labour Office Geneva.

Feng, Y., & Chen, F. (2022). Nonintrusive objective measurement of speech intelligibility: A review of methodology. *Biomedical Signal Processing and Control*, *71*, Article 103204. doi: 10.1016/J.BSPC.2021.103204

Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language and Speech*, *4*(3), 171–174. doi: 10.1177/002383096100400305

Hou, N., Tian, X., Chng, E. S., Ma, B., & Li, H. (2018). Improving air traffic control speech intelligibility by reducing speaking rate effectively. In *Proceedings of the 2017 International Conference on Asian Language Processing* (pp. 197–200). doi: 10.1109/IALP.2017.8300578

Kantowitz, B. H., & Casper, P. A. (1988). Human workload in aviation. In *Human factors in aviation* (pp. 157–187). doi: 10.1016/B978-0-08-057090-7.50012-6

Kantowitz, B. H., & Sorkin, R. D. (1983). *Human factors: Understanding people-system relationships*. Wiley.

Kopparapu, S. K. (2015). Non-linguistic speech processing. In *Non-linguistic analysis of call center conversations* (pp. 35–45).

Molesworth, B. R., & Estival, D. (2015). Miscommunication in general aviation: The influence of external factors on communication errors. *Safety Science*, *73*, 73–79. doi: 10.1016/J.SSCI.2014.11.004

O'hare, D., Wiggins, M., Batt, R., & Morrison, D. (1994). Cognitive failure analysis for aircraft accident investigation. *Ergonomics*, *37*(11), 1855–1869. doi: 10.1080/00140139408964954

Prinzo, O. V., & Britton, T. W. (1993). *ATC/pilot voice communications: A survey of the literature* (Tech. Rep.). doi: 10.21949/1503647

Prinzo, O. V., Lieberman, P., & Pickett, E. (1998). *An acoustic analysis of ATC communication* (Tech. Rep.). Office of Aviation Medicine Federal Aviation Administration.

Rothkrantz, L. J., Wiggers, P., Van Wees, J. W. A., & Van Vark, R. J. (2004). Voice stress analysis. In *Proceedings of the Seventh Conference on Text, Speech and Dialogue* (pp. 449–456). doi: 10.1007/978-3-540-30120-2_57

Sørensen, C., Boldt, J. B., & Christensen, M. G. (2019). Harmonic beamformers for non-intrusive speech intelligibility prediction. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 4260–4264). doi: 10.21437/Interspeech.2019-2929

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks.* Retrieved from `http://arxiv.org/abs/1409.3215`

Traunmüller, H. (1994). Conventional, biological and environmental factors in speech communication: A modulation theory. *Phonetica*, *51*(1-3), 170–183. doi: 10.1159/000261968

Yaruss, J. S. (2000). Converting between word and syllable counts in children's conversational speech samples. *Journal of Fluency Disorders*, *25*(4), 305–316. doi: 10.1016/S0094-730X(00)00088-7