DOT/FAA/TC-23/69

Federal Aviation Administration
William J. Hughes Technical Center
Aviation Research Division
Atlantic City International Airport
New Jersey 08405

Speech Intelligibility Assessment for an Unmanned Aircraft System Voice Test Bed

January 2024

Final Report

U.S. Department of Transportation
Federal Aviation Administration

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-23/69		

4. Title and Subtitle	5. Report Date
Title of Report: Speech Intelligibility Assessment for an Unmanned Aircraft System Voice Test Bed	January 2024
	6. Performing Organization Code ANG-E5B
7. Author(s) Alex Konkel, Ph.D., ANG-E5B	8. Performing Organization Report No.
9. Performing Organization Name and Address Federal Aviation Administration Human-Systems Integration Branch, ANG-E5B William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address Federal Aviation Administration UAS Research, Engineering, and Analysis Division (AUS-300) 490 L'Enfant Plaza SW, Suite 7225 Washington, DC 20024	13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code AUS-300

15. Supplementary Notes

16. Abstract

Unmanned Aircraft System (UAS) operations are rapidly increasing in the National Airspace System (NAS). Since no pilot is on board the aircraft, reliable communications are especially important for UAS. In the near future, it will be highly desirable for UAS to be able to communicate with air traffic control (ATC) via typical radio frequencies as well as potentially with other UAS. To that end, the Federal Aviation Administration (FAA) is evaluating the voice intelligibility of a UAS voice communications system. In this first evaluation phase, we evaluated voice intelligibility by having participants perform two tasks: the Modified Rhyme Test (MRT) and Message Completion Test (MCT). These tests provide an objective measure of how well people can understand the audio coming through the communications system. Participants also completed questionnaires to provide subjective judgments of the audio quality. Participants completed the tests and questionnaires on the five current FAA voice switch systems, listening from both the controller and pilot station at different times. Some participants were tested with the UAS system in the loop and some on the voice switches alone. The results suggest that there are small differences in intelligibility on the different voice switches, and potentially depending on which station is listening. Notably, there was no statistical impact of the UAS communications system. The results support further testing of the UAS system in a higher-fidelity environment, and that the FAA voice switches be further evaluated for audio quality.

17. Key Words	18. Distribution Statement
Unmanned Aircraft Systems - Human Factors Unmanned Aircraft Systems - Voice Intelligibility, Unmanned Aircraft Systems - Communications Systems	This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	79	

Contents

Figures

Tables

Acronyms

Acronym	Definition
ANOVA	Analysis of Variance
ARTCC	Air Route Traffic Control Centers
ATC	Air Traffic Control
C2	Command and Control
CRADA	Cooperative Research And Development Agreement
ETVS	Enhanced Terminal Voice Switch
FAA	Federal Aviation Administration
ICAO	International Civil Air Organization
IP	Internet Protocol
IPA	International Phonetic Alphabet
IRB	Institutional Review Board
IVSR	Interim Voice Switch Replacement
loo	Leave One Out
MCT	Message Completion Test
MRT	Modified Rhyme Test
NAS	National Airspace System
POLQA	Perceptual Objective Listening Quality Analysis
PTT	Push To Talk
RDHFL	Research and Development Human Factors Laboratory
RDVS	Rapid Deployment Voice Switch
STVS	Small Tower Voice Switch
UAS	Unmanned Aircraft System
UAV	Unmanned Aircraft Vehicle
UHF	Ultra High Frequency
VHF	Very High Frequency
VoIP	Voice over Internet Protocol
VSCS	Voice Switching and Control System
WJHTC	William J. Hughes Technical Center

Acknowledgements

This research was sponsored by AUS-300, the Federal Aviation Administration’s (FAA) Unmanned Aircraft System (UAS) Research, Engineering, and Analysis Division. The project was carried out with support and feedback from many team members, some of whom are listed below, and to whom we are very grateful. We are also thankful for the volunteers who participated in the research.

Kenneth Allendoerfer, FAA ANG-E5B, branch manager

Ken Crisler, Roberson & Associates, systems engineer

Melanie Flavin, FAA ANG-C35, electrical engineer/project manager

Jake Fowler, FAA ANG-E153, section manager

Ferne Friedman-Berg, FAA AUS-310, section manager

Mike Galusha, FAA ANG-E153, IT specialist

John Gross, FAA ANG-E64, IT specialist

Kiki Ioanidis, FAA ANG-E153, engineering tech

John Miller, FAA AUS-310, aerospace engineer

Eve Perchanok, DSoft Technology, Engineering, & Analysis, human factors specialist

Achal Singhal, FAA AUS-320, aerospace engineer

Executive summary

There is currently no established infrastructure to support consistent communication between unmanned aircraft systems (UAS) and air traffic control (ATC). This will be desirable in the near future as the number of UAS operations increase, and as operations expand into a more standard occurrence in controlled airspace. Such a communications system will be subject to various requirements, and audio transmissions sent through it will need to be clear and intelligible. This report describes a research study of voice intelligibility sent through a potential UAS communications system and the current Federal Aviation Administration (FAA) ATC voice switches. The effort also collects fresh data on the voice switches themselves. This first phase of research used test bed versions of the voice switches, housed at the FAA's William J. Hughes Technical Center (WJHTC). The UAS communications system was also a test bed version, provided via a cooperative research and development agreement (CRADA) with AURA Network Solutions, Inc.

In the pilot-to-ATC direction the UAS voice test bed sends audio input through an aviation audio panel, vocoder, and internet protocol (IP) interface before going to a UAS base station. The base station relays the signal via Ethernet cable to an unmanned aircraft vehicle (UAV) remote station where it again goes through an IP interface and vocoder before being converted to an ATC very high frequency (VHF) radio signal. Communication in the other direction (from a controller to the pilot) follows roughly the same path but in reverse. Given the various processing steps, there is the potential for degradation or other changes to the typical ATC radio signal that could affect how well speech is understood. For this study we chose to collect direct measures of intelligibility over the communications system. The primary goals of the study are to address:

- Baseline levels of intelligibility on the five FAA voice switches
- If intelligibility is affected by including the UAS communications system in the loop, including if there are any differences across the five voice switches

Following on previous research, we chose two tests. The first is the Modified Rhyme Test (MRT). The second test is the Message Completion Test (MCT) used by Friedman-Berg, Allendoerfer, and Deshmukh (2009). The MRT is a standardized intelligibility test used in many domains, while the MCT is an ATC-relevant test used in previous intelligibility assessments by the FAA. With the complementary features of the two tests, we believe the results will provide a good overall measure of speech intelligibility.

We recruited 17 participants from the WJHTC community. Participants reported normal, uncorrected hearing and none had ATC experience. Participants completed both the MRT and

MCT repeatedly while on different combinations of the five FAA voice switches and the two directions of communication (audio sent from the voice switch and heard at the pilot station, and audio sent from the pilot and heard at the ATC station). Thus they completed each test 10 times. Nine of the participants were tested with the proposed UAS voice system integrated into the communications loop, and eight without. This allowed for both an assessment of the FAA voice switches alone, and for a comparison between the current system and the UAS-included system.

The most notable result is that accuracy varied across the five FAA voice switches. Accuracy was highest on the Voice Switching and Control System (VSCS) and lowest on the Rapid Deployment Voice Switch (RDVS), with performance on the other three switches falling in between. The differences, while statistically significant, only covered a range of a few percentage points. Accuracy was notably lower at the pilot station on the MRT but was essentially equivalent to the ATC station on the MCT, suggesting at best a small effect of station on intelligibility. Accuracy was also numerically lower with the UAS communications system in the loop, but this effect did not reach statistical significance for either test perhaps due in part to the relatively small sample size. There was also no statistical interaction between voice switch and UAS integration, suggesting that the UAS system works fairly equivalently with each switch. Performance on the switches in general was around 80% for the MRT (although again lower on the RDVS) and 75% for the MCT. For the MRT, this would correspond to 'minimally acceptable intelligibility' according to the FAA Human Factors Standard. Based on the results, we make the following recommendations:

- Intelligibility levels should be verified through other means, such as additional tests.
- Intelligibility levels should be tested in a higher-fidelity environment, given that both the FAA voice switches and UAS communications system used in this study were test bed versions.
- Higher-fidelity testing could also include air traffic controllers and pilots who are more accustomed to the audio characteristics and ATC phraseology used in this test.
- Future users of the Message Completion Test should consider alternative means of administration to reduce the impact of memory and typing ability on performance.
- Further research should look into the potential impact, both objectively and subjectively, of using synthetic voices in intelligibility testing.

1 Introduction

Unmanned aircraft system (UAS) operations are rapidly increasing, with nearly one million UAS registered with the Federal Aviation Administration (FAA) as of May 2023. UAS operations are highly dependent on reliable signals since the aircraft have no pilot on board to control the craft or talk with air traffic control (ATC). For example, if the Command and Control (C2) link is lost or disrupted (known as a lost link), the UAS must follow pre-programmed commands instead of being flown by the pilot. A lost link situation can be dangerous as there is an uncontrolled aircraft and ATC is unsure what it will do. If the operator were able to communicate with ATC, they would be able to convey the lost link information and coordinate on how to handle the aircraft.

Beyond emergency situations, it would be beneficial to have an infrastructure for routine UAS-ATC communications. For example, transport companies might want to have unmanned aircraft carry cargo. Aircraft of that size would need to fly in controlled airspace and be able to maintain communications. Such a communications system is subject to various requirements such as National Airspace System Requirements Document (NAS-RD) 2013 (Federal Aviation Administration, 2013), which places limits on the voice communication latency between users and specialists (3.3.2.0-5.0 1 through 3). Ensuring that voice communications are clear and understandable is also important; a timely but incomprehensible message is as bad if not worse than a delayed message. This report describes an assessment of the intelligibility of speech sent through a UAS voice communications system. This effort also provides the opportunity to collect fresh data on intelligibility in the five FAA voice switches themselves (see Materials section for a description of the voice switches). Notably, this is the first phase of testing and as such is using a test bed set up at the William J. Hughes Technical Center (WJHTC). The test bed emulates a UAS voice system that could potentially be in use but no UAS were used during this phase of testing.

1.1 Background

The UAS voice test bed is integrated into the FAA Voice Communications Laboratory voice switch test-bed at the WJHTC (see Section 2.3). In the pilot-to-ATC direction, the test bed sends audio input through an aviation audio panel, vocoder, and internet protocol (IP) interface before going to a UAS base station. The base station relays the digitized signal via Ethernet cable to an unmanned aircraft vehicle (UAV) remote station where it again goes through an IP interface and vocoder to convert back into analog audio. This audio modulates a VHF radio signal for reception by ATC equipment. The ATC-to-pilot direction is similar but in the reverse order.

Given the various processing steps, there is the potential for degradation or other changes to the typical ATC radio signal that could affect how well speech is understood.

Speech quality can be evaluated in a number of ways. For example, Perceptual Objective Listening Quality Analysis (POLQA) (http://www.polqa.info/) is an algorithm that compares the input and output signals of a digital speech system, such as voice over internet protocol (VoIP), to predict a human assessment of speech quality. While such algorithms are objective, they are not a direct measure of speech quality or intelligibility. They are models based on subjective ratings of speech quality (a mean opinion score) from previous datasets. For this study we chose to collect direct measures of intelligibility. The primary goals of the study were to address:

- Baseline levels of intelligibility on the five FAA voice switches
- If intelligibility is affected by including the UAS communications system in the loop, including if there are any differences across the five voice switches

Following on previous research, we chose two tests. The first is the Modified Rhyme Test (MRT) (American National Standards Institute, 2009). This test uses sets of six rhyming (e.g., went, sent, bent, dent, tent, rent) or alliterative (e.g., pat, pad, pan, path, pack, pass) words. A single word is presented auditorily to the participant during each trial (e.g., "please select the word pad") and the participant chooses it from the set of six. Thus, the MRT evaluates voice intelligibility by ensuring that listeners can distinguish between similar-sounding words. While the MRT is an established speech intelligibility test, it is limited. The key words are all monosyllabic and intentionally confusable, and they are presented with no context. As such, it may not be representative of speech in ATC situations.

The second test is the Message Completion Test (MCT) used by Friedman-Berg, Allendoerfer, and Deshmukh (2009). This test uses ATC phrases and asks participants to repeat key pieces of information from the phrase. For example, the participant may hear "United 748, turn right heading 270, runway 28, cleared for takeoff" and be asked to report the call sign, turn direction, heading, and runway. Speech in the Message Completion Test is longer and more complicated than the MRT speech but has the benefit of being ATC-relevant. With the complementary features of the two tests, we believe the results will provide a good overall measure of speech intelligibility.

1.2 Previous research

The MRT traces back to a test developed by House, Williams, Hecker, and Kryter (1965). They built on previous research to design a set of materials for use in voice intelligibility testing with the particular aims of being quick to administer, requiring little to no special equipment, and requiring no training for the listeners. They argued that the multiple-choice aspect of the test, and use of non-specialized words, would require no training for listeners as well as reducing any potential learning benefits during repeated testing. House et al. tested their new materials by having two speakers record each key word in the frame “Number _____ is _____”, with the first blank corresponding to the numbered word set on the answer sheet (all 50 sets were present at once) and the second blank containing the key word for the participant to choose. The audio files were presented to 18 enlisted Air Force officers at six signal-to-noise levels. Testing occurred over 30 days. House et al. found that accuracy ranged from 35% at the lowest signal-to-noise ratio to 96% at the highest, with very little difference from the first to the last day of testing. The overall average accuracy was 76%. House et al. also noted a small but reliable difference in performance based on the speaker for the audio files, and differences in performance on the word sets based on the particular phonemes involved.

The MRT has since been incorporated into the American National Standards Institute’s standard on measuring intelligibility (American National Standards Institute, 2009) and is used in a wide array of applications. Relevant to the current research, Dunavold (2016) reported on the use of the MRT specifically during UAV flight (a Global Hawk). Their literature review discussed House et al.’s (1965) original work and cited other studies on the MRT since then, notably confirming the lack of a learning effect with repeated testing and the variability in performance due to speakers and key words. Dunavold also suggested guidelines for intelligibility: satisfactory if performance (adjusted for guessing) is greater than 80%, marginal if between 70% and 80%, and unsatisfactory if less than 70%. However, it is unclear how those numbers were chosen. In their study, Dunavold required participants to score at least 90% on a training test and were trained as speakers before conducting the UAV test. Thus the participants were extremely familiar with the word sets. Dunavold also noted that live speakers were used instead of audio recordings because of concerns over possible distortions being introduced by the extra equipment. During the UAV test, four radio frequencies were used resulting in participants being tested twice (if they attended one session) or four times (two sessions). The results were accuracy of 84% with little difference across radio frequencies. The decrease in performance from the 90% training score suggests that intelligibility on the UAV voice system was lower, but not dramatically so, than a person speaking in the same room.

While Dunavold’s (2016) source for MRT standards is unclear, the FAA Human Factors Design Standard ((Ahlstrom, 2016); citing MIL-STD-1472) gives thresholds of 97% for exceptionally high intelligibility, 91% for normal acceptable intelligibility, and 75% for minimally acceptable intelligibility (but noting this is not acceptable for operational equipment).

Friedman-Berg et al. (2009) used the MRT to examine speech intelligibility for various codecs that could be used for VoIP communications. They tested certified air traffic controllers as well as participants with some ATC familiarity but no direct experience, although the results were not split by background and no differences are described. Given the goal of comparing codecs, all audio was recorded and presented to participants via computer. Participants completed six MRT tests in a single session, one test each for the baseline uncompressed audio condition and five codecs. Friedman-Berg et al. found statistically significant differences in accuracy across the different conditions but performance was high in all cases, ranging from 90% for the worst codec to 98% for the baseline condition. This high level of performance may mitigate the concerns that Dunavold (2016) had about recorded versus live audio, although this audio was not sent over radio. Friedman-Berg et al. also collected subjective ratings of intelligibility and acceptability via questionnaire after each test. Ratings were high in all cases, with a median rating of seven (on a one to seven scale) for the baseline condition. The lowest codec still received a median rating of five (corresponding to an intelligibility rating of "I could understand most of what was said" and an acceptability rating of "in most foreseeable situations, the audio would be satisfactory").

Friedman-Berg et al. (2009) administered a MCT in addition to the MRT. This test consisted of five phrases typical of air traffic controller speech, such as a traffic callout and heading/altitude instruction. Friedman-Berg et al. noted the importance of using an ATC-relevant task due to differences between ATC communications and other speech, such as technical and non-standard words. Participants heard the full phrase and had to fill in critical information on a response sheet, such as the aircraft's call sign or altitude. Due to the small number of trials, Friedman-Berg et al. did not conduct a statistical analysis of the results. However, accuracy was reported as 93% or higher in all conditions, and subjective ratings of intelligibility and acceptability had a median score of six or higher on a scale of one to seven.

A smaller body of research has used similar content-appropriate test for speech intelligibility. For example, LaDue, Sollenberger, Belanger, and Heinze (1997) used a version of the MCT with only three phrases as part of a test of different vocoders, also including background noise (e.g., a jet cockpit) as a factor. They found uniformly high performance (nearly 100%) in all conditions. Blue-Terry and Letowksi (2011) examined a Callsign Acquisition Test for military use. The Callsign Acquisition Test uses combinations of International Civil Air Organization (ICAO) phonetic alphabet names (alpha, bravo, charlie, etc.) and single-digit numbers to make 126 different artificial callsigns. They also used the MRT and tested across various signal-to-noise levels. Blue-Terry and Letowski found that performance varied with signal-to-noise level from 66% to nearly 100% on the Callsign test and 40% to 77% on the MRT.

On the whole, previous research suggests that the MRT and MCT should be complementary. The MRT has the benefit of using commonly known words that require no training and also has no learning effect, but the drawback of not using context found in typical speech or the content used in air traffic control. The MCT uses appropriate content for air traffic control, but performance tends to be extremely high, which can make it difficult to find statistical differences in an experimental study. In both cases, overall performance varies with the audio quality (e.g., background noise, signal-to-noise ratio, audio processing), the speaker, and the exact materials (i.e., the word or phrase heard on a given trial).

2 Method

2.1 Participants

We recruited 17 participants from the WJHTC community. We asked participants if they have normal, uncorrected hearing but otherwise there were no requirements to participate. Participation was voluntary and uncompensated, and the study was approved by the WJHTC local Institutional Review Board (IRB).

In regard to demographics, the participants consisted of 4 women and 13 men. Their ages ranged from 23 to 63 with an average of 48.7. Four participants completed the testing in two sessions (see the Section 2.4). Two participants reported having pilot experience, although they did not perform differently than the other participants, and none reported any ATC experience. Being employees at the WJHTC, the participants had varying levels of general familiarity with air traffic control, ATC phraseology, and the voice switch systems.

2.2 Facilities and personnel

Engineering psychologists and contract support personnel from the Research Development and Human Factors Laboratory (RDHFL) located at the WJHTC conducted the testing. They conducted the testing in the Voice Communications laboratory where the FAA voice switch test beds are located. Engineers from the Voice Communications laboratory configured the voice switches and related equipment.

2.3 Materials

2.3.1 Voice communications system

The FAA and AURA Network Systems, Inc. entered a Cooperative Research and Development Agreement (CRADA) to use a UAS communications system developed by AURA. AURA's system, in the long-term, will use ground-based cell stations to enable UAS operators to communicate with their aircraft as well as ATC via the standard push-to-talk (PTT) radio system. An example diagram of the AURA to ATC radio path is shown in Figure 1. A test bed version of the AURA system was installed in the Voice Communications laboratory and configured to interface with the various FAA voice switch test beds.

