ote technical note technical

Voice Over Internet Protocol: Speech Intelligibility Assessment

Ferne Friedman-Berg, Ph.D., Human Factors Team – Atlantic City, ATO-P
Kenneth Allendoerfer, Ph.D., Human Factors Team – Atlantic City, ATO-P
Atul Deshmukh, Ph.D., Hi-Tec Systems, Inc.

December 2009

DOT/FAA/TC-TN-09/04

U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-TN-09/04		

4. Title and Subtitle	5. Report Date
Voice Over Internet Protocol: Speech Intelligibility Assessment	December 2009
	6. Performing Organization Code AJP-6110
7. Author(s) Ferne Friedman-Berg, Ph.D., Human Factors Team – Atlantic City, ATO-P Kenneth Allendoerfer, Ph.D., Human Factors Team – Atlantic City, ATO-P Atul Deshmukh, Ph.D., Hi-Tec Systems, Inc.	8. Performing Organization Report No. DOT/FAA/TC-TN-09/04
9. Performing Organization Name and Address Federal Aviation Administration Human Factors Team – Atlantic City, ATO-P 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 Human Factors Research and Engineering Group 800 Independence Avenue, S.W. Washington, DC 20591	13. Type of Report and Period Covered Technical Note 14. Sponsoring Agency Code ATO-P

15. Supplementary Notes

16. Abstract

In this study, we examined the speech intelligibility of five codecs and parameter settings to determine their suitability for Air Traffic Control use. These codecs (g711r64, g726r32, g729r8, g723r53, and g726r16) convert analog communications to digital communications. We used standard intelligibility tests – the Modified Rhyme Test (MRT) and the Message Completion Task (MCT) – to examine intelligibility differences. For the MRT, codec type had an effect on intelligibility. Performance in an uncompressed audio condition was the best, followed by codecs g711r64 and g726r32, then g729r8, and then g723r53 and g726r16, respectively. Reaction times were consistent with the performance data, with the better performing codecs having shorter reaction times. Subjective intelligibility and acceptability ratings matched the performance and reaction time data. For the MCT, the intelligibility and acceptability ratings for all codecs were high, but we found no clear pattern in the performance data. We recommend that the National Airspace System Voice Switch (NVS) program further evaluate the three best performing codecs, g711r64, g726r32, and g729r8. These codecs delivered the best performance, the shortest reaction times, and the most positive feedback. We also recommend that the NVS Program Office investigate the effect of channel noise, ambient noise, and level of network utilization on the intelligibility of different codecs.

17. Key Words	18. Distribution Statement
Message Completion Task Modified Rhyme Test NAS Voice Switch Speech Codec Voice over Internet Protocol	This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161. A copy is retained for reference at the William J. Hughes Technical Center Library.

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	36	

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

Table of Contents

List of Illustrations

Acknowledgments

We thank the National Airspace System Voice Switch Program Office for sponsoring this project. We also thank Ken Schulz, Ph.D. (Hi-Tec Systems, Inc.) for his help building the data collection program and for his expert advice and guidance. Thanks to Otto Smith (Engility Corporation) for setting up and monitoring the routers, enabling the codecs, and writing the software that allowed us to run the codecs. We thank Robert Bortu (Engility Corporation) for his help with computer hardware and software. Thanks to all of the participants who graciously agreed to volunteer for this study. We also thank April Jackman (Northrop Grumman) for her careful review and editing of this document.

Executive Summary

In the Next Generation Air Transportation System (NextGen) timeframe, voice communications will remain an important part of Air Traffic Control (ATC). The National Airspace System (NAS) Voice Switch (NVS) is one of the enabling technologies for voice communication in NextGen. Over the next two decades, the Federal Aviation Administration (FAA) will replace the legacy voice switch systems at ATC facilities with the NVS. The NVS will replace the Voice Switching and Control System (VSCS), the Enhanced Terminal Voice Switch (ETVS), the Rapid Deployment Voice Switch (RDVS), the Interim Voice Switch Replacement (IVSR), and other legacy systems. The NVS will provide controllers with their primary means of operational ground-to-ground and air-to-ground voice communications. The FAA anticipates that operational use of the NVS at FAA key sites will begin in the year 2014.

The NVS will incorporate new capabilities and will provide enhanced interconnectivity for controllers within and across ATC facilities. The NVS will allow controllers to access resources remotely, and facilities will be able to offload work to other facilities in response to changing operational conditions. The FAA expects to take advantage of the flexible routing provided by the NVS to support real-time resource reallocation and other NextGen enhancements like dynamic resectorization.

Voice over Internet Protocol (VoIP) is one proposed NVS technology. VoIP is used for digital voice communications. In this study, we examined the speech intelligibility of different VoIP codecs and parameter settings to determine their potential suitability for ATC use.

Before considering VoIP to be a viable technology, the FAA must identify speech codecs and parameter settings that provide an acceptable level of intelligibility for ATC communications. A *speech codec* is an algorithm that converts analog voice signals to digital data (a process known as encoding) and digital data to analog signals for playback (a process known as decoding). Speech codecs use a variety of compression techniques, and the resulting voice signals vary in quality. In addition, each codec may have more than one associated bitrate. The NVS program would prefer using codecs with low bitrates because they require less bandwidth and can reduce costs. However, lower bitrates are often associated with lower voice quality because they use heavier compression. This study examined differences in intelligibility for five codecs:[1] G.711 (64 bits), G.726 (32 bits), G.729 (8 bits), G.723.1 (53 bits), and G.726 (16 bits) using a standard speech intelligibility test, the Modified Rhyme Test (MRT). ATC participants also completed a second intelligibility test, the Message Completion Task (MCT), to examine the intelligibility of standard ATC words and phraseology.

[1] Hereafter, we will refer to the codecs using the following abbreviations: G.711 (64 bits) = g711r64, G.726 (32 bits) = g726r32, G.729 (8 bits) = g729r8, G.723.1 (53 bits) = g723r53, and G.726 (16 bits) = g726r16.

For the MRT, codec type had an effect on intelligibility. Performance in the uncompressed audio condition was the best, followed by codecs g711r64 and g726r32. The performance for codec g729r8 was in the middle, and the two lowest performing codecs were g723r53 and g726r16. We also examined reaction times to determine whether the best performing codecs had correspondingly short reaction times. We found that the reaction times were consistent with the performance data, with the better performing codecs having shorter reaction times. Subjective intelligibility and acceptability ratings matched the performance and reaction time data. For the MCT, the intelligibility and acceptability ratings for all of the codecs were high, indicating high perceived intelligibility and acceptability when listening to standard controller speech. However, we were unable to discern any clear pattern for the error data, most likely due to the small number of participants.

On the basis of our assessment, we recommend that the NVS Program Office consider using the three best performing codecs for VoIP. Codecs g711r64, g726r32, and g729r8 delivered the best performance, the shortest reaction times, and received the most positive feedback. We also recommend that the program office make speech intelligibility a top priority when evaluating performance and cost reduction trade-offs related to lowering the bitrate. As the program office moves closer to selecting a VoIP codec for operational use, we believe it is important to investigate how factors like channel noise, ambient noise, and varying levels of network utilization affect speech intelligibility for the different codecs. Although these factors may affect intelligibility or acceptability, we were unable to evaluate them in this assessment.

Furthermore, we recommend that the program office examine the suitability of using objective algorithms to measure speech quality (e.g., the Perceptual Evaluation of Speech Quality or the PESQ) and speech intelligibility (e.g., Speech Intelligibility Index or SII). The benefits of these metrics are that they can be measured objectively and some can be measured using Commercial-Off-The-Shelf tools. If we can demonstrate that these metrics are reliable in an ATC context, they would provide us with a simpler way to measure speech quality reliably. However, these algorithms have limitations, so we also recommend performing a human factors assessment to identify their benefits and limitations.

1. INTRODUCTION

Over the next two decades, the National Airspace System (NAS) Voice Switch (NVS) will replace the legacy voice switch systems at Federal Aviation Administration (FAA) Air Traffic Control (ATC) facilities. The NVS will replace the Voice Switching and Control System (VSCS) in Air Route Traffic Control Centers (ARTCCs). The NVS will also replace the Enhanced Terminal Voice Switch (ETVS), the Rapid Deployment Voice Switch (RDVS), the Interim Voice Switch Replacement (IVSR), and other legacy systems in Terminal Radar Approach Controls (TRACONs) and Airport Traffic Control Towers (ATCTs). The FAA anticipates that operational use of the NVS at FAA key sites will begin in the year 2014.

In the current NAS, there are many different forms of controller communications. For example, controllers use air-to-ground radios to issue voice instructions and receive voice responses from pilots. Controllers use telephones and intercoms to coordinate operations with other controllers at other sectors or facilities. Controllers use touchscreens and physical controls located at their working positions to access and configure radios, phones, and intercoms. Controllers also use headsets, handsets, or speakers to talk and listen.

In the Next Generation Air Transportation System (NextGen), many controller-to-pilot communications will occur via Data Communications (DataComm) and many controller-to-controller communications will occur via the ATC automation systems. For instance, in NextGen, a controller may uplink a new heading to an aircraft via DataComm, thereby eliminating several air-to-ground voice communications. The controller may coordinate with an adjacent sector regarding a new heading by making a pointout entry on the automation system, thereby eliminating several ground-to-ground communications.