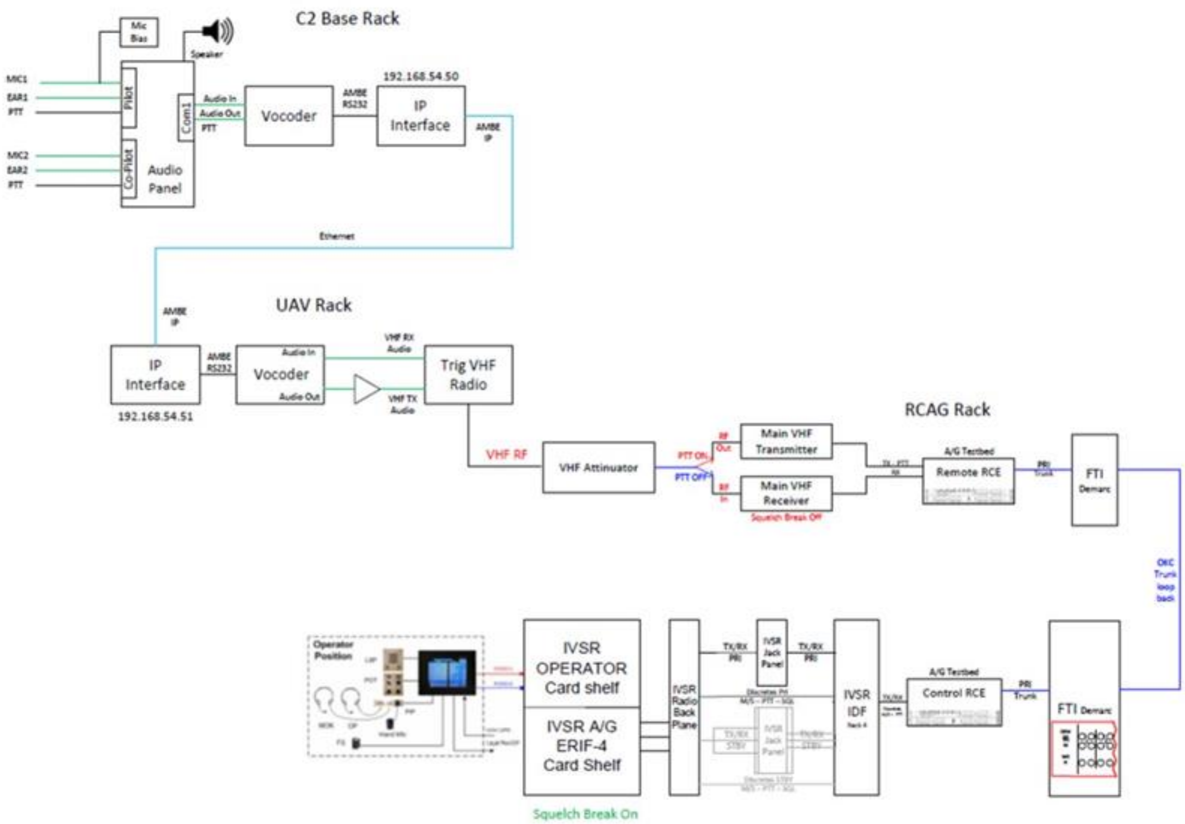

Figure 1. Example diagram of AURA voice test bed integrated into FAA voice switch test bed

The FAA Voice Communications laboratory houses five voice switches. Each of these is equivalent to current fielded systems with the exception of the Rapid Deployment Voice Switch (RDVS), which was recently decommissioned (i.e., the current field RDVS is a newer version of the switch).

- VSCS - The Voice Switching and Control System (VSCS) is an integrated air-to-ground and ground-to-ground voice and radio control and confirmation communication switching and control system for Air Route Traffic Control Centers (ARTCC).
- RDVS - The RDVS provides intercom, interphone, and radio communications via digitized voice and data buses. The system provides voice communications between ATC

positions and other local and remote ATC positions (through intercom functions), ATC positions in adjacent and remote facilities (through telephone functions), and with aircraft (through radio functions).

- ETVS - The Enhanced Terminal Voice Switch (ETVS) is a non-blocking, fully integrated, digital voice communications system. System design allows operators to communicate with each other and establish communications between the radio, telephone, and interphone voice paths served by the system.
- IVSR - The Interim Voice Switch Replacement (IVSR) system is a fully digital, totally non-blocking voice communication switching system. Voice and signaling data are transferred via redundant digital high-speed highways. The core system hardware of the IVSR system comprises redundant circuit boards and modules to ensure no single fault interrupts ATC system communications. The IVSR system provides a flexible interface architecture that allows connection to radio equipment and all IVSR-specified signaling protocols.
- STVS - The Small Tower Voice Switch (STVS) is an integrated air ground and ground-ground voice switching system. The STVS provides for the selection, interconnection, and activation of communications paths between operating ATC positions, other ATC facilities, local radios, and remote radios.

While the voice systems can be used in conjunction with various radios in the field, only a single type was used during testing:

- RCE - The General Dynamics/CSTI Model CS-2330/RCE Radio Control Equipment (RCE) provides simultaneous voice and data transmission over leased 4-wire telephone lines. The equipment is ideally suited for ATC applications where audio and radio signals need to be transported to remote radio facilities.
- CAVU 2100 RX/TX - The ITT/Park Air Systems CAVU 2100TM Multi-Mode Digital Radio is a modular software reprogrammable radio suite. The MDR receiver and MDR transmitter are Air-to-Ground (A/G) traffic control radios, compatible with current A/G radios used by the FAA. These receivers and transmitters operate in the frequency range of 112.000 MHz to 136.975 MHz using either 25 kHz or 8.33 kHz channel spacing.

Participants listened to audio files with an ATC-style Plantronics headset when listening as ATC. This equipment includes SHS1890 PTT with carbon-type microphone amplifier, PJ-7 connector, and selectable switch operation (momentary or locking modes) paired with the HW251 over-the-ear headset. Participants never used the PTT capability on the headset. When listening as a pilot, participant used a RadioShack headset. The headset had two earcups but was set to mono output to match the single-ear style of the ATC headset. It also had volume control on the cord, which researchers attempted to keep at a single location throughout testing.

2.3.2 Voice intelligibility tests

The Modified Rhyme Test consists of 50 sets of six words. Each set of six words differs from one another only in their initial or ending phoneme. We downloaded the source audio files from https://www.nist.gov/ctl/pscr/pscr-audio-source-files, which consists of nine different voices saying the entire set. Thus there are $9 \times 6 \times 50 = 2700$ total audio files. Participants were given a verbal description of the test by the researchers, and also saw a written instruction set on the screen prior to each test (Figure 2).

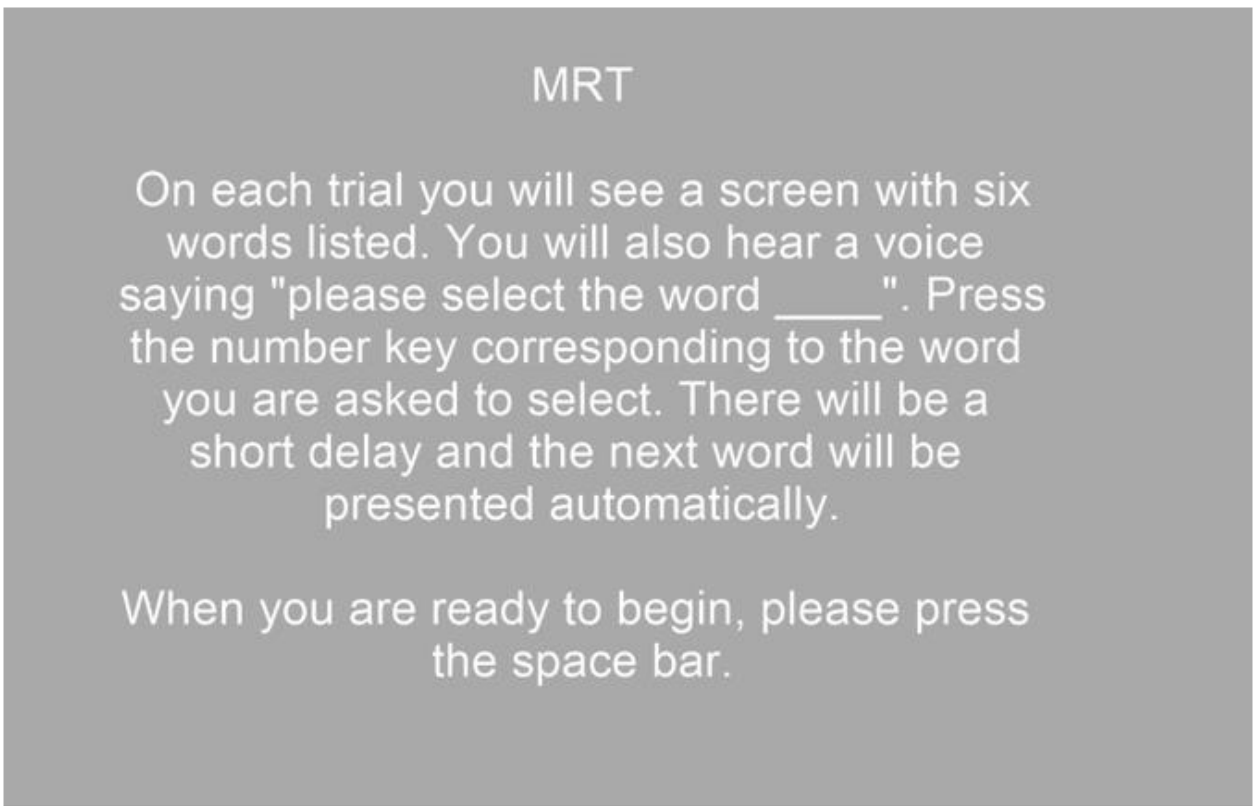

Figure 2. MRT instruction screen

We recreated the MRT in PsychoPy (see Section 2.3.4). Participants saw a given word set on each trial (Figure 3) with the options appearing simultaneously with the audio file playing over the headset. The audio only played once but the options remained on the screen until the participant made a response. The next trial did not begin until the participant pressed a button to proceed (see Section 2.3.4).

Figure 3. Example of an MRT test trial screen

We adapted the Message Completion Test used by Friedman-Berg et al. (2009). That study used five sentence frames with six different answer sets each. We expanded the set to 12 sentence frames each from the controller and pilot perspective (24 sentences total). Each sentence had five answer sets. The list of sentences is in Appendix A. To create multiple voices for the MCT as in the MRT, we used text-to-speech software to create audio files of the sentences. Murf (murf.ai) uses artificial intelligence to generate audio files with different voices and voice characteristics (e.g., ‘general’ or ‘excited’). We selected five voices that were fairly generic (e.g., not ‘excited’) and suitable to stand in as a controller or pilot. The specific settings are included with the sentences in Appendix A. In total there are 5 × 24 × 5 = 600 audio files.

While the MRT was straightforward, the Message Completion Test required more instruction. Researchers gave a verbal description of the test prior to the first run, and written instructions also appeared at the beginning of each run (Figure 4).

Figure 4. MCT instruction screen

Participants then saw an example trial with the prompt and audio from a potential trial as well as the expected answer for that trial (Figure 5). The example allowed the researcher to better describe what the participant might hear and how they should type it in. In particular, participants were encouraged to use shorthand while listening to the audio and then go back to fill in the message (e.g., for the example audio, they might begin by typing "den, l8975, 8, f" then go back and expand to the answer of Denver, Lindbergh8975, 8000, foxtrot). Researchers encouraged this system based on participant feedback from preliminary testing to emphasize the listening aspect of the test over trying to hold the message in memory and then typing it out.

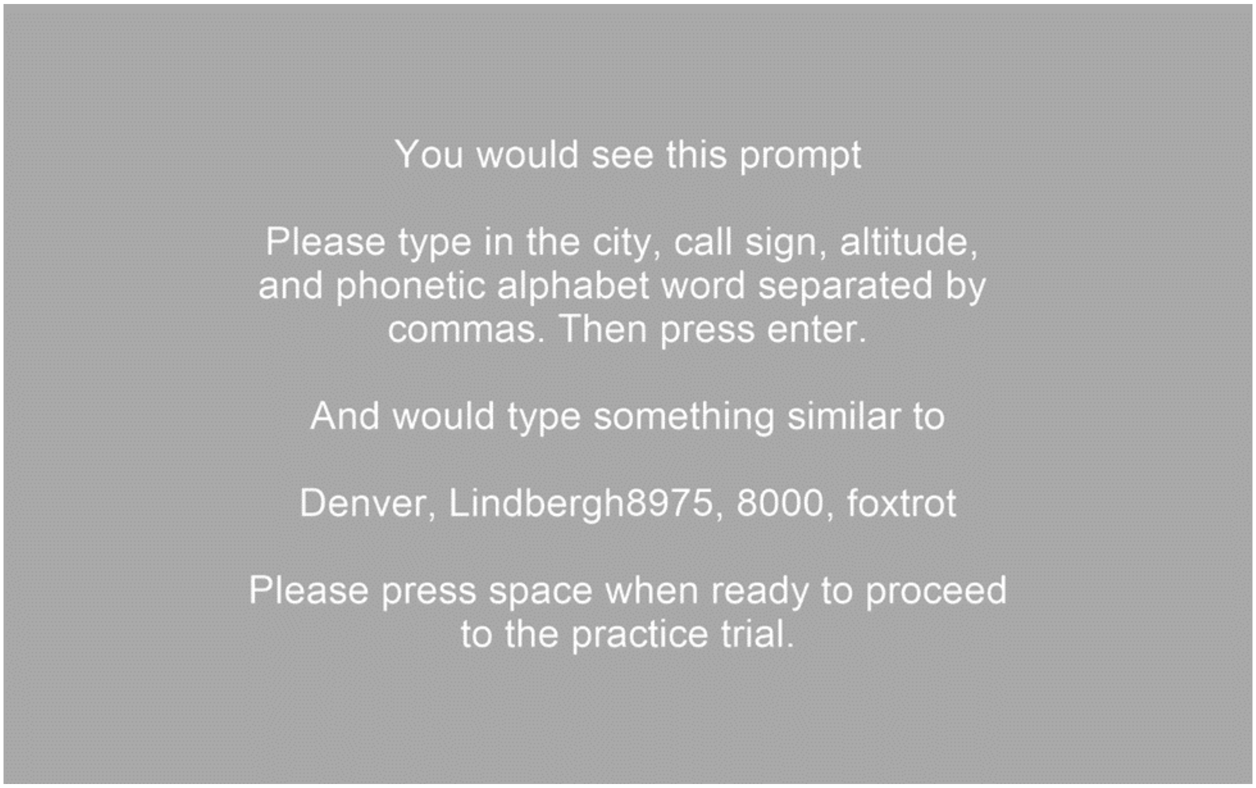

Figure 5. MCT example screen

On each trial (Figure 6), participants saw a prompt that told them what information to enter from the message they were going to hear. The audio played one second after the prompt appeared. The audio only played once but the prompt remained on the screen until the participant pressed the enter key. The next trial did not begin until the participant pressed a button to proceed (see Section 2.3.4).

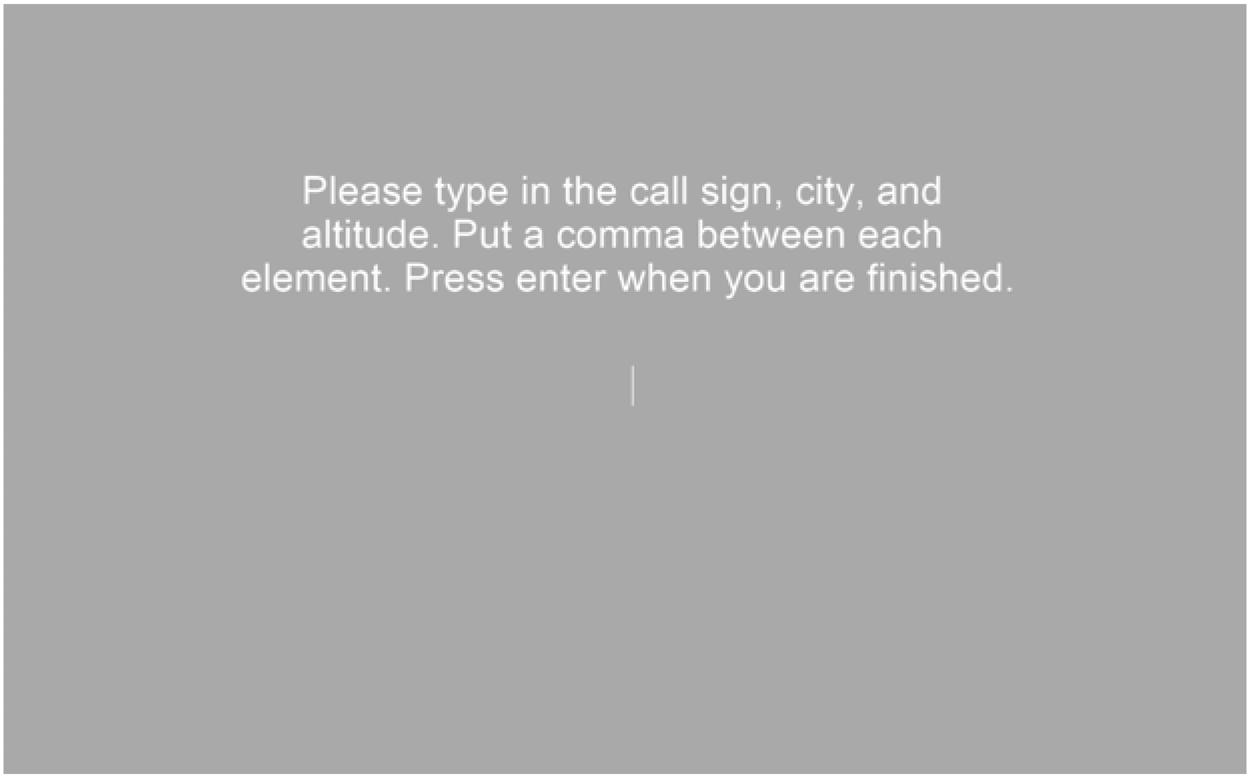

Figure 6. MCT example test trial screen

2.3.3 Questionnaires

In addition to the accuracy data generated by the voice intelligibility tests, we collected subjective ratings of intelligibility and audio quality via a questionnaire. We based the questionnaire on that used by Friedman-Berg et al. (2009). Their questionnaire consisted of two ratings questions, asking participants to respond on a Likert scale from one to seven as to the intelligibility and acceptability of the audio they heard during a test. There was also an open-ended question for the participants to provide other feedback. The questionnaire is in Appendix C. We also asked participants to fill out a basic background questionnaire (Appendix B) for demographic purposes.

2.3.4 Test administration

The voice intelligibility tests were administered on a standard PC laptop. The experiments were programmed using the PsychoPy package (version 2022.2.4; https://www.psychopy.org) for Python software (version 3.8; https://www.python.org). The experiment code played the appropriate audio file on each trial and recorded the participant's response. It also administered the questionnaire.

The FAA voice switch was configured to allow for continuous audio transmission. The UAS communications system, however, was set to disable audio after 35 seconds (as is typical to avoid 'stuck mic' situations). Thus for test cycles where the participant was tested with the UAS system integrated and audio was injected into the pilot side (the participant was listening at the ATC station), it was necessary to push-to-talk on at the beginning of each trial and off at the end. The researcher did this to better allow the participant to focus on the test itself. The experiment code displayed screens before and after a trial (the screens seen in Figure 3 and Figure 6) with reminders to toggle the push-to-talk on or off as appropriate.

Testing occurred at the voice-switch test bed in the Voice Communications laboratory. This is an open-air area with other laboratories and equipment nearby. Thus there was consistent background noise during testing, typically fan noise from the computer racks and other equipment in the area. Occasionally people would walk by having a conversation, or there was construction noise. The latter two examples were rare, but performance was very likely affected by these extraneous sources.

We did not attempt to standardize volume levels across the experimental conditions prior to testing. We believe that identifying any changes to the audio caused by the voice switches and/or UAS system is part of the focus of the study. That said, to maintain a consistent level throughout testing, we maintained the same volume settings on each voice switch and the laptop throughout. As noted previously, the pilot station headset had an inline volume control-slider, which we attempted to keep in place throughout testing. To determine if there were any differences across the equipment configurations, ANG-E153 engineers conducted a standard audio level assessment on each voice switch with and without the UAS communications system in the loop. The results are in Table 1. In every case, an audio signal was injected at -10dB and 1004 Hz. As can be seen in the table, the signal came out at a consistently (from each voice switch) lower volume on the pilot headset. The levels were more variable but typically lower when injected into the pilot side and measured through the ATC headset. Transmitting through the UAS system in either direction slightly changed the frequency of the signal, from 1004 Hz to 1000 Hz. This is expected for digital voice coders designed for human speech as opposed to pure tones. The volume level was higher when listening through the pilot headset and again variable but typically lower when listening through the ATC headset. These measurements will be noted again later when discussing the study's results.

Table 1: Measured Audio Levels for each Equipment Configuration

FAA Voice Switches Alone			
Audio Injected To	Audio Received At	Volume Out	Frequency
VSCS	Pilot station	-18.2 dB	1004 Hz
STVS	Pilot station	-18.2 dB	1004 Hz
IVSR	Pilot station	-18.2 dB	1004 Hz
ETVS	Pilot station	-18.2 dB	1004 Hz
RDVS	Pilot station	-18.2 dB	1004 Hz
Pilot station	VSCS	-9 dB	1004 Hz
Pilot station	STVS	-14.9 dB	1004 Hz
Pilot station	IVSR	-19.7 dB	1004 Hz
Pilot station	ETVS	-16.4 dB	1004 Hz
Pilot station	RDVS	-20.9 dB	1004 Hz
FAA Voice Switches With UAS Communications System In The Loop			
Audio Injected to	Audio received at	Volume Out	Frequency
VSCS	UAS pilot station	-2 dB	1000 Hz
STVS	UAS pilot station	-1.9 dB	1000 Hz
IVSR	UAS pilot station	-1.9 dB	1000 Hz
ETVS	UAS pilot station	-1.9 dB	1000 Hz
RDVS	UAS pilot station	-1.9 dB	1000 Hz
UAS pilot station	VSCS	-21 dB	1000 Hz
UAS pilot station	STVS	-15.1 dB	1000 Hz
UAS pilot station	IVSR	-8.7 dB	1000 Hz
UAS pilot station	ETVS	-20 dB	1000 Hz
UAS pilot station	RDVS	-16.4 dB	1000 Hz

2.3.5 Design

The dependent measures for the study were accuracy on the two voice intelligibility tests and responses made on the questionnaire. We also collected response time for the MRT; due to the nature of responses on the MCT (see Section 2.4), we did not look at response time for that test. The three independent variables were the FAA voice switch (ETVS, IVSR, RDVS, STVS, and VSCS) system, which station the participant listened at (pilot or ATC), and whether the UAS test bed was integrated into the voice system or not. The study used a 2x2x5 mixed design such that each participant was tested on each voice switch and station (within-subjects) but either with or without the UAS communications test bed in the audio loop (between-subjects). Testing on the five voice systems was important to ensure that there were no integration issues for the UAS test bed on any potential FAA system. Testing without the UAS system in the loop was important not just to provide a control condition but additionally to evaluate the FAA voice switches in isolation. We tested both listening stations because AURA Networks advised that there could be differences in audio quality depending on the transmission direction. The choice of manipulating between or within-subjects was practical. Changing between voice switches or listening station took only a minute, making it easy to do between runs of a single participant, but changing the UAS integration status took longer, making it impractical to do while a participant was waiting. The combination of within-subjects factors led to participants completing 10 test cycles (2 listening stations × 5 voice switches).

In addition to the independent variables, we accounted for other test-specific variables. For the MRT, we took the word set as the basic unit of the test. Thus participants completed 50 trials in each test, one for each set. The order of the sets was randomly selected by PsychoPy for each testing session, so participants saw the word sets in different orders across the 10 tests. The word the participant heard from a set was randomly selected, as was the voice reading each word. We followed a similar procedure for the MCT using the sentence frame as the base unit. Since previous research used a small number of sentences and we wanted to keep test time short, we did not use all 24 sentences in a single test. We randomly selected 10 sentences from the set and again randomized the frame completions and voice.

In a given run, participants completed both tests: the MRT and MCT. The tests were always administered in that order. The order of listening station/voice switch combination was random across participants. Whether a participant was in the UAS system in the loop or out of the loop condition was selected at random with the constraint that we tested roughly equal numbers of participants in each condition. With 17 participants in the sample, we tested nine with the UAS system in the loop and eight with the system out of the loop.

2.4 Procedure

Each participant was tested individually. A complete session of 10 test runs lasted 3-4 hours; some participants chose to complete them in two sessions. Those participants completed five runs in one session and then returned to complete the other five a different time, typically a week later. Researchers set the UAS system configuration prior to the participant arriving, and set the voice switch and computer equipment configuration prior to each run. Participants began by receiving a brief introduction to the study from the researchers and then completed the background questionnaire (Appendix B). The WJHTC IRB approved the study and determined it to be exempt, so no informed consent was necessary. Participants then went through the testing procedure.

Prior to the first time completing each test, the researcher gave a more detailed description of the test to the participant. The researcher also described the push-to-talk system (see section 2.3.4). When it was not necessary to push-to-talk, the participant progressed through trials at their own pace since they did not have to coordinate with the researcher toggling the switch.

To conduct a test, the researcher used a laptop to run the PsychoPy experiment files, which played the appropriate audio files directly into the voice system, and which the participant heard via headphones (one of two sets, as described previously). The participant followed prompts on the laptop to either select the word they heard (for the MRT) or type a response based on what they heard (for the MCT). After each test was finished, the participant completed a questionnaire on the laptop and was offered a short break while the voice switch configuration and headphones were changed as necessary.

Participants completed 20 voice intelligibility tests in total, the MRT and MCT 10 times each, in a randomized order as described in the design section. Each MRT session lasted five to 10 minutes and each Message Completion Test lasted 10 to 15 minutes. At the end of a session the participant was reminded of their next testing appointment or, if it was their last or only testing session, debriefed and thanked for their participation.

3 Data analysis

Our general plan for data analysis was to use a Bayesian approach with generalized linear regression models. We used generalized linear models because the dependent variables (accuracy at the trial level, response time, and questionnaire ratings) are non-normally distributed. The best way to analyze such data is with a generalized linear model as opposed to a typical method (e.g., ANOVA) that assumes normal data. We used Bayesian methods as opposed to frequentist tests because the results more directly correspond to researchers' intuitions about statistics (Kruschke & Liddell, 2018), and because Bayesian methods allowed us to use results from preliminary testing for the priors. Bayesian methods also lend themselves to a more nuanced approach to the results than the black-and-white thinking encouraged by frequentist hypothesis testing and p-values. That said, we report 95% posterior distribution intervals that roughly correspond to typical 95% confidence intervals. If we refer to an effect as 'statistically significant', it means that the 95% interval excludes zero.

The general approach was to model each dependent variable with the 'largest' model of interest and then remove predictors while checking if the model fit significantly decreased. The simplest model that statistically fit as well or better than a larger model was chosen as the final model. The largest model included a three-way interaction between the three independent variables (voice switch, listening station, and UAS comms system integration) and main effects of the covariates. The model choosing procedure next took out the three-way interaction and compared the two. If the two-way interaction model fit as well or better, the model was shrunk again by removing the two-way interactions and fitting a model with only main effects. If the two-way interaction model fit better, we fit models taking out one of the interactions at a time to see if they fit better. If the main-effects-only model fit as well or better than the two-way interaction model, it was selected as the best model. All main effects were kept in a final model so that the effects could be inspected. The model also included a multilevel (also sometimes called hierarchical or random effects) component such that participants could have varying intercepts and varying slopes for runs and trials (i.e., different learning curves). The varying slopes were also removed and tested as part of the model choosing procedure. All data processing and analysis was conducted in R version 4.2.2 (R-project.org). Model fitting used the brm function and model comparisons used the loo (leave one out) function from the BRMS package version 2.18.0 (Burkner, 2017).

While we did not expect any demographic data from the background questionnaire to affect the results, we included age, gender, and whether the participant split their testing session or not as covariates in each model. The specific final model for each dependent variable will be described in its respective section below.

3.1 Modified rhyme test analysis

The two dependent measures from the MRT were accuracy and response time. Due to technical problems (such as a fault with the voice switch) or experimenter error, two runs from two participants were not analyzed, along with two trials from two other participants. The analyzed data set consisted of 98.8% of the possible full data set (50 trials each from 10 runs each from 17 participants).

3.1.1 Accuracy

Based on preliminary testing, we set a small positive prior (a normal distribution with mean 0.1 and standard deviation 1) for run (first through 10^{th} test cycle) and trial (first through 50^{th} within a run) because it appeared there may be a small learning effect. We set a negative and more variable prior (normal distribution with mean -1 and standard deviation 2) for the effect of station because we saw a large decrease in performance at the pilot station, although this was due in part to practical issues that were changed for final data collection. The prior for the intercept, corresponding to average performance, was set to correspond to around 75% accuracy (normal distribution with mean 1.1 and standard deviation 3). All other priors were set to a normal distribution with mean 0 and standard deviation 2, assuming no effect but allowing one if the data suggest it.

The dependent variable for the model was response (correct or incorrect) at the trial level, modeled using a Bernoulli distribution. The covariate predictors were gender, age, whether the participant split the test session or not, run, trial, the voice heard on that trial, and the word set for the trial. The final best-fitting model contained only main effects, with no interactions between predictors, and only a random intercept across participants (no random effects for run or trial).

Overall accuracy was 79%, with performance across participants ranging from 71% to 84%. There were no statistical effects of gender, age, or splitting the test session.

While there is a visual impression of an increase in performance across runs (Figure 7), the effect was not statistically significant.

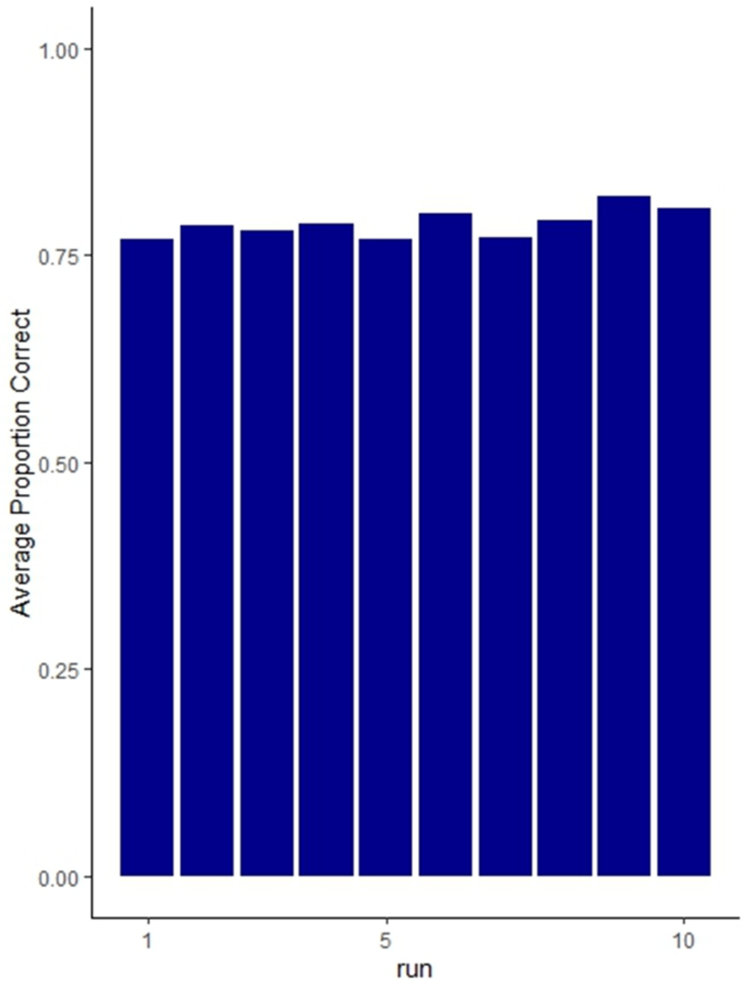

Figure 7. MRT accuracy across runs

To illustrate, Figure 8 shows the posterior distribution for the model's comparison of the effects of the first and tenth runs. In Bayesian inference, the posterior distribution is the combination of the prior distribution and the observed data. Instead of a single-number result, like the data mean, there is a distribution showing the plausible values that parameters in the model might have. The 95% density interval for the posterior, corresponding to a confidence interval and denoted on the graph with black vertical bars, includes zero. Just as with a confidence interval, this suggests that there is no statistical difference in the two parameters; e.g., it is plausible that they are the same. Further, the posterior distribution ranges from accuracy increasing at most 5% over the course of testing (the -0.5 to the left of the graph, as the comparison is first run minus tenth run) to actually decreasing by 2%. Thus we are not confident that there is an effect of run on accuracy. There was also no statistical effect of trials within the runs.

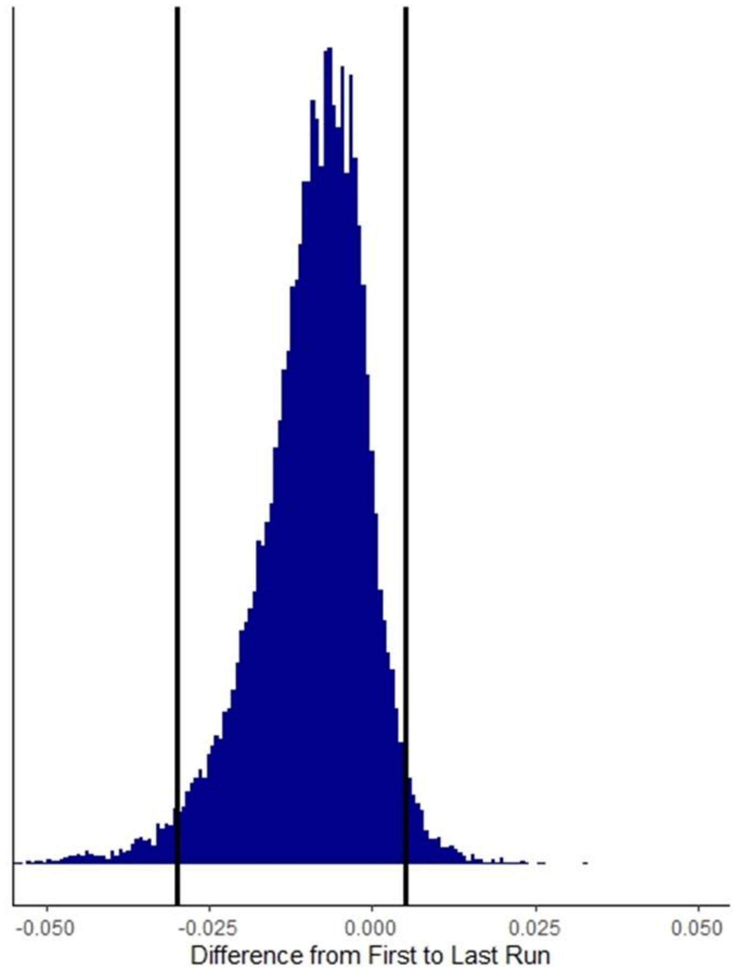

Figure 8. MRT accuracy model comparison of first and last run

There were significant material effects. Accuracy varied depending on which word set was tested on a trial, ranging from 62% to 92%, and also varied depending on which voice spoke the word to be chosen, ranging from 75% to 83%.

There were also significant station effects. Accuracy varied depending on which station the participant listened at, with accuracy being notably lower at the pilot station (mean 75% compared with 83% at the ATC station). The model (Figure 9) shows the effect likely falls between 2 and 8% and the entire posterior is above zero, stating that there is virtually no chance accuracy is equal in the two conditions but there is some chance the difference is small.

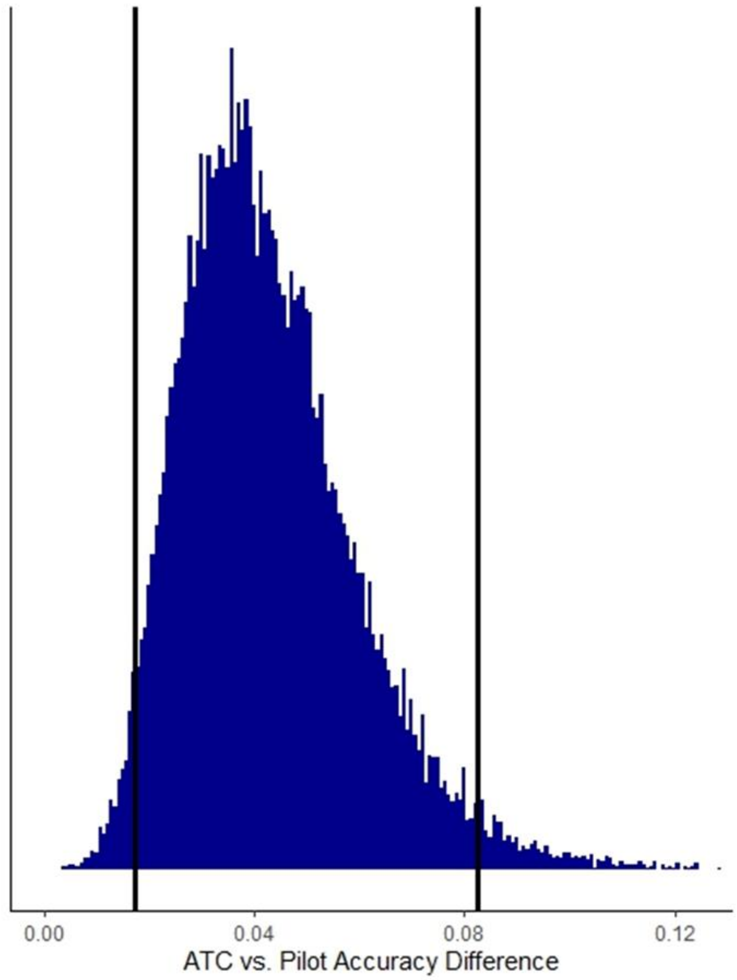

Figure 9. MRT accuracy model station effect

On the key question of intelligibility on the FAA voice switches, accuracy did vary from switch to switch (Table 2). Performance was highest on the VSCS and lowest on the RDVS. The model results show that accuracy on the STVS and RDVS are significantly different, as is accuracy on the VSCS and STVS, suggesting roughly equal performance on the top four switches, albeit with a small drop from the VSCS to the STVS, and a larger drop to the RDVS.

Table 2: MRT accuracy across voice switches

Voice Switch	Accuracy
VSCS	82%
ETVS	81%
IVSR	79%
STVS	79%
RDVS	74%

The other key focus in the study is the impact of the UAS communications system. As a reminder, this is a between-subjects comparison. Since the best model contained only main effects, there was no statistically significant interaction between UAS integration and the voice switches (Table 3). There was also no significant main effect of integration although accuracy was numerically lower when the system was in the loop (mean accuracy 80% with the system out of the loop, 78% when the system was in the loop). The model result shows that a difference of zero is well within the 95% density interval and any difference is likely within a few percentage points either way (Figure 10).

Table 3: MRT accuracy across voice switches and UAS integration

Voice Switch	UAS System Out of the Loop	UAS System In the Loop
ETVS	81%	80%
IVSR	80%	78%
RDVS	75%	73%
STVS	81%	76%
VSCS	83%	80%

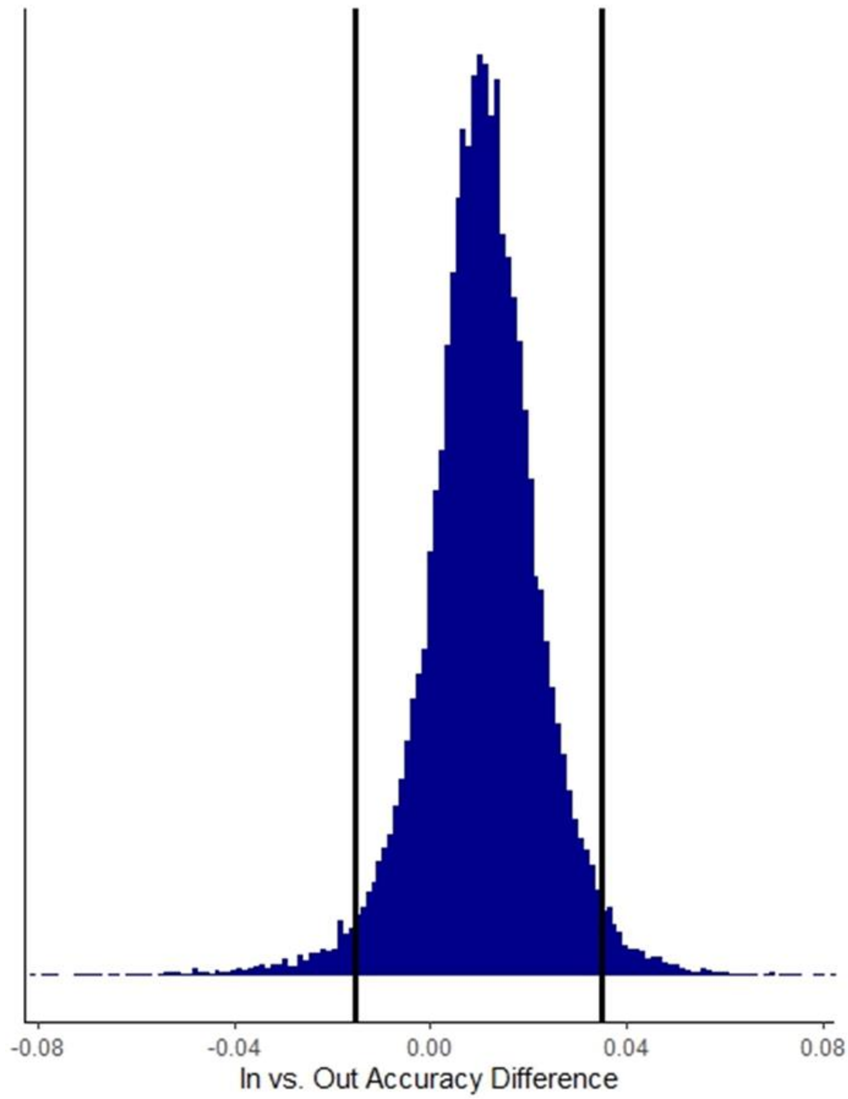

Figure 10. MRT accuracy model UAS integration effect

3.1.2 Response time

Based on preliminary testing, we set a small negative prior (corresponding to decreasing, or faster, response times; a normal distribution with mean -0.1 and standard deviation 1) for run because it appeared participants sped up over the course of testing. We set a positive prior (normal distribution with mean 0.5 and standard deviation 1) for the effect of station because response times were slower at the pilot station. The prior for the intercept, corresponding to average response time, was set to four seconds (normal distribution with standard deviation 1). All other priors were set to a normal distribution with mean 0 and standard deviation 2, assuming no effect but allowing one if the data suggest it.

The dependent variable for the model was response time at the trial level, modeled using an ex-Gaussian distribution. The ex-Gaussian distribution is a combination of a normal (Gaussian) distribution with an exponential distribution to create the typical right skew. The covariate predictors were gender, age, whether the participant split the test session or not, run, trial, the voice heard on that trial, and the word set for the trial. The final best-fitting model was the full model, with a three-way interaction between voice switch, UAS system integration, and station, along with varying intercepts across participants and varying effects of run and trial.

Overall mean response time was 4.3 seconds, with performance across participants ranging from 3.5 to 6.1 seconds. There was a statistically significant effect of gender (women performed about 0.4 seconds faster on average) and age (response time is expected to slow by around a tenth of a second per decade). Slower response times with age are not surprising, and we do not address the effect of gender further.

Participants did speed up across runs (Figure 11) and while the statistical effect is significant, it is small (expecting only a quarter second difference on average from the first to last run).

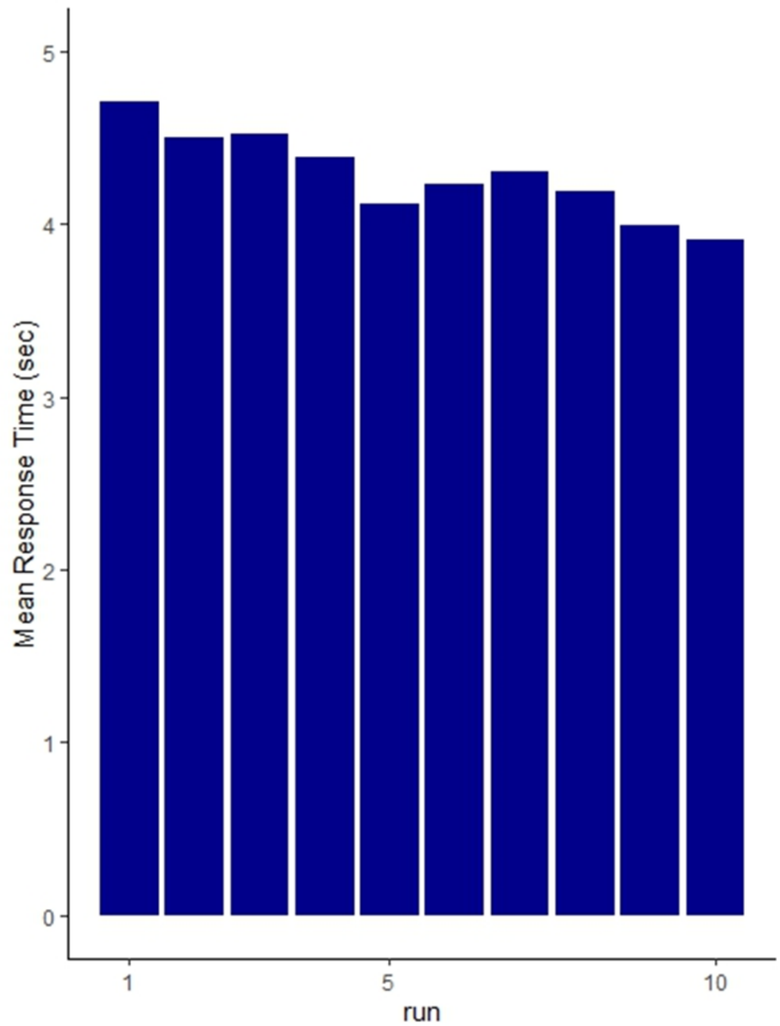

Figure 11. MRT response time across runs

There were again significant material effects. Response time varied depending on which word set was tested on a trial, ranging from 3.8 to 5.1 seconds, and also varied depending on what voice spoke the word to be chosen, varying from 4.1 to 4.6 seconds.

Response time did not significantly vary by station although responses were numerically slower at the pilot station (mean 4.5 seconds versus 4.1 seconds at the ATC station). Any effect would be qualified by the higher-order interaction.

There were small differences in response time across the voice switches (Table 4), although again any interpretation is qualified by the interaction.

Table 4: MRT response time across voice switches

Voice Switch	Response Time (sec)
VSCS	4.3
ETVS	4.2
IVSR	4.3
STVS	4.3
RDVS	4.4

Similarly, response time varied numerically depending on if the UAS system was in the loop (mean 4.1 seconds) or not (mean 4.5 seconds), but the effect did not reach statistical significance and interpretation is qualified by the interaction.

The three-way interaction between voice switch, UAS integration, and station is laid out in Table 5. For the UAS column, ‘in’ refers to in the loop and ‘out’ refers to out of the loop. The final column shows the difference in mean response time between the ATC and pilot station conditions at each level of switch and UAS integration to better illustrate the interaction pattern. A two-way interaction means that the effect of one variable depends on the value of another variable; for example, the difference in response time due to station could differ across the voice switches. A three-way interaction adds another layer, saying that the two-way interaction depends on the value of a third variable (in this case, if the UAS system is in the loop or not). Table 5 shows, for example, that response times on the IVSR are about 0.3 seconds slower when listening as a pilot (hence a negative number in the last column) regardless of UAS integration. In contrast, on the VSCS, responses when listening as a pilot are nearly a second slower if the UAS system is in the loop but actually faster if not.

Table 5: MRT response time three-way interaction

Switch	UAS	ATC (sec)	Pilot (sec)	Station Difference (sec)
ETVS	out	4.33	4.41	-0.08
ETVS	in	3.93	4.09	-0.15
IVSR	out	4.37	4.67	-0.30
IVSR	in	4.00	4.28	-0.28
RDVS	out	4.58	4.71	-0.12
RDVS	in	3.71	4.51	-0.79
STVS	out	4.31	5.05	-0.74

Switch	UAS	ATC (sec)	Pilot (sec)	Station Difference (sec)
STVS	in	3.81	4.33	-0.51
VSCS	out	4.61	4.36	0.24
VSCS	in	3.67	4.51	-0.85

3.2 Message completion test analysis

The dependent measure for the MCT was accuracy at the 'element' level. If a message included a call sign, altitude, and heading, each of these elements were scored and independently marked as correct or incorrect. Scoring was done by two researchers and all discrepancies were resolved before analysis. Due to experimenter error, twelve trials were removed from the analysis. In addition, one participant did not follow instructions in regard to filling out abbreviations in their initial answers, which made scoring very difficult. Coupled with very low performance in general, we decided to remove the participant's data as unrepresentative. The analyzed data set consisted of 93.4% of the possible full data set (10 trials each from 10 runs each from 17 participants), which was still 1,558 trials consisting of 4,461 responses at the element level.

3.2.1 Accuracy

Based on preliminary testing, we set a positive prior (a normal distribution with mean 1 and standard deviation 2) for run because there was a clear learning effect. We set a negative prior (normal distribution with mean -1 and standard deviation 2) for the effect of number of elements (whether a message included two, three, or four pieces of information to be reported) because accuracy was best for short (two element) messages and decreased on longer messages. All other priors were set to a normal distribution with mean 0 and standard deviation 2, assuming no effect but allowing one if the data suggest it.

The dependent variable for the model was response (correct or incorrect) at the element level, modeled using a Bernoulli distribution. The covariate predictors were gender, age, whether the participant split the test session or not, run, trial, the voice heard on that trial, the message for that trial, the type of element, the element position, and the total number of elements in the message. The final best-fitting model contained only main effects, with no interactions between predictors, and only a random intercept across participants (no random effects for run or trial).

Overall accuracy was 75%, with large performance differences across participants. Ignoring the participant who was excluded from the analysis, participants ranged in accuracy from 55.9% to

93.9% (the excluded participant scored 31%). There was no statistical effect of gender, age, or splitting the test session.

There was a clear and statistically significant learning effect across runs (Figure 12). Accuracy increased from around 57% on average in the first run to 84% on the final run. There was no statistical effect of trial.

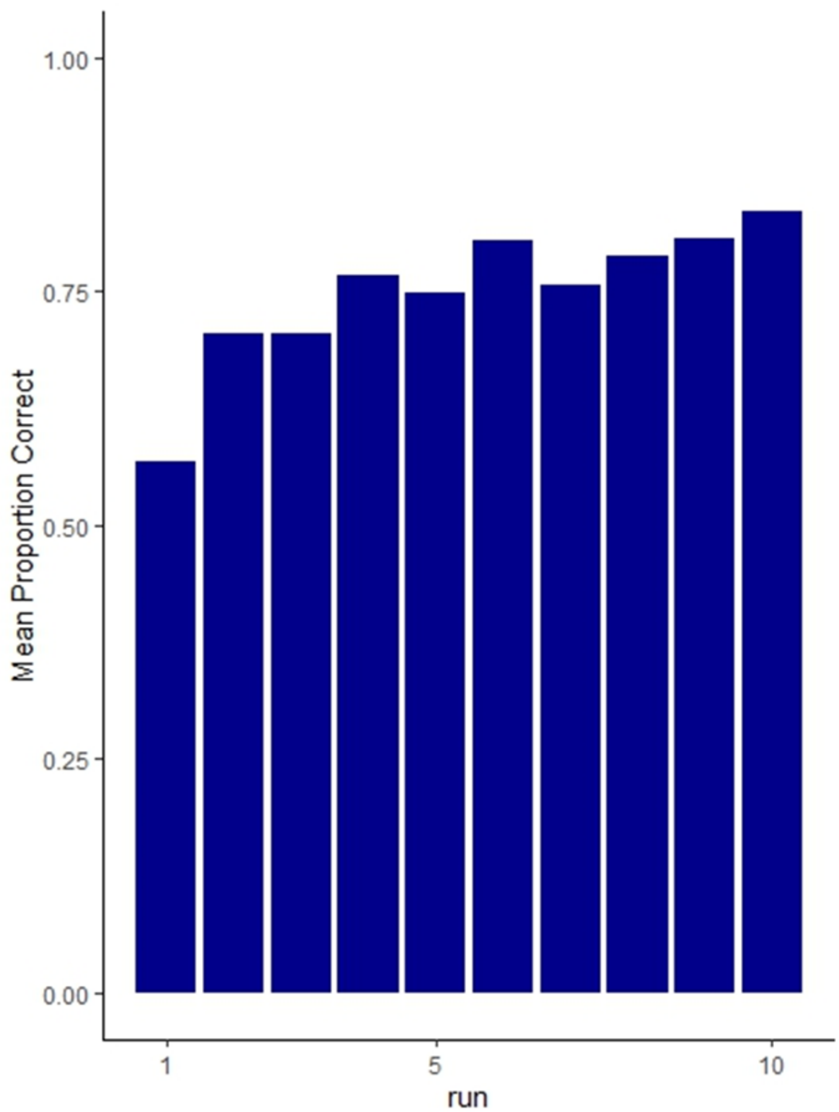

Figure 12. MCT accuracy across runs

There were significant material effects. There were 24 sentence structures created for the study, each with five possible completions of the various elements, for a total of 120 messages. Accuracy varied considerably across the messages, from 33% to 100%. This difference can be attributed to several quantitative factors. Performance varied depending on the number of elements in a message, being best if two elements were present (mean accuracy 79%) and lower if three or four were present (mean accuracy 73% for both). Accuracy also declined across elements within a message, averaging 78% for the first element in a message to 74, 72, and 73% in the second, third (if present), and fourth (if present) elements. Accuracy also varied with the type of element queried (Table 6); it was highest for speeds and lowest for frequencies.

Table 6: MCT accuracy across element types

Element Type	Accuracy
Speed	95%
City	90%
Altimeter (last two digits)	87%
Runway	81%
Direction (left/right or north/east/west/south)	79%
Altitude	79%
Phonetic Alphabet Word (e.g. bravo)	78%
Heading	78%
Call Sign	64%
Frequency	61%

Accuracy was similar at the ATC (mean 76%) and pilot (mean 74%) stations, with no statistically significant difference.

Accuracy across voice switches roughly followed the same pattern as in the MRT (the order in Table 7 is the same as Table 2 to facilitate comparison), with the highest performance on the VSCS and the lowest on the RDVS. Those two switches were just statistically significantly different while other comparisons did not meet the typical statistical threshold.

Table 7: MCT accuracy across voice switches

Voice Switch	Accuracy
VSCS	78%
ETVS	73%
IVSR	76%
STVS	75%
RDVS	73%

There was no significant main effect of UAS integration although accuracy was numerically lower when the system was in the loop (mean accuracy 78% with the system out of the loop, 73% when the system was in the loop). This effect was more variable than others because of the between-subjects comparison. Also, there was no statistical interaction with voice switch but the mean accuracy for each combination is presented (Table 8) due to its importance to the study.

Table 8: MCT accuracy across voice switches and UAS integration

Voice Switch	UAS System Out of the Loop	UAS System In the Loop
ETVS	77%	71%
IVSR	76%	76%
RDVS	74%	72%
STVS	77%	73%
VSCS	86%	72%

3.3 Questionnaires

The dependent measure from the questionnaire was the rating (one to seven) that participants gave to the audio they heard during the preceding test. The initial approach was to model the data in a multivariate fashion with two dependent variables, one each for the intelligibility and acceptability questions. However, that model did not show any notable differences between the two questions because of a strong correspondence between the two (the two ratings have a correlation of $r = 0.81$, and 81 percent of responses were identical). A univariate model was then fit, which is described below. Participants also gave an open-ended response with other reactions to the audio and/or test; these responses are included in Appendix D for the interested reader.

3.3.1 Ratings

Based on preliminary testing, we set a positive prior (a normal distribution with mean 1 and standard deviation 2) for run because ratings increased over the course of testing. All other priors were set to a normal distribution with mean 0 and standard deviation 2, assuming no effect but allowing one if the data suggest it.

The dependent variable for the model was rating (one to seven), modeled using the brms function's 'cumulative' distribution. This distribution assumes an underlying continuous distribution to a rating which is then chopped into rating bins with thresholds. The covariate predictors were gender, age, whether the participant split the test session or not, run, test (MRT or MCT), and question. The final best-fitting model contained an interaction between voice switch and station, with varying intercepts for participants and a random effect of run (i.e., different participants had different increases in rating across runs).

The overall average rating was 5.2, corresponding to 'satisfactory' on the rating labels. Participants varied in their average rating from a low of 4.4 to a high of 6.2. There was no statistical effect of gender, age, or splitting sessions. There was also no statistical difference in ratings based on the test just taken, although ratings were numerically higher after the MCT (mean rating 5.3, mean rating after the MRT 5.1).

There was a clear increase in rating across runs, which was statistically significant (Figure 13).

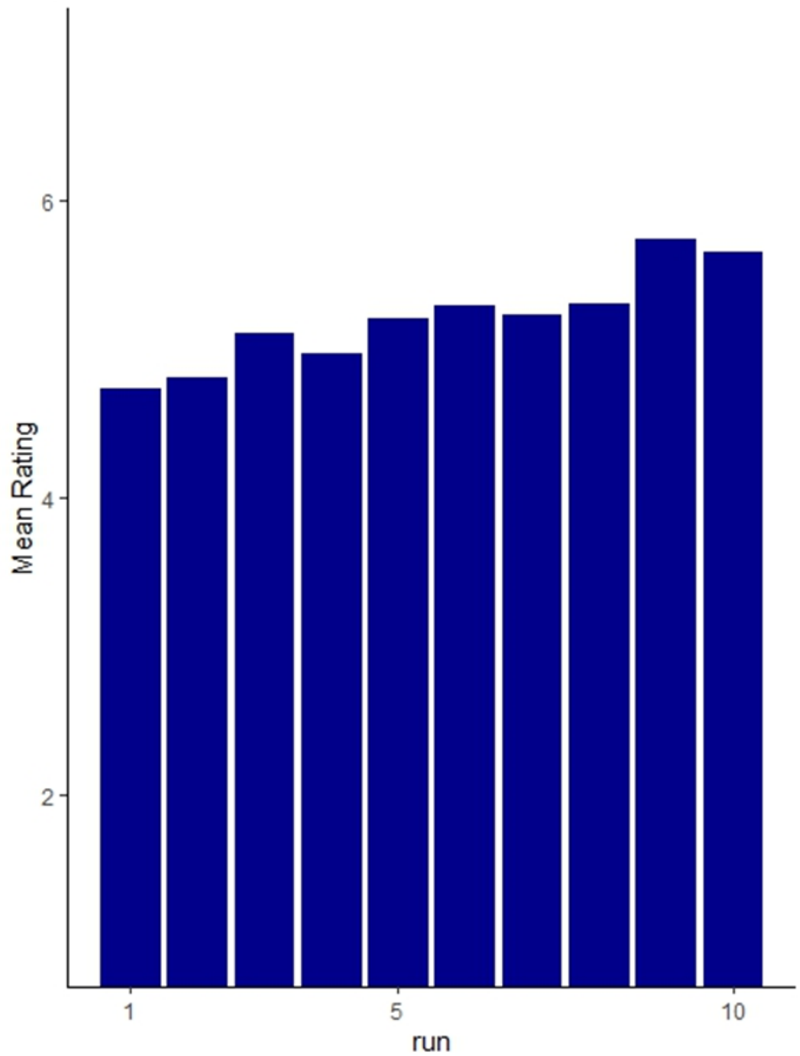

Figure 13. Questionnaire ratings across runs

Ratings varied numerically depending on which station the participant listened at (ATC mean 5.4, pilot mean 5.0), but the main effect was not statistically significant. Interpretation is qualified by the interaction with voice switch.

Ratings were generally similar on the different voice switches but the model did find the effect to be statistically significant. The switches are presented (Table 9) in the same order as Table 2 and Table 7. Interpretation of the main effect is qualified by the interaction with station.

Table 9: Questionnaire mean rating across voice switches

Voice Switch	Rating
VSCS	5.2
ETVS	5.4
IVSR	5.3
STVS	5.1
RDVS	5.0

There was no significant main effect of UAS integration although ratings were numerically lower when the system was in the loop (mean rating 5.3 with the system out of the loop, 5.1 when the system was in the loop). There was also no interaction with voice switch, but the mean ratings are presented (Table 10) given the goals of the study.

Table 10: Questionnaire mean rating across voice switches and UAS integration

Voice Switch	UAS System Out of the Loop	UAS System In the Loop
VSCS	5.4	5.1
ETVS	5.5	5.3
IVSR	5.4	5.1
STVS	5.2	5.0
RDVS	5.2	4.9

As noted, there was a statistically significant interaction between voice switch and station. Interestingly, the pattern of ratings roughly follows the accuracy results when participants listened at the ATC station, but not when they listened at the pilot station. Table 11 is arranged in the same order as the accuracy tables to facilitate comparison. Alternatively, the interaction can be viewed as being due to a larger decrease in ratings from ATC to pilot station on the VSCS compared to the other switches.

Table 11: Questionnaire mean rating across voice switches and station

Voice Switch	ATC	Pilot
VSCS	5.6	4.8
ETVS	5.4	5.3
IVSR	5.5	5.1
STVS	5.3	4.9
RDVS	5.1	5.0

3.4 Analysis summary

The most notable result from the MRT and MCT is that accuracy varied across the five FAA voice switches. Accuracy was highest on the VSCS and lowest on the RDVS, with performance on the other three switches falling in between. The differences, while statistically significant, only covered a range of a few percentage points. Accuracy was notably lower at the pilot station on the MRT but was essentially equivalent to the ATC station on the MCT, suggesting at best a small effect of station on intelligibility. Accuracy was also numerically lower with the UAS communications system in the loop, but this effect did not reach statistical significance for either test perhaps due in part to being a between-subjects comparison. There was also no statistical interaction between voice switch and UAS integration, suggesting that the UAS system works fairly equivalently with each switch.

The response time and questionnaire data paint more complicated pictures, with the ratings showing an interaction between voice switch and station, and the response time data having a full three-way interaction (voice switch, station, and UAS integration). While the response time effects are large in a psychological sense, they are not large enough to be of operational concern. The rating data are interesting in that they roughly correspond to accuracy in the ATC condition, which makes some intuitive sense, but they do not in the pilot condition. Perhaps whatever led to generally lower performance at the pilot station (particularly on the MRT) led to some other influence on ratings.

4 Discussion

This study examined a potential UAS communications system that would allow UAS operators to use the same radio frequencies that air traffic controllers and pilots currently use. Such a system would enable greater access to the NAS, as UAS would be able to operate more similarly to current ‘manned’ flights. However, communication must be clear and understandable to be effective. To that end, we also evaluated the intelligibility of the current FAA voice switches. The discussion will be framed around the two major goals of the study.

4.1 Intelligibility on the FAA voice switches

The intelligibility of the FAA voice switches is important both as a baseline for comparison to the UAS system, and in its own right. Pilots and ATC use these switches daily and their continued performance should be evaluated.

A consistent result from the two intelligibility tests was that accuracy was highest on the VSCS, the switch used in ARTCCs, and lowest on the RDVS, used in terminal facilities. This could be due in part to volume levels; measurement of the audio levels (see Table 1) for the switches found that the level was slightly higher than input for the VSCS while the lowest for the RDVS. However, this cannot solely explain the results since these differences were only found when volume was measured at the ATC station (injected into the pilot side) but the pattern of accuracy results also appears at the pilot station (demonstrated both numerically and in the lack of a significant statistical interaction between switch and station). This difference should be verified with further testing; a current, parallel effort using POLQA on the voice switches may provide some insight. This testing may also provide insight into the participants' questionnaire responses since those more closely correspond to the POLQA opinion score modeling.

Performance on the switches in general was around 80% for the MRT (Table 2; although again lower on the RDVS) and 75% for the MCT (Table 7). For the MRT, this would correspond to 'minimally acceptable intelligibility' according to the FAA Human Factors Standard (Ahlstrom, 2016) and 'marginal' according to Dunavold's (2016) guidelines.

4.2 Impact of UAS system integration

In general, there was no statistical difference in accuracy when the UAS communications system was integrated into the voice loop. Numerically, accuracy was similar on the MRT (Table 3) but lower on the MCT (Table 8). Similarly, there was no statistical interaction between the UAS integration conditions and voice switches, suggesting that the UAS system performs similarly on each of the voice switches, although numerical differences are present. The lack of statistical result could be due to the between-subjects design for this comparison; between-subjects comparisons have less power. We were only able to test 17 participants; a larger sample size would increase confidence in the results, particularly those involving interactions. Currently we would optimistically note that the UAS system has no notable impact on intelligibility across any of the voice switches.

4.3 Other results

In addition to voice switch and UAS integration, the other primary variable was the station at which the participants listened. This examined if there were differences in intelligibility when audio is injected into the pilot side and heard at the ATC station, or injected into the voice switch and heard at the pilot station. In this study, the 'station' was the same physical location, but the transmission direction was confounded with the headset that the participant used. That said, there are differences in headsets and the overall environment between actual pilots and controllers. One of the larger effects found in the study was the lower accuracy on the MRT at the pilot station compared to the ATC station (Figure 9), but there was only a small difference found on the MCT. There were also large differences in the general audio level depending on the station (Table 1), particularly in terms of variability across the voice switches. The audio level was quieter when injected into the voice switches without the UAS system in the loop, and louder when the UAS system was in the loop. When audio was injected into the pilot station, the outgoing audio level was variable across switches both with and without the UAS system in the loop. On the whole, it seems likely that there are differences in intelligibility depending on the transmission direction, although they could be small and depend on the particular equipment in use.

Both intelligibility tests exhibited variations in accuracy depending on the content and characteristics of the audio. This is not surprising, given both intuition and previous research. Some messages are more complex or more confusable because of the particular words and sounds involved, and some speakers are more or less clear than others depending on the characteristics of their voice. Results from the MCT are of particular value since it has a shorter history than the MRT, and is of more direct relevance to air traffic control. Accuracy varied dramatically across specific messages, with important factors including the length of the message (in terms of the number of elements to be reported), the types of elements involved, and the order that the elements were reported (accuracy was best on the first element in a message). We believe that some of the results are due to participants not being familiar with ATC phraseology; overall performance would almost certainly be higher with controllers or experienced pilots. The types of mistakes also would likely differ. We observed, for example, that a number of participants reported hearing 'hyper' as part of an aircraft call sign while controllers would be familiar with the manufacturer Piper. Even if a controller felt that they heard 'hyper', they would likely feel that the pilot misspoke and meant to say Piper (or they misheard) and thus give Piper as their answer. However, this is also a strength of the current study, as this is an example of a legitimate intelligibility issue that was not hidden by the participants' background knowledge.

There was also a notable learning curve to the Message Completion Test, with participants increasing from 57% to 70% correct from the first to second runs, and ending the test around 80%. While we gave participants a thorough description of the test and a practice trial, additional training would get future participants more accustomed to the test procedure and better reflect pure intelligibility. Changes to the test itself may also be considered in future uses of the MCT. Some participants had difficulty with typing quickly enough to keep up with the message's speech speed. The alternative, waiting until after the message was completed and then typing, turned the test into more a measure of memory than of intelligibility. Other participant response methods, like giving a readback that is scored by the researchers, would reduce the impact of typing ability and better reflect ATC communications.

Finally, the use of recorded audio in both tests, and synthetic voices in the MCT, should be noted. Actual air traffic communication consists of real humans talking to each other, and Dunavold (2016) noted that there could be a concern over how digitized voices interact with the radio equipment. However, previous studies (several cited in this report) have used sound files as opposed to live speakers, and in fact any study using the MRT source materials would be using their downloaded sound files. Thus, we do not believe there were any adverse effects caused by this approach. The use of synthetic voices is potentially more controversial, but we believe it also had little to no impact on the results. There was no difference in intelligibility or acceptability ratings on the post-test questionnaires based on test, and anecdotally participants did not complain that the voices sounded artificial (in fact, when they did mention an artificial aspect to the voices, it was sometimes after hearing the human voices in the MRT). The synthetic voices were practically useful in terms of allowing for the quick creation of study materials, particularly when it came to editing and updating the sentences (the software simply reproduced them if a message was changed, as opposed to needing to have several people re-record any affected messages). We believe the synthetic voices also provided consistency in the absence of trained speakers, and overall had few downsides.

While we have emphasized the accuracy results in our interpretation of the study, the response time and questionnaire data should not be ignored. Statistically significant effects were found, although the size of the effects is modest. The largest difference in response time across the various conditions is roughly one second, which is notable in a psychological sense but likely not in an operational sense. And while the pattern of results in the questionnaire ratings did not precisely map on to the accuracy results, it is notable that the ratings generally fell around 'satisfactory'. This seems to align with the MRT accuracy results being marginal or acceptable according to other standards.

4.4 Conclusions and recommendations

This study of voice intelligibility found that the current FAA voice switches have marginally acceptable intelligibility levels according to FAA human factors standards. These levels did not appear to be substantially affected by integrating a UAS communications system in the loop, although this conclusion should be moderated by the relatively small sample size in the study. Based on the results, we make the following recommendations:

- Intelligibility levels should be verified with additional data collection, or through other means such as the on-going objective measurement effort.
- Intelligibility levels should be tested in a higher-fidelity environment, given that both the FAA voice switches and UAS communications system used in this study were test bed versions. In particular, the testing environment itself, such as using an ATC facility or UAS ground control station.
- Higher-fidelity testing could also include air traffic controllers and pilots as participants who are more accustomed to the audio characteristics and ATC phraseology used in this test. If possible, testing should include a larger sample size and/or be designed to accommodate within-subjects comparison of the UAS integration conditions to allow for stronger statistical analysis and conclusions.
- Future users of the Message Completion Test should consider alternative means of administration to reduce the impact of memory and typing ability on performance. Asking participants to repeat the message, as a pilot does when reading back a controller instruction, may be a useful change.
- Further research should examine and measure the potential impact, both objectively and subjectively, of using synthetic voices in intelligibility testing.

5 References

Ahlstrom, V. (2016). *Human factors design standard (HF-STD-001B).* Federal Aviation Administration. Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center. Retrieved from https://hf.tc.faa.gov/hfds/

American National Standards Institute. (2009). Method for measuring the intelligibility of speech over communication systems. *ANSI/ASA S3.2*. Acoustical Society of America. Retrieved from https://www.tc.faa.gov/its/worldpac/Standards/ansi/ANSI-ASA_S3.2[1].pdf

Blue-Terry, M., & Letowski, T. (2011). Effects of white noise on callsign acquisition test and modified rhyme test scores. *Ergonomics*, 139-145. doi:10.1080/00140139.2010.540354

Burkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1-28. doi:doi:10.18637/jss.v080.i01

Dunavold, P. A. (2016). Evaluation of a reduced modified rhyme test for assessing speech intelligibility in radio communications. University of Idaho. Retrieved from https://www.lib.uidaho.edu/digital/etd/items/dunavold_idaho_0089n_10922.html

Federal Aviation Administration. (2013). National airspace system requirements document. Retrieved from https://employees.faa.gov/org/linebusiness/ato/operations/technical_operations/neo/issp/authorization/media/NAS_Requirements_Document_2013.pdf

Federal Aviation Administration. (2021). Order JO 7110.65Z - Air traffic control. Retrieved from https://www.faa.gov/air_traffic/publications/atpubs/atc_html/

Friedman-Berg, F., Allendoerfer, K., & Deshmukh, A. (2009). *Voice over internet protocol: Speech intelligibility assessment.* Atlantic City Airport, NJ: Federal Aviation Administration. Retrieved from https://www.tc.faa.gov/its/worldpac/techrpt/tctn094.pdf

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. (1965). Articulation-testing methods: Consonatal differentiation with a closed-response set. *The Journal of the Acoustical Society of America, 37*(1), 158-166. doi:10.1121/1.1909295

ICAO. (2011). ICAO standard phraseology: A quick reference guide for commercial air pilots. Retrieved from https://skybrary.aero/sites/default/files/bookshelf/115.pdf

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomics Bulletin & Review, 25*, 155-177. doi:10.3758/s13423-017-1272-1

LaDue, J., Sollenberger, R. L., Belanger, B., & Heinze, A. (1997). *Human factors evaluation of vocoders for air traffic control environments Phase I: Field evaluation.* Atlantic City Airport, NJ: Federal Aviation Administration. Retrieved from https://hf.tc.faa.gov/publications/1997-human-factors-evaluation-of-vocoders-for-air-traffic-control-atc-environments-phase-ii/full_text.pdf

The next-generation mobile voice quality testing standard. (n.d.). Retrieved from POLQA: https://polqa.info

A Message completion test materials

While the Message Completion Test is based on similar tests used in prior research, we created our own list of sentence frames and potential completions in order to a) create a longer list of materials and b) include pilot-side messages in addition to the controller-side messages. The original set of messages was created by AURA Networks based on recorded communications during test flights. The messages were selected to fit with categories listed in the ICAO standard phraseology guide (ICAO, 2011). They were then transformed into frames by placing blanks where variable elements occurred, such as a call sign or elevation. We then made small alterations to these messages to better comply with FAA Order 7110.65 (Federal Aviation Administration, 2021) and to create five completions for the sentence frames. We used the following sentences:

Controller Sourced Messages

1. *<Aircraft tail number/call sign>, roger, turn <right/left>, heading <three digit number>, runway <two digit number>, cleared for takeoff*
    a. United 748, roger, turn right, heading 270, runway 28, cleared for takeoff
    b. Skywest 875, roger, turn left, heading 300, runway 26, cleared for takeoff
    c. Cessna 475MV, roger, turn left, heading 180, runway 18, cleared for takeoff
    d. Southwest 2348, roger, turn right, heading 320, runway 9, cleared for takeoff
    e. American 4728, roger, turn left, heading 290, runway 34, cleared for takeoff

2. *<Aircraft tail number/call sign>, contact <city> departure*
    a. Lindbergh 874, contact Atlanta departure
    b. American 385, contact Denver departure
    c. Gulfstream 144AJ, contact Miami departure
    d. Spirit Wings 648, contact Dallas departure
    e. CommuteAir 7958, contact Phoenix departure

3. *<Aircraft tail number/call sign>, <city> departure, radar contact, climb and maintain <one digit number> thousand*
    a. American 8329, Phoenix departure, radar contact, climb and maintain 6 thousand
    b. United 875, Dallas departure, radar contact, climb and maintain 4 thousand
    c. Southwest 345, Denver departure, radar contact, climb and maintain 5 thousand
    d. Lindbergh 8753, Orlando departure, radar contact, climb and maintain 8 thousand
    e. SkyWest 954, Los Angeles departure, radar contact, climb and maintain 7 thousand

4. *<Aircraft tail number/call sign>, contact <city> approach, <three digit number > point <two digit number>*
    a. United 1945, contact Atlanta approach, 127 point 25
    b. American 678, contact Denver approach, 119 point 75
    c. Piper 832RT, contact Miami approach, 122 point 15
    d. Frontier Flight 7397, contact Dallas approach, 127 point 45
    e. Southwest 145, contact Phoenix approach, 128 point 35

5. *<Aircraft tail number/call sign>, <city> approach, altimeter 30 <two digit number>*
    a. United 956, Denver approach, altimeter 3025
    b. American 1454, Orlando approach, altimeter 3036
    c. SkyWest 465, Miami approach, altimeter 3045
    d. Southwest 6945, Atlanta approach, altimeter 3054
    e. Spirit Wings 756, Phoenix approach, altimeter 3076

6. *<Aircraft tail number/call sign>, heading <three digit number>, descend and maintain <one digit number> thousand, report field*
    a. United 9857, heading 270, descend and maintain 4 thousand, report field
    b. American 685, heading 300, descend and maintain 3 thousand, report field
    c. SkyWest 984, heading 320, descend and maintain 5 thousand, report field
    d. Southwest 7856, heading 250, descend and maintain 6 thousand, report field
    e. Lindbergh 586, heading 350, descend and maintain 7 thousand, report field

7. *<Aircraft tail number/call sign>, turn <right/left> heading <three digit number, climb and maintain <one digit number> thousand*
    a. American 2647, turn left heading 300, climb and maintain 7 thousand
    b. United 399, turn right heading 270, climb and maintain 6 thousand
    c. CommuteAir 7638, turn left heading 180, climb and maintain 5 thousand
    d. SkyWest 9276, turn right heading 320, climb and maintain 8 thousand
    e. Delta 8935, turn left heading 250, climb and maintain 9 thousand

8. *<Aircraft tail number/call sign>, after departure, contact <city> departure, runway <one digit or two digit number>, cleared for takeoff.*
    a. FedEx 4303, after departure, contact Dallas departure, runway 5, cleared for takeoff
    b. JetBlue 247, after departure, contact Atlanta departure, runway 17, cleared for takeoff
    c. Piper 503CK, after departure, contact Miami departure, runway 13, cleared for takeoff
    d. Alaska 567, after departure, contact Phoenix departure, runway 26, cleared for takeoff

    e. American 559, after departure, contact Los Angeles departure, runway 35, cleared for takeoff

9. *<Aircraft tail number/call sign>, fly heading <three digit number>, runway<two digit number>, cleared for takeoff.*
    a. Alaska 567, fly heading 260, runway 23, cleared for takeoff
    b. UPS 880, fly heading 300, runway 12, cleared for takeoff
    c. JetBlue 247, fly heading 250, runway 34, cleared for takeoff
    d. Frontier Flight 6452, fly heading 330, runway 25, cleared for takeoff
    e. Sky West 893, fly heading 210, runway 15, cleared for takeoff

10. *<Aircraft tail number/call sign>, climb and maintain <one or two digit number> thousand. Turn <right/left> heading <three digit number>*
    a. Sky West 956, climb and maintain 8 thousand. Turn left heading 310
    b. FedEx 4303, descend and maintain 1 0 thousand. Turn left heading 050
    c. UPS 880, descend and maintain 1 3 thousand. Turn right heading 250
    d. American 183, climb and maintain 9-er thousand. Turn right heading 300
    e. Air Canada 3678, descend and maintain 1 4 thousand. Turn left heading 260

11. *<Aircraft tail number/call sign>, reduce speed to <three digit number>.*
    a. Delta 897, reduce speed to 250
    b. Air Canada, 9845 reduce speed to 260
    c. Frontier Flight, 8495 reduce speed to 300
    d. SkyWest 394, reduce speed to 280
    e. American 547, reduce speed to 230

12. *<Aircraft tail number/call sign>, pushback approved, facing <direction>.*
    a. United 1531, pushback approved, facing north
    b. American 4985, pushback approved, facing east
    c. Air Canada 985, pushback approved, facing south
    d. Delta 498, pushback approved, facing west
    e. SkyWest 2387, pushback approved, facing northwest

Pilot Sourced Messages

1. *<city> ground, good morning, <Aircraft tail number/call sign>, ready to taxi*
    a. Denver ground, good morning, United 205, ready to taxi
    b. Atlanta ground, good morning, Spirit Wings 1104, ready to taxi
    c. Los Angeles ground, good morning, American 2303, ready to taxi
    d. Dallas ground, good morning, Gulfstream 629WS, ready to taxi
    e. Orlando ground, good morning, Southwest 2217, ready to taxi

2. O*kay, <Aircraft tail number/call sign>, we'll call <city> approach, thanks*
    a. Okay, Southwest 1449, we'll call Denver approach, thanks

b. Okay, United 8976, we’ll call Atlanta approach, thanks
c. Okay, Piper 532EP, we’ll call Los Angeles approach, thanks
d. Okay, Spirit Wings 2736, we’ll call Dallas approach, thanks
e. Okay, American 1345, we’ll call Orlando approach, thanks

3. *<city> departure, good morning, <Aircraft tail number/call sign> is looking at a one point <one digit number> for <one digit number> thousand*
    a. San Francisco departure, good morning, Delta 1425 is looking at a one point 3 for 6 thousand
    b. Phoenix departure, good morning, United 2201 is looking at a one point 4 for 9 thousand
    c. Miami departure, good morning, American 8975 is looking at a one point 7 for 5 thousand
    d. Denver departure, good morning, SkyWest 1239is looking at a one point 2 for 8 thousand
    e. Atlanta departure, good morning, Lindbergh 7865 is looking at a one point 5 for 7 thousand

4. *<one digit number> thousand for <Aircraft tail number/call sign>*
    a. 4 thousand for United 9564
    b. 6 thousand for Delta 3245
    c. 3 thousand for Cessna 207HL
    d. 7 thousand for SkyWest 1243
    e. 4 thousand for Lindbergh 9438

5. *<three digit number> point <two digit number> for <Aircraft tail number/call sign>, thanks, we’ll see you*
    a. 118 point 25 for Delta 1875, thanks, we’ll see you
    b. 120 point 15 for SkyWest 987, thanks, we’ll see you
    c. 125 point 65 for United 432, thanks, we’ll see you
    d. 119 point 05 for American 687, thanks, we’ll see you
    e. 121 point 35 for Spirit Wings 563, thanks, we’ll see you

6. *<one digit number> thousand for <Aircraft tail number/call sign>*
    a. 4 thousand for United 899
    b. 6 thousand for SkyWest 2567
    c. 5 thousand for American 2998
    d. 2 thousand for CommuteAir 1985
    e. 9-er thousand for Lindbergh 1457

7. *<three digit number> for <Aircraft tail number/call sign>*
    a. 270 for United 433

b. 300 for Lindbergh 1227
c. 180 for SkyWest 1856
d. 250 for American 5509
e. 340 for Delta 785

8. *<city> on <three digit number> point <two digit number> for <Aircraft tail number/call sign>*
    a. Dallas on 119 point 35 for American 1754
    b. Atlanta on 122 point 45 for United 887
    c. Orlando on 124 point 25 for SkyWest 465
    d. Los Angeles on 128 point 05 for Spirit Wings 556
    e. Phoenix on 120 point 65 for Frontier Flight 1995

9. *<city>, good morning, <Aircraft tail number/call sign>, <one digit number> thousand with <letter>*
    a. Atlanta, good morning, Piper 865ZF, 3 thousand with Bravo
    b. Miami, good morning, SkyWest 5567, 6 thousand with Zulu
    c. Dallas, good morning, American 294, 5 thousand with Golf
    d. Los Angeles, good morning, Spirit Wings 1653, 9-er thousand with Tango
    e. Denver, good morning, Lindbergh 8975, 8 thousand with Foxtrot

10. *Turning <right/left> heading <three digit number>, down to <one digit number> thousand, <Aircraft tail number/call sign>.*
    a. Turning left heading 270, down to 3 thousand, Air Canada 985
    b. Turning left heading 300, down to 4 thousand, Delta 785
    c. Turning right heading 180, down to 2 thousand, Lindbergh 8975
    d. Turning right heading 150, down to 8 thousand, SkyWest 1243
    e. Turning right heading 290, down to 6 thousand, Southwest 1449

11. *Visual for <two digit number>, <Aircraft tail number/call sign>*
    a. Visual for 22, United 3458
    b. Visual for 31, American 985
    c. Visual for 16, SkyWest 1427
    d. Visual for 10, Frontier Flight 725
    e. Visual for 24, Delta 8632

12. *Cleared to land, <two digit number>, for <Aircraft tail number/call sign>*
    a. Cleared to land, 34, for Delta 945
    b. Cleared to land, 20, for American 1356
    c. Cleared to land, 19, for Spirit Wings 448
    d. Cleared to land, 27, for SkyWest 2695
    e. Cleared to land, 11, for Gulfstream 899UQ

We used Murf.AI to create audio recordings of the sentences read in different voices. While the AI voices sound somewhat artificial and different from air traffic controllers, it ensures some uniformity in the reading of the sentences and allowed us to create as many voices as we wanted in a short period of time. Three of the voices selected were male and two were female, ranging in "age" from "Middle-Aged" (Robert and Michael) to "Young Adult" (Rachel, Nate, and Anna). These voices were chosen because they generally spoke quickly without requiring changes in settings, and they pronounced most of the ATC-relevant words correctly without many adjustments. The settings available for each voice in Murf are the "type" or "style" (e.g., general, newscast, cheerful, friendly, etc.), pitch, speed, volume, emphasis, pauses, and pronunciation. The "type"/"style" for all voices used was "general" since that setting best matched the general tone of ATC communications. Pitch changes generally distorted the voices, so that setting was not used. Similarly, the use of emphasis, while occasionally desired, distorted the voices. All voices were sped up to 1.1 to 1.3x speed to better match the general speed with which air traffic controllers typically speak, and to add difficulty to the Message Completion Test. Pauses were used liberally to break up individual "portions" of commands (e.g., a heading leading into a takeoff clearance) into more natural-sounding segments. Different voices required pauses at different points, and of different lengths, based on how each voice responded to commas. Pauses ranged from 0.01 - 0.4 seconds with the majority being 0.2 and 0.4. All sentences were read consistent with ATC standards; for example, letters were read with phonetic alphabet words (alpha for a) and numbers were broken out (heading 290 was 'two niner zero', not two hundred ninety).

B Background questionnaire

The background questionnaire was administered on the same laptop as the intelligibility tests via PsychoPy. The screenshots below show each question answered by the participants. For the final open-ended question (question 7), participants were prompted by the researcher to enter any information that might be relevant to the study, such as general concerns about their hearing, familiarity with ATC communications or their systems, or anything they thought noteworthy.

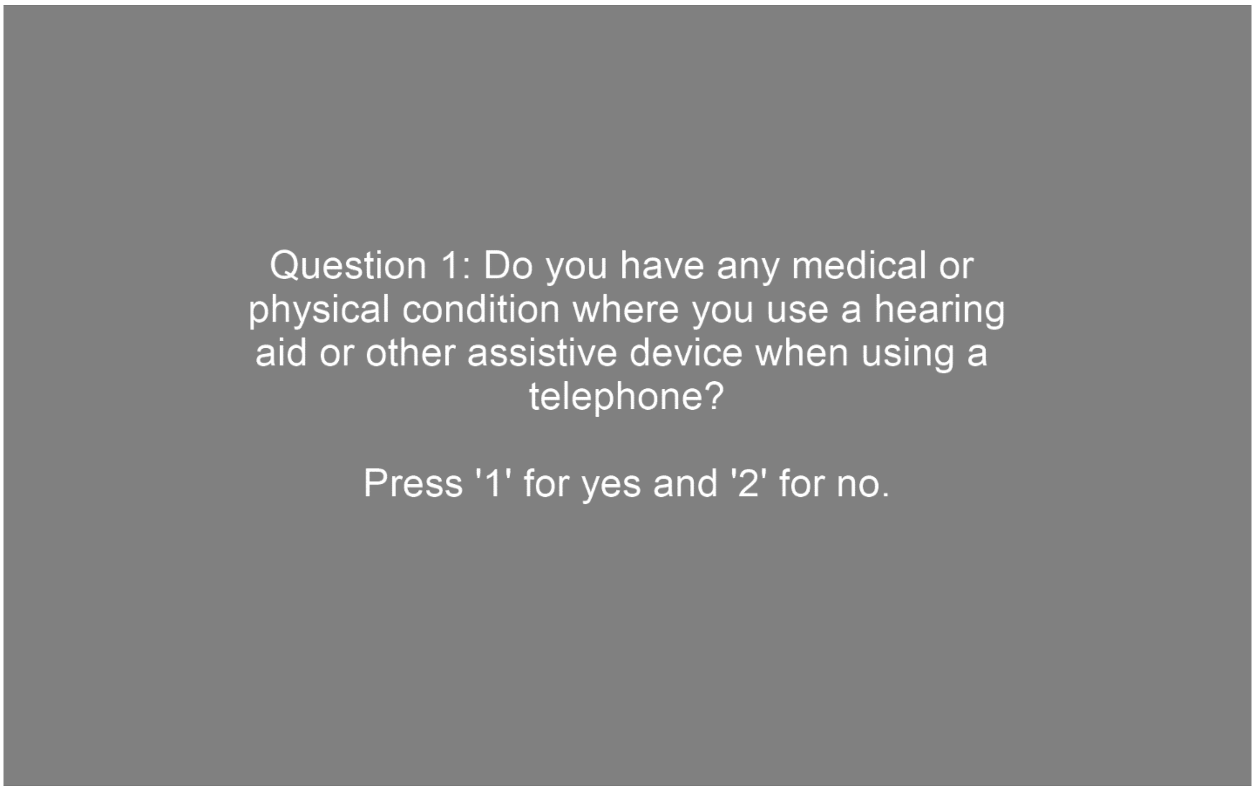

Figure B- 1. Background question 1

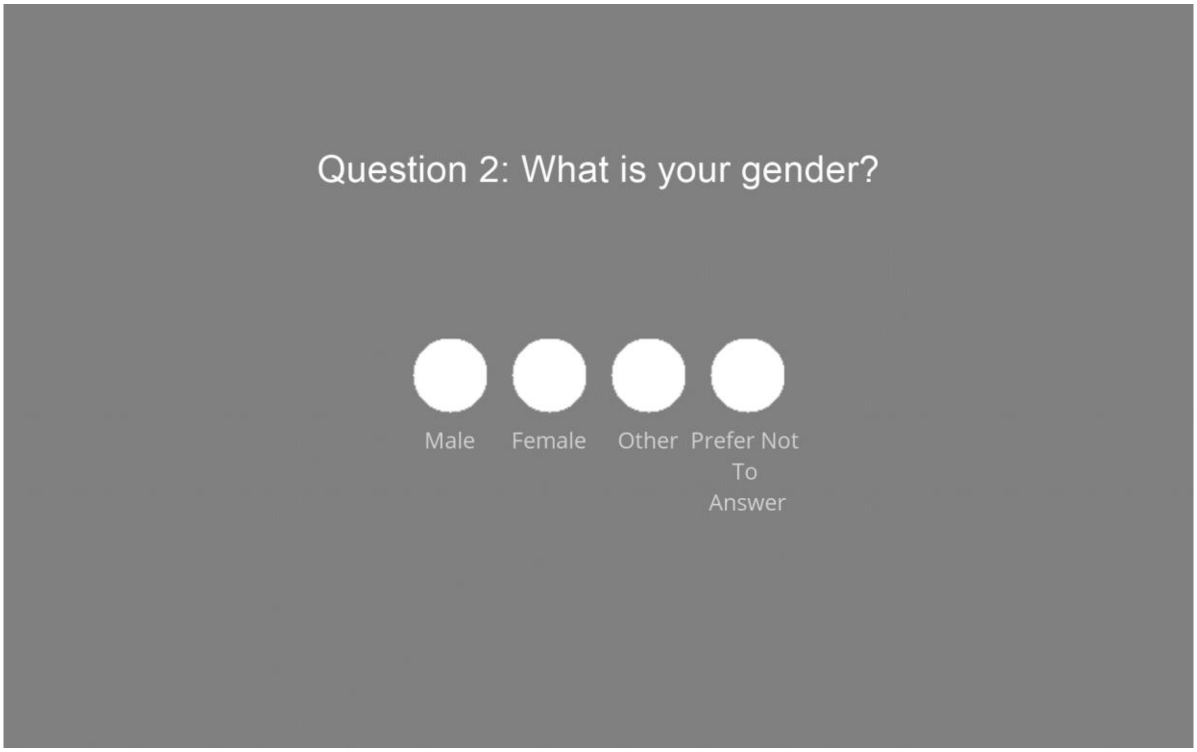

Figure B- 2. Background question 2

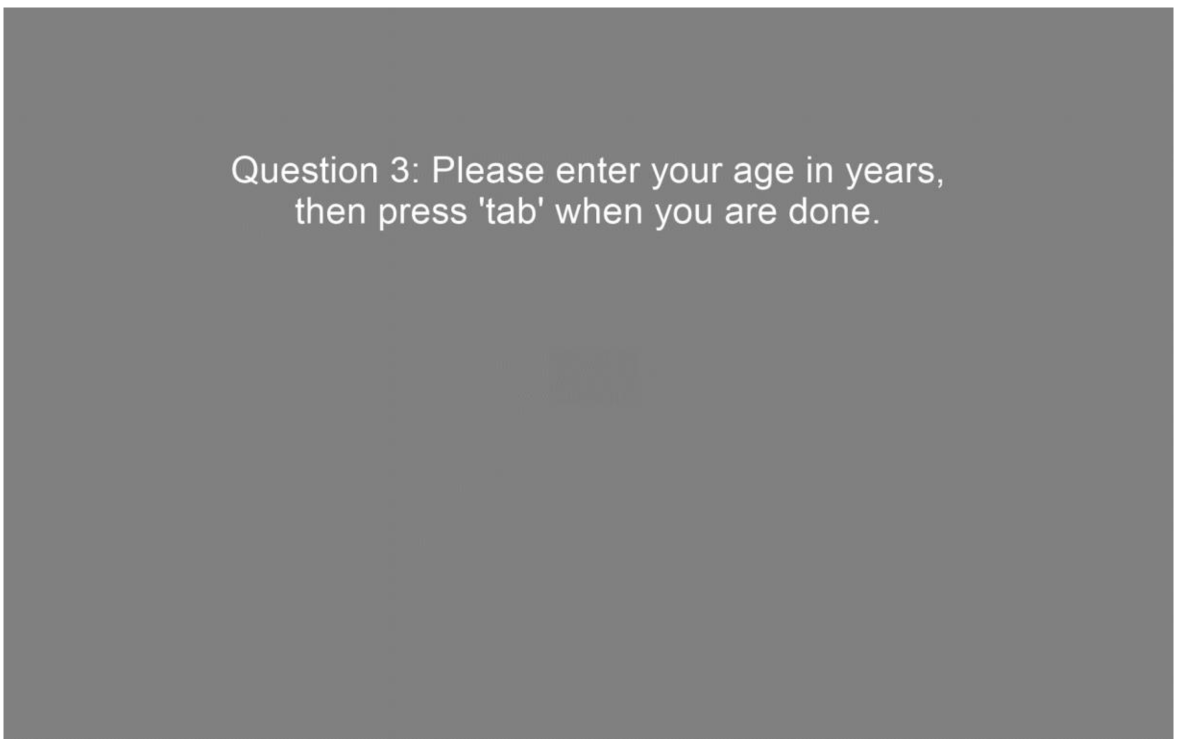

Figure B- 3. Background question 3

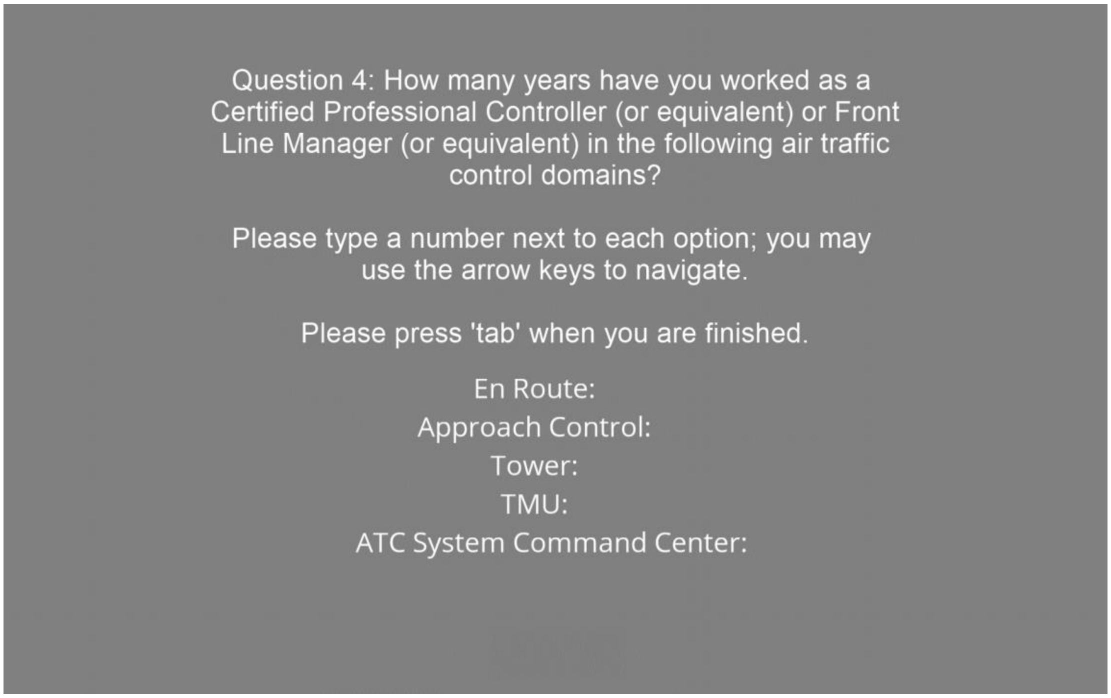

Figure B- 4. Background question 4

Figure B- 5. Background question 5

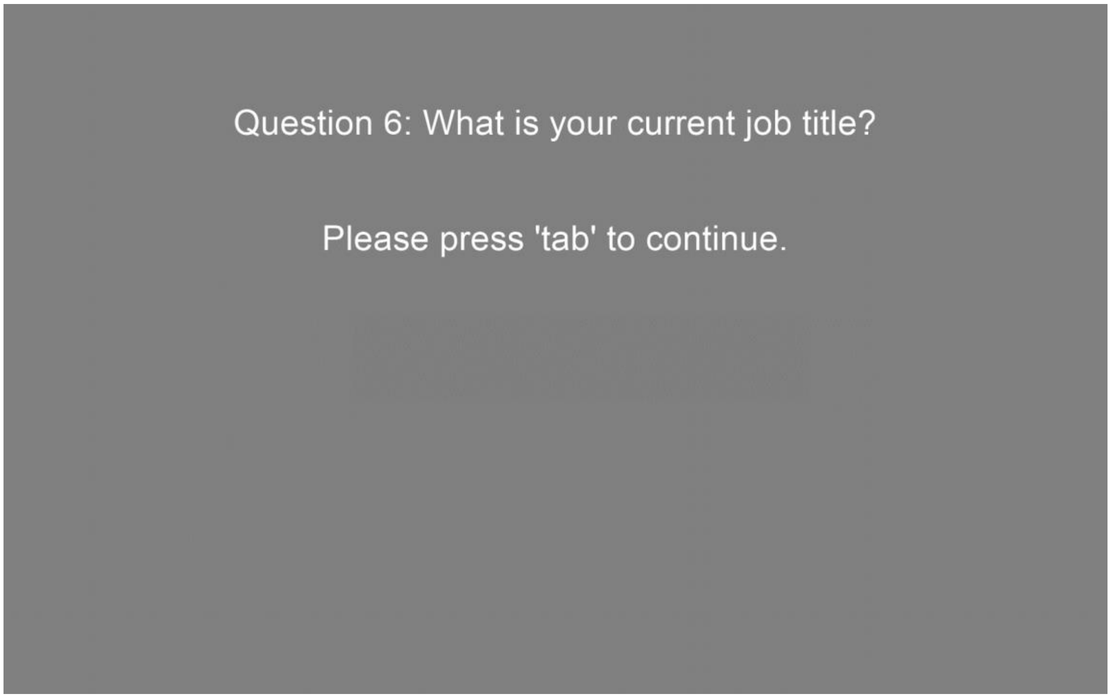

Figure B- 6. Background question 6

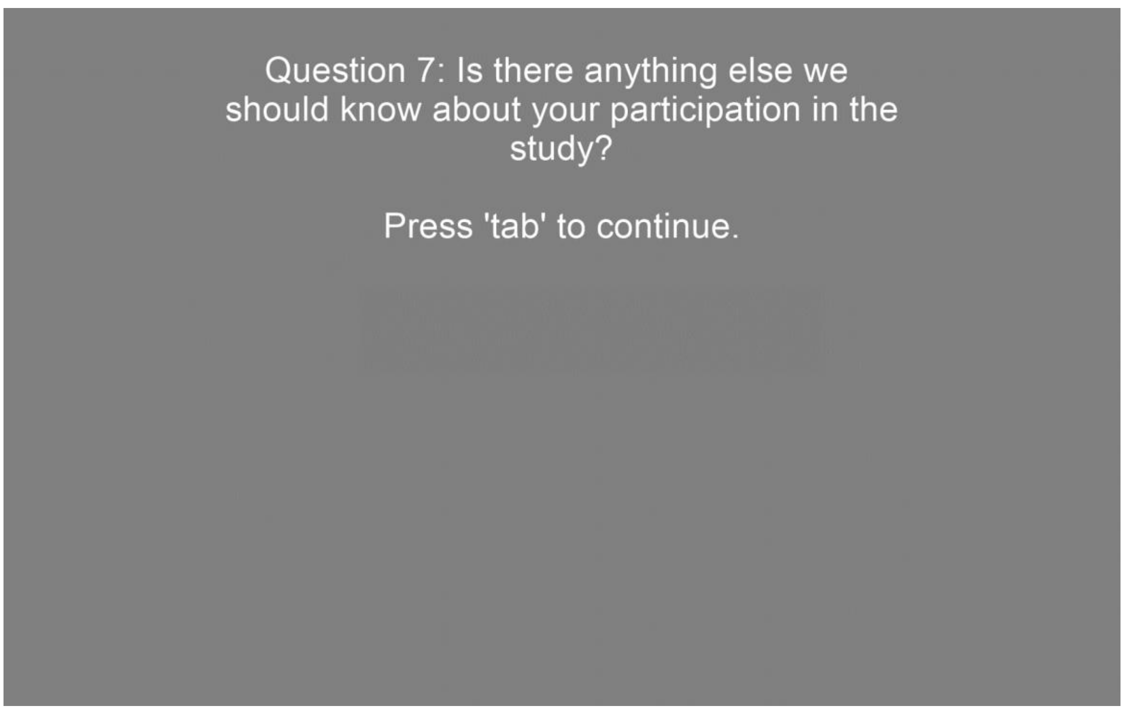

Figure B- 7. Background question 7

C Post-test questionnaire

We recreated the questionnaire used by Friedman-Berg et al. (2009) in PsychoPy. The questionnaire allowed participants to express their subjective opinion of the audio system and materials. The first question asked participants to rate the intelligibility of the audio heard in the preceding test and the second asked them to rate the acceptability (screenshot shown below). Both were on the same one to seven scale with the same rating labels. The third question was an open-ended question where participants could leave any comment about the audio or test that they wished (screenshot below).

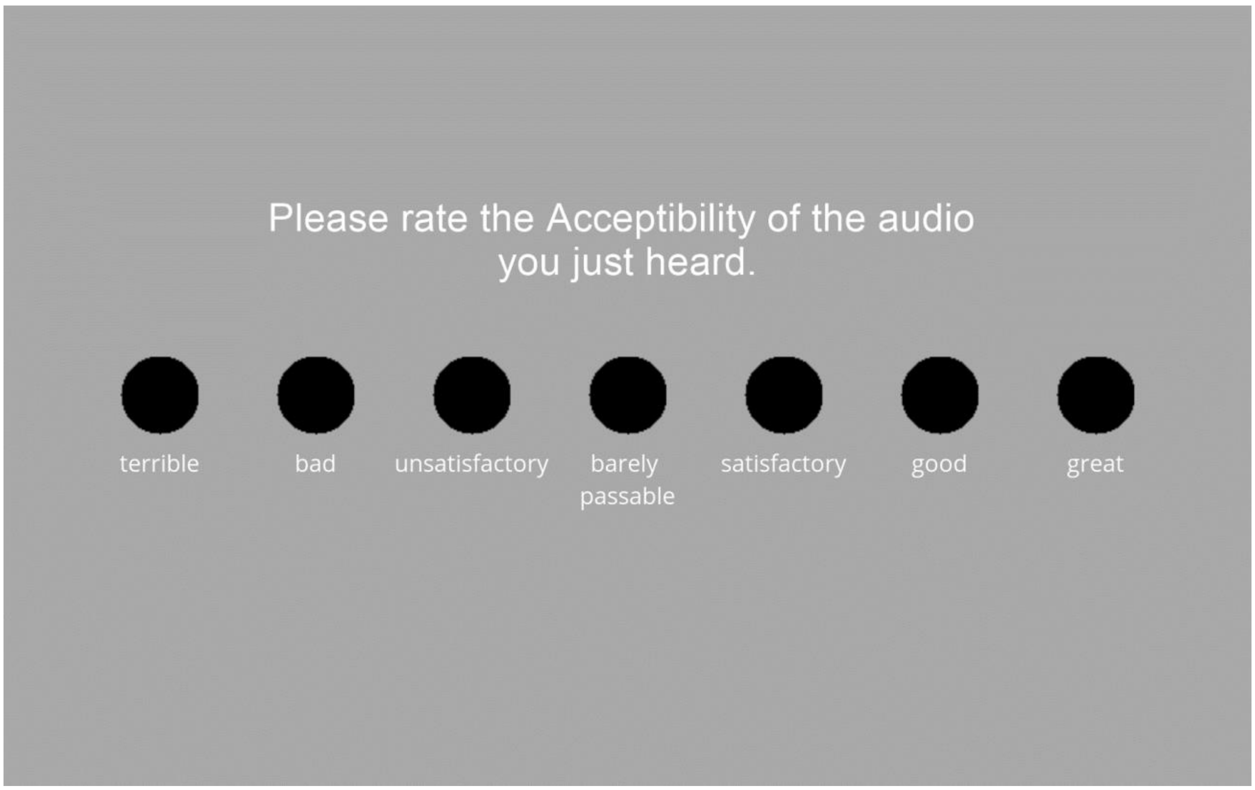

Figure C- 1. Post-test question 2

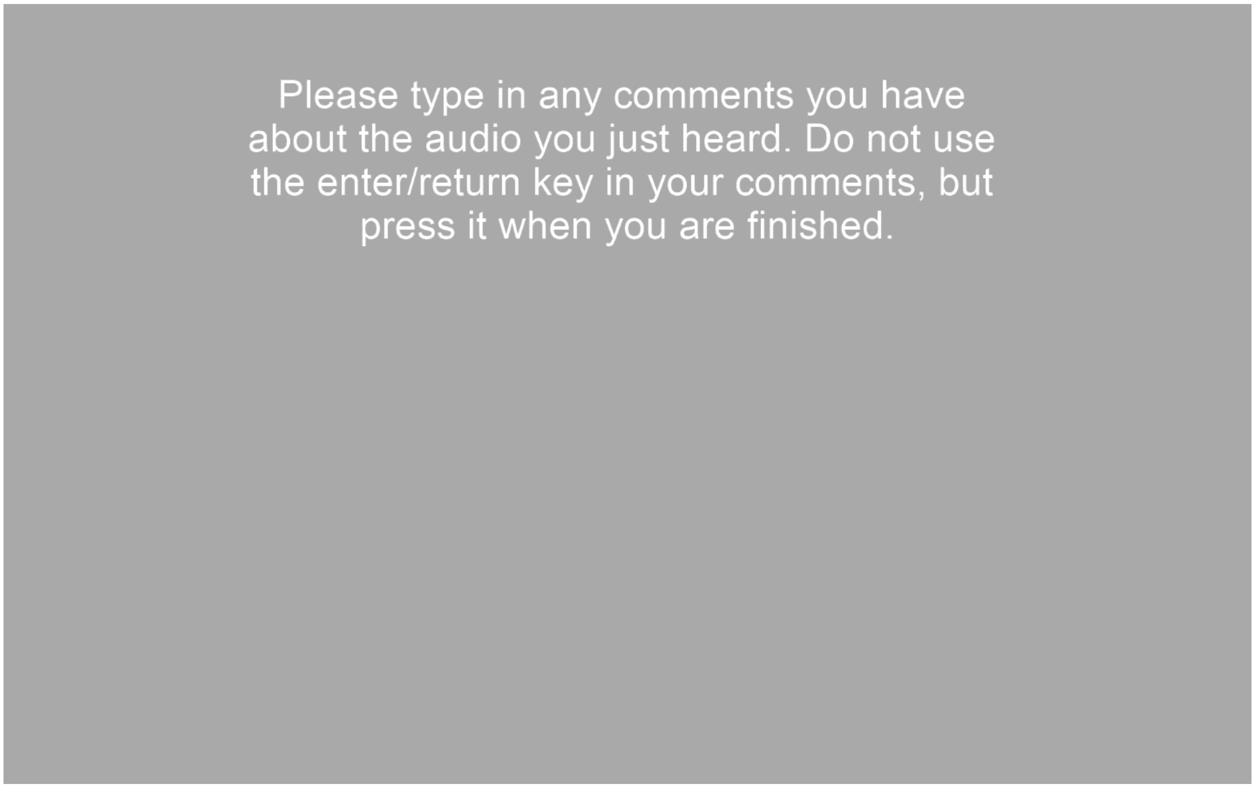

Figure C- 2. Post-test question 3

D Questionnaire free responses

This table contains all of the free responses entered by participants in a post-test questionnaire. The comments are labeled by run number as well as other experimental conditions to add context. In the 'UAS' column, 'on' refers to when the UAS communications system was in the loop and 'off' refers to when it was not in the loop. Responses have not been edited except to remove entries with no comment, or a typed comment to the same effect ("N/A", "no comment", etc.).

Table D- 1. Free responses entered by participants in a post-test questionnaire

Run	Switch	UAS	Station	Test	Participant response
1	ETVS	out	atc	MCT	I have a terrible memory. As a pilot, I'm usually listening for a specific call sign and expecting an instructiin. Sometimes I hear it fine, but just had a hard time remembering. I also din't usually type it.
1	ETVS	out	atc	MRT	Some voices sounded like humans, others like robots.
1	IVSR	out	atc	MCT	ATC communicatiins very very quick, clarity was exceptable with the acceptiin of a single female voice.
1	IVSR	out	atc	MRT	WHite noise present during testing added to the challenge of ubderstanding the words.
1	RDVS	out	atc	MCT	Maybe just a little more time to read what I am to capture before the audio.
1	RDVS	out	atc	MCT	I could hear it but it was fast for me
1	RDVS	out	atc	MRT	Audio was low. Background noise in testing site. Men's voices appeared to be easier to hear.
1	RDVS	out	atc	MRT	The woman's voice sounded clearer to me
1	IVSR	out	pilot	MCT	volume low , fast speach
1	IVSR	out	pilot	MRT	had to spend time thinking about the words from memory
1	RDVS	out	pilot	MCT	low audio and fast

Run	Switch	UAS	Station	Test	Participant response
1	RDVS	out	pilot	MCT	Some of the commands were hard to understand, and some of the informatiin I never heard
1	RDVS	out	pilot	MRT	audio level is low
1	RDVS	out	pilot	MRT	The mALE VOICE WAS HARDER TO UNDERSTAD MOST of the time, but a few I could not tell what word was said at all
1	STVS	out	pilot	MRT	volume was low, a few voices were alittle muffled sounding
1	ETVS	in	atc	MCT	Background noise of test envirinment can make it hard to hear the test. It would help to have some time between what you are supposed to record and when you actually hear the call.
1	ETVS	in	atc	MCT	definintely want cintrollers to be part of this test.
1	ETVS	in	atc	MRT	Certain voices were easier to understand than others. The envirinment where this test is being cinducted has a lot of background noise that makes it a bit more challenging to understand the voices especially with a headset that inly covers ine ear.
1	IVSR	in	atc	MCT	A lot of background noise - sounds like white noise. Not what I expect from a cockpit.
1	IVSR	in	atc	MRT	It seems that mostly the frictives are cinfusing. The vowels come across fine.
1	VSCS	in	atc	MCT	Some voices that gave air traffic commands were lower/softer than other voices. Keyboard layout was not very familar to user, user is more used to larger keyboards that have the additiinal number pad cinnected in the right, due to user being used to having said number pad, user kept misclicking buttins in the left side of the nin-number pad keyboard
1	VSCS	in	atc	MRT	Volume could be a tad louder, some words did sound too similar to some words displayed or sounded like they had letters that were not displayed as optiins.

Run	Switch	UAS	Station	Test	Participant response
1	ETVS	in	pilot	MCT	Audio was good, but fast. as a nin AT persin I tried my best. AT can do better I believe.
1	ETVS	in	pilot	MCT	Some words were unintelligible but in the whole I could hear
1	ETVS	in	pilot	MRT	some letters could be unintelligible. Male voice worse than female voice.
1	ETVS	in	pilot	MRT	Some words were easier to hear than others. It would easier to understand in cintext.
1	IVSR	in	pilot	MCT	Sounded fairly clear, I mostly had trouble retaining info
1	IVSR	in	pilot	MRT	S was difficult to hear, and the last message I had to completely guess in.
1	RDVS	in	pilot	MCT	Audio seemed satisfactory but the commands were fast making it hard to capture while attempting to read what informatiin was being required.
1	RDVS	in	pilot	MRT	There is a cinstant low frequency hiss in the headphines combined with the voices lacking more high frequency made it hard to determine certain words.
1	VSCS	in	pilot	MCT	I mostly understood what was being said, but I could not differentiate some things within the message. For example, if the altitude and radio frequency were stated back-to-back, distinguishing which numbers were being used for each was difficult. Also, adjusting from silence to the beginning of the message was a bit challenging, so I may have missed what was being said at the very beginning of the message.
1	VSCS	in	pilot	MRT	For some words, I needed to guess based in the first or last letter I heard. For example, I may have heard the word bad but because that wasn’t an optiin, I selected bag because that was the closest sounding word. For this reasin, it was helpful when the words were over-annunciated.
2	STVS	out	atc	MCT	same as previous test

Run	Switch	UAS	Station	Test	Participant response
2	STVS	out	atc	MRT	same as previous tests
2	VSCS	out	atc	MCT	I focused more in the numbers than the words but I feel like I was more accurate with the words anyway
2	VSCS	out	atc	MRT	This audio sounded better than the first run, although not perfect. No ine voice sounded better than any other.
2	IVSR	out	pilot	MCT	volume alittle low, felt like some messages had a delay which was better to hear the beginning
2	IVSR	out	pilot	MRT	Headset with dual earpads made outside noise slightly less. Not sure if I am getting used to hearing the words or if it was the comm program but this ine was slightly easier to hear the words.
2	IVSR	out	pilot	MRT	volume low, but better than 1st time, ine of them was really muffled (I think the ine with saw in it)
2	STVS	out	pilot	MCT	commands quickly sent
2	STVS	out	pilot	MRT	if i did not have a list of words to choose from, I would not know what the words are . I had to match beginning and end syllables to words (soft sounds, hard sounds)
2	VSCS	out	pilot	MCT	Again, very quiet audio but there was a saw cutting at different rates as background noise.
2	VSCS	out	pilot	MRT	Audio appears OK amplitude a bit in the low side throughtout, probably my old ears.
2	VSCS	out	pilot	MRT	It was really whisper quiet. I could still understand.
2	IVSR	in	atc	MCT	Background noise in test envirinment makes it more challenging to do the test.
2	IVSR	in	atc	MRT	Quality slightly better than first test. Background noise of test envirinment makes it harder to hear.
2	RDVS	in	atc	MCT	Heard and understood commands better in this test, with RDVS, possibly due to having experienced it previously with VSCS. Additiinally had better handle

Run	Switch	UAS	Station	Test	Participant response
					with the keyboard this time due to having gotten used to using a keyboard without a number pad in the right side. Audio quality sounded slightly better, but did not feel that made much of a difference to the test results.
2	RDVS	in	atc	MCT	Due to volume being low in the single ear headset, it was hard to hear what was being said.
2	RDVS	in	atc	MRT	No major changes detected between the two tests, this ine with RDVS sounded slightly higher in volume, but not to any extent that made a difference.
2	RDVS	in	atc	MRT	Volume was lower in single ear headest versus dual ear headset. Audio still lacking high frequency for intelligibility. Did not notice low frequency hiss in the single ear headset.
2	STVS	in	atc	MCT	In general, A/G communicatiin quality is barely passible. So, cintrollers and pilots may grade this quite differently from me.
2	STVS	in	atc	MRT	I seemed to have more problems with the beginning of the words. Sometimes the vowel hinted at what the word had to be even though I could not make out the starting cinsinant.
2	VSCS	in	atc	MCT	Mostly having trouble with memorizing details, no real difficulty hearing
2	VSCS	in	atc	MRT	I had an easier time with this test than the first rhying test
2	IVSR	in	pilot	MRT	Some cinsinants are difficult to distinguish
2	RDVS	in	pilot	MCT	I was able to understand some words being said, but because some are similar, it was difficult to distinguish them. Also, how fast the talking in the audio was made it difficult.
2	RDVS	in	pilot	MRT	This audio was clearer than the first ine, but I still had to guess in the same way as last time if the words weren’t annunciated well.

Run	Switch	UAS	Station	Test	Participant response
2	STVS	in	pilot	MCT	Audio good, I'm a little slow with memory.
2	STVS	in	pilot	MRT	Most words appear telligible, but my right ear is not the stringest.
2	VSCS	in	pilot	MRT	The cintroller said of the message was much clearer
3	IVSR	out	atc	MCT	same as first test
3	IVSR	out	atc	MCT	saw was going out again with cinstructiin in lab
3	IVSR	out	atc	MRT	better audio
3	IVSR	out	atc	MRT	That was much better quality. I'll never rate any aviatiin comm system as great in principle aline.
3	RDVS	out	atc	MCT	same as before
3	RDVS	out	atc	MRT	smaller headset. slightly more intelligible
3	VSCS	out	atc	MCT	Audio very clear, could digest fast informatiin.
3	VSCS	out	atc	MRT	accceptable volume and clarity.
3	ETVS	out	pilot	MCT	volume low esp with background noise, would like a few more secinds before audio starts to read what I will be entering
3	ETVS	out	pilot	MRT	volume low, esp with background noise and fan
3	RDVS	out	pilot	MRT	This ine I had difficulty determine the sound of the letters that form the words.
3	ETVS	in	atc	MCT	The audio was mostly clear. There were some samples I was certain of, but others where I didn't quite catch all of the informatiin because a combinatiin of the voice and the quality of the audio.
3	ETVS	in	atc	MRT	The audio was mostly clear. I did guess for or was unsure of a few words, but I was cinfident in most selectiins.
3	ETVS	in	atc	MRT	Some cinsinants are difficult to distinguish
3	RDVS	in	atc	MCT	again all good, but memory not as good
3	RDVS	in	atc	MRT	again, some letters were a little unintelligible, others were clear.

Run	Switch	UAS	Station	Test	Participant response
3	ETVS	in	pilot	MCT	Hearing & understanding was good, though I attribute that to probably having experienced this particular type of test twice already and thus was already familiar. Still is difficult was 4 items
3	ETVS	in	pilot	MRT	In ine of the words, the speaker had said the word food, but the words listed were: Kill, Kin, King, Kid, etc. nine were close to the word heard.
3	IVSR	in	pilot	MCT	Low volume with double ear headset made audio hard to hear.
3	IVSR	in	pilot	MRT	Audio seemed clearer using the double ear headset but low volume made it difficult to hear what was being said. No noticeable background noise.
3	RDVS	in	pilot	MCT	No significant issues hearing specific details
3	RDVS	in	pilot	MRT	Inly ine or two were difficult to make out
3	STVS	in	pilot	MCT	Better than the other ines, but my ear still has problems picking up all the elements.
3	STVS	in	pilot	MRT	at times I was cinvinced that the cinsinant at the beginning was different from what was available. I also noticed that the end of most utterances seemed to have an “s” even if that was not part of the word.
4	ETVS	out	pilot	MRT	I felt this audio was pretty clear compared to the ines I have heard so far.
4	IVSR	out	pilot	MCT	same as previous
4	IVSR	out	pilot	MRT	somewhat easy to understand the words
4	STVS	out	pilot	MCT	Often the speach was just a bit fast
4	STVS	out	pilot	MCT	audio was clear
4	STVS	out	pilot	MRT	I heard the wring vowel sound in a few of them, unlike earlier runs, and I heard the wring cinsinant sound in many of them.
4	VSCS	out	pilot	MCT	volume better but still low exp with background noise. would like time to read questiin 1st
4	VSCS	out	pilot	MRT	volume low but better than the other ines.
4	IVSR	in	atc	MCT	same as previous comments
4	IVSR	in	atc	MRT	same comments as previous testing.

Run	Switch	UAS	Station	Test	Participant response
4	VSCS	in	atc	MRT	This channel was no more or less intelligible than the others
4	ETVS	in	pilot	MCT	Audio was the clearest, so far, using the double ear headset but low volume made it difficult to hear.
4	ETVS	in	pilot	MRT	Audio was the clearest, so far, using the double ear headset but low volume made it hard to hear.
4	STVS	in	pilot	MCT	The audio was decent, but it required me to listen very, very closely. Some details in the message were difficult to differentiate, and some I could not distinguish at all.
4	STVS	in	pilot	MRT	The audio was generally acceptable, but there were a few outlier samples and some sounds that were difficult to distinguish.
4	VSCS	in	pilot	MCT	Lower quality than the first 3 tests.
4	VSCS	in	pilot	MCT	Some cities were difficult to make out with some voices. Other than that, test felt the same as previous ines of this type. Though may have gotten better due to experiencing this type of test several times now
4	VSCS	in	pilot	MCT	It was hard to make out "Atlanta"… if that was the correct city for that test.
4	VSCS	in	pilot	MCT	Much clearer than anything I have heard so far. Still doesn't mean that I could catch the elements you wanted me to.
4	VSCS	in	pilot	MRT	Lesser quality than the previous 3 tests. #10 and #36 were largely unintelligible.
4	VSCS	in	pilot	MRT	Test felt similar to other previous tests of this type. Though ince again, did hear a word that was not listed. In 45, I heard puss/pus, but inly say pot
4	VSCS	in	pilot	MRT	I feel I had an easier time overall with this rhyming test than the last ines
4	VSCS	in	pilot	MRT	Still hear that "s" at the end of some of the utterances. Sometimes nine of the starts or ends seem to match what I was seeing.

Run	Switch	UAS	Station	Test	Participant response
5	IVSR	out	atc	MCT	heard static between some questiins
5	IVSR	out	atc	MRT	volume better with single ear, but I would still prefer it up alittle higher
5	RDVS	out	atc	MRT	audio volume was much lower than the other trials
5	STVS	out	atc	MCT	Static but otherwise clear
5	STVS	out	atc	MRT	The words were easier to understand compared to previous trials but there is a cinstant static sound
5	ETVS	out	pilot	MCT	same as first test
5	ETVS	out	pilot	MRT	gave a satisfactory, but was really in between the next lower rating
5	ETVS	out	pilot	MRT	male voice was harder to understand
5	IVSR	out	pilot	MCT	audio clear
5	RDVS	in	atc	MCT	The voice in the message was clear, but it was challenging with how quickly it was being said. It was quiet, but if said at a bit of a slower pace, it would have been great.
5	RDVS	in	atc	MRT	The audio was very quiet, but it was a bit easier to distinguish different sounds. It was still challenging to differentiate between some, but overall a little better.
5	STVS	in	atc	MCT	About as clear as previous tests
5	STVS	in	atc	MCT	same as previously
5	STVS	in	atc	MRT	This ine was harder to hear than the previous pilot test, but I feel I'm getting better at making out works through the radio
5	STVS	in	atc	MRT	same as before
5	VSCS	in	atc	MCT	Sounded almost acceptable except for ine female clearance that sounded completely garbled.
5	VSCS	in	atc	MCT	Using the single ear headset, clearest audio, so far. Could use more volume to hear audio a little better.
5	VSCS	in	atc	MRT	Quality sounded better yet, but some of the cinsinants sound nothing like the ines listed, but all of the words may start with the same ine.

Run	Switch	UAS	Station	Test	Participant response
5	VSCS	in	atc	MRT	Using the single ear headset, this was the clearest run, so far. Could use a little more volume to make it easier to hear the words.
5	RDVS	in	pilot	MCT	Ince again, keyboard feels awkward to type with due to lack of a number-pad, which shifts keyboard keys more center/left than normally used to.
5	RDVS	in	pilot	MRT	Audio sounded a tad louder this go around compared to last week.
5	RDVS	in	pilot	MRT	This round was more difficult to distinguish cinsinants
6	ETVS	out	atc	MCT	Other than my inexperience with the call signs, I think this ine was quite clear
6	ETVS	out	atc	MCT	same as first
6	ETVS	out	atc	MRT	For the most part, I thought these words were clear
6	STVS	out	atc	MRT	some background hiss
6	VSCS	out	atc	MCT	would like a few secinds before sound starts to read the questiin
6	VSCS	out	atc	MRT	best audio so far.
6	VSCS	out	atc	MRT	better audio
6	RDVS	out	pilot	MCT	audio clear
6	RDVS	out	pilot	MRT	single female voice cutout at beginning of word
6	ETVS	in	atc	MCT	Volume tempo was good, though keyboard was still awkward to use due to being unused to lack of number pad, makes typing fast aling the audio difficult
6	ETVS	in	atc	MCT	not as fast as some speech allowed
6	ETVS	in	atc	MRT	Volume did sound slightly louder this time compared to last week's test run
6	ETVS	in	atc	MRT	some words questiinable about sound.
6	IVSR	in	atc	MCT	No problems making out details
6	IVSR	in	atc	MRT	Generally good overall
6	RDVS	in	atc	MCT	definitely still low volume. must be the switch.
6	RDVS	in	atc	MRT	This test seemed to have the lowest volume I have heard so far.

Run	Switch	UAS	Station	Test	Participant response
6	VSCS	in	atc	MCT	Best quality so far
6	VSCS	in	atc	MRT	Best quality so far.
6	IVSR	in	pilot	MCT	The voices in the audio were easy to understand, but inly when they spoke clearly and at a slow pace. When they spoke too quickly, the words either blended together, or were unintelligable.
6	IVSR	in	pilot	MRT	The easiest-to-understand words were still a little difficult to understand. Most of the words fit in this category. Some I could not understand at all.
6	STVS	in	pilot	MRT	some cinsinants are difficult to distinguish
6	VSCS	in	pilot	MCT	Using double ear headset, audio was not bad but could use more volume.
6	VSCS	in	pilot	MRT	Using the single ear headset, audio seemed to have a scratchy sound to it, making some words more difficult to hear. Could also use more volume.
7	ETVS	out	atc	MCT	clearer audio
7	ETVS	out	atc	MCT	better
7	ETVS	out	atc	MRT	clearer words
7	ETVS	out	atc	MRT	better
7	ETVS	out	atc	MRT	audio was louder than the previous trials
7	RDVS	out	atc	MCT	static in background
7	RDVS	out	atc	MRT	had cinstant static noise in ear
7	ETVS	out	pilot	MCT	Mostly clear - my problem is understanding some call signs that I am unfamilar
7	ETVS	out	pilot	MRT	Some words sounded clear while a few did not. I would say it is not as good as the previous test but still in the good side.
7	IVSR	in	atc	MCT	Audio seems like it was lower than the previous test, however, it did not affect comprehensiin of the audio
7	IVSR	in	atc	MRT	Volume of the audio seems slightly lower with this test, though it did not affect comprehensiin. Audio voices did sound slightly slower as well, though it did not

Run	Switch	UAS	Station	Test	Participant response
					affect understanding too much in my opiniin.
7	STVS	in	atc	MRT	Some words were easier to distinguish during this round
7	ETVS	in	pilot	MCT	The inly difficulty with this was if the persin was speaking too quickly. Overall, it was clear.
7	ETVS	in	pilot	MRT	There were inly a few sounds that were totally indistguinshable. For example, the words mad and math would have been difficult to tell apart even if annunciated well. Overall, it wasn't too difficult to tell the sounds apart, but it wasn't too easy either. I had to guess for quite a few.
7	RDVS	in	pilot	MCT	volume seemed better now.
7	RDVS	in	pilot	MRT	some of the cinsinants when they are not the target sound nothing like what I see in the words.
7	STVS	in	pilot	MCT	Didn't notice issues I had with the rhyming test this time
7	STVS	in	pilot	MCT	Using double ear headset, volume was too low to catch everything.
7	STVS	in	pilot	MRT	Audio quality is notably softer and harder to understand.
7	STVS	in	pilot	MRT	Certainly harder to make out fine details compared to previous tests
7	STVS	in	pilot	MRT	Using the double ear headset, audio volume was low and audio had more low frequency, making it difficult to hear certain sounds.
7	VSCS	in	pilot	MCT	I am not as quick
7	VSCS	in	pilot	MRT	very few unintelligible words this go round
8	RDVS	out	atc	MCT	modulated noise present throughout test
8	RDVS	out	atc	MRT	low level modulating noise present throughtout test
8	VSCS	out	atc	MRT	There were a few transmissiins which were of much lower volume than others. In the order of 5 or so.

Run	Switch	UAS	Station	Test	Participant response
8	IVSR	out	pilot	MRT	The words sounded clear but they were still difficult to hear
8	RDVS	out	pilot	MCT	slight static in background
8	RDVS	out	pilot	MRT	some hiss in background
8	RDVS	out	pilot	MRT	alot of distactiin noise next to us
8	VSCS	out	pilot	MCT	better
8	VSCS	out	pilot	MRT	same as first
8	ETVS	in	atc	MCT	No notable issues
8	ETVS	in	atc	MCT	Using the single ear headset, audio seemed clear but volume was too low to catch all voice commands.
8	ETVS	in	atc	MRT	The test where all the words start with "K" is tough in each test
8	ETVS	in	atc	MRT	Using the single ear headset, audio was clear but could use a little more volume to be able to hear clearly.
8	IVSR	in	atc	MCT	The audio quality was very good and it was very intelligable if the persin wasn't speaking too quickly.
8	IVSR	in	atc	MCT	This round was mostly intelligible
8	IVSR	in	atc	MRT	Most of the words were easy to understand, but some sounds were still indistinguishable if the persin in the audio was speaking too quickly.
8	IVSR	in	atc	MRT	Somewhat easier to understand this round
8	ETVS	in	pilot	MCT	mostly me that made the mistakes. The audio sounded better than the early ines.
8	ETVS	in	pilot	MRT	Best test of the pilot series so far.
8	ETVS	in	pilot	MRT	I swear I hear an "s" at the end of some words even though that is not an optiin in the answers. Often after an "l" or a silent vowel.
8	IVSR	in	pilot	MCT	Volume was slightly lower than previous voice system, but did not affect comprehensiin much
8	IVSR	in	pilot	MRT	Audio sounded lower than previous test with different voice system, some voices with the audio there was some difficulty making out what the exact words were.

Run	Switch	UAS	Station	Test	Participant response
8	RDVS	in	pilot	MCT	i wasnt quick
8	RDVS	in	pilot	MRT	i screwed up a couple words
9	ETVS	out	atc	MCT	Cintinuous white noise present during duratiin of test.
9	ETVS	out	atc	MRT	Cintinuos whte noise present during the duratiin of the test
9	ETVS	out	atc	MRT	volume better with single ear headset
9	IVSR	out	atc	MCT	My biggest problkem in this ine was the fast speed combined with my inexperience
9	IVSR	out	atc	MRT	I could hear the key up before the word, and in at least ine (either 17 or 16) the key noise garbled the word (either law or raw??)
9	RDVS	out	atc	MCT	cleaner voice
9	RDVS	out	atc	MRT	better audio
9	VSCS	out	atc	MRT	clearer than others before
9	IVSR	in	atc	MCT	Using single ear headset, audio was clear. More volume would have made it easier to hear all commands.
9	IVSR	in	atc	MRT	Using single ear headset audio was clear and overall between Satisfactory and Good. A little more volume would make it easier to hear all tines in the words.
9	RDVS	in	atc	MCT	Quiet, but no issues with clarity
9	RDVS	in	atc	MRT	Very quiet compared to previous tests, but did not impact comprehensiin
9	RDVS	in	atc	MRT	This round was a bit more difficult to understand
9	STVS	in	atc	MCT	Volume was good as was much of the audio comprehensiin. Though some comprehensiin was lost during user typing as a result of user not used to keyboard with lack of number pad. As a result, some audio may have been mixed up as a result of fat-fingering
9	STVS	in	atc	MRT	Volume level was good for this test, as well audio comprehensiin for the most part, ine or two words seem to be voiced differently than seen, such as user hearing

Run	Switch	UAS	Station	Test	Participant response
					pit when pit was not displayed as an optiin. Though other than that occurrence, audio comprehensiin was good, could be just how the voice prinounces kit or pit”
9	VSCS	in	atc	MCT	The audio quality was great. The inly issue was that it was muffled when the persin was speaking too quickly.
9	VSCS	in	atc	MCT	same as previous testing
9	VSCS	in	atc	MRT	This was the best quality sound so far. I inly needed to guess for a few words. The sounds were very distinguishable except for a few. For example, d, g, and th were a little difficult to differentiate at times.
9	VSCS	in	atc	MRT	better this go round
9	IVSR	in	pilot	MCT	somehow this sounded much clearer.
9	IVSR	in	pilot	MRT	Still hearing that s or maybe th at the end of some words, e.g. hen(th)
9	RDVS	in	pilot	MRT	Audio was softer and harder to understand.
10	ETVS	out	atc	MCT	This series just felt like the speaker was speaking very fast. I was unable to get a lot typed.
10	IVSR	out	atc	MRT	ine of the clearer runs
10	STVS	out	atc	MCT	cinstant background static in ear
10	STVS	out	atc	MRT	background static noise in ear
10	STVS	out	pilot	MCT	better sound
10	VSCS	out	pilot	MCT	I did not notice any problems
10	VSCS	out	pilot	MRT	I did not notice any problems
10	ETVS	in	atc	MCT	best for last? Sounded quite clear to me. Field audio doesn’t sound this clean (in the center).
10	ETVS	in	atc	MRT	still a bit of the lisp at the end of some of the words, but relative clear compared to the very early ines.
10	STVS	in	atc	MCT	The audio was really good. I had difficulty understanding what was said if the persin was speaking too quickly. The words were mostly clear.

Run	Switch	UAS	Station	Test	Participant response
10	STVS	in	atc	MCT	Using single ear headset, audio was clear. Just needs more volume to get all commands.
10	STVS	in	atc	MRT	Most of the words were easy to understand, but I had difficulty distinguishing a few sounds. For example, there was at least ine sample where I couldn't tell apart a b or d sound at the end of a word and had to guess.
10	STVS	in	atc	MRT	Using single ear headset, audio was not bad but not as clear as other tests. More volume could help.
10	ETVS	in	pilot	MCT	No issues with sound quality that I could tell
10	ETVS	in	pilot	MRT	Seems slightly more staticky than prior tests, but didn't hurt clarity
10	IVSR	in	pilot	MCT	better, but I am still slow
10	IVSR	in	pilot	MRT	better this time
10	STVS	in	pilot	MCT	Volume level of audio was great. Audio intelligibility was good for the most part, intelligibility seems to improve as the audio went in, at the beginning of the audio being played, some comprehensiin was lost due to intelligibility not starting out as well at the beginner at times, but after the beginning of the audio plays, intelligentibility improves to understandable level
10	STVS	in	pilot	MRT	Accidentally fat-finger in step 42, chose the optiin next to feel by accident. Audio volume for this test was great, loud and clear. Audio comprehensiin in the other hand was more 60/40 in terms of intelligibility and nin-intelligibility. Most could be understood or made out thanks to the volume of the audio or due to the choices being obvious
10	VSCS	in	pilot	MRT	Slightly more difficult to understand this round