However, even in the NextGen timeframe, ATC procedures will require controllers to communicate certain time-critical or safety-critical information by voice. Controllers will require a backup communication method in case of a DataComm failure and more complicated, between-sector coordination will still require direct communications. Finally, not all aircraft will be DataComm-equipped. Therefore, voice communications will remain an important part of ATC in the NextGen timeframe. One of the enabling technologies for voice communication in NextGen is the NVS. The NVS will provide controllers with their primary means of operational ground-to-ground and air-to-ground voice communications.

1.1 Purpose

The NVS will incorporate new technology and capabilities and will provide enhanced interconnectivity for controllers both within and across ATC facilities. For example, the NVS network architecture will allow controllers to access resources remotely, and facilities will be able to offload work to other facilities in response to changing operational conditions. In the long term, the FAA expects to take advantage of the flexible routing provided by the NVS to support real-time resource reallocation and other NextGen enhancements such as dynamic resectorization. The FAA will likely need to modify operational procedures when implementing these new capabilities and will also need to evaluate how these changes affect controller performance, workload, training, and staffing requirements. For these reasons, the NVS Program Office has sought the involvement of FAA human factors personnel to ensure that the NVS program incorporates human factors standards and best practices.

1.2 Background

In the current study, we performed an assessment of the speech intelligibility of one proposed NVS technology, Voice over Internet Protocol (VoIP), to determine its potential suitability for ATC use. VoIP is a technique for digital voice communications that uses standard network protocols and hardware. This assessment will examine the speech intelligibility of different VoIP codecs and parameter settings.

In simple terms, a *speech codec* is an algorithm that converts analog voice signals to digital data, (a process known as encoding) and digital data to analog signals for playback (a process known as decoding). These codecs can exist in hardware or software. Different speech codecs use different encoding and decoding techniques, and the resulting voice signals vary in their audio characteristics, which may affect the resulting audio quality. In particular, speech codecs differ in the amount and type of compression they apply. Heavy compression, especially *lossy* compression that permanently removes some portion of the audio signal, normally reduces voice quality. This is especially true when the compression introduces so-called *compression artifacts*. For example, heavily compressed audio can sound wavy or jumpy. In addition to employing different compression techniques, each codec has one or more associated bitrate(s). The bitrate is the number of bits that are required to transmit a voice call. Low bitrates require less network bandwidth and offer potential cost savings. Because they use heavy compression, lower bitrates are normally associated with lower voice quality. The NVS program and system vendors may prefer to use codecs with lower bitrates because of their potential to reduce program costs.

Before the FAA can consider VoIP to be a viable technology for use by the NVS program, the program must identify VoIP codecs and parameter settings that provide a level of speech intelligibility necessary for ATC communications. Although previous FAA studies have evaluated vocoder intelligibility (La Due, Sollenberger, Belanger, & Heinze, 1997; Sollenberger, La Due, Carver, & Heinze, 1997; Sollenberger, McAnulty, Kerns, 2003; Zingale, McAnulty, & Kerns, 2003), none have examined the intelligibility of different codec compression algorithms or bitrates in an ATC context. This study examines differences in intelligibility for five speech codecs. We believe this study will help the NVS program select an appropriate codec and will provide them information about any performance trade-offs between bitrate and intelligibility.

1.3 Controller Speech

Controller speech differs from regular speech in several ways. For example, controllers often speak very quickly. Normal speech rates are about 156 words per minute (wpm), whereas controller speech may be as fast as 210 wpm (GAIN Working Group E: Flight Ops/ATC Ops Safety Information Sharing, 2004). Controller speech also uses a restricted vocabulary and syntax, including the phonetic alphabet (e.g., Alpha, Bravo, Charlie) and phraseology mandated by FAA Order 7110.65, *Air Traffic Control.* It contains many technical terms and nonstandard words, such as geographic fix names, that are unfamiliar to the general public. Controller speech conveys precise operational data, such as altitudes, that must be perceived accurately. This level of required precision, where one misheard syllable could lead to a major problem, is not typical of regular speech.

It is possible that factors such as these might make some VoIP codecs, even those used for other applications, unsuitable for ATC communications. We believe this assessment will help us identify codecs and parameter settings that are clearly unacceptable and will allow us to guide the program office toward those options most suitable for operational use.

2. METHOD

2.1 Participants

Eighteen FAA employees assigned to the William J. Hughes Technical Center (WJHTC), or temporarily located at the WJHTC for the NVS or other FAA projects, served as participants. These employees included engineers, scientists, managers, and administrative staff. Most of the participants had some familiarity with ATC terminology and concepts, but none of them had any operational ATC experience.

We also recruited six current Certified Professional Controllers (CPCs) and Front Line Managers (FLMs). We recruited CPCs and FLMs temporarily located at the FAA WJHTC for other activities (e.g., testing, meetings, and training); we took all necessary steps to avoid disrupting or inconveniencing outside activities.

Participation in the study was voluntary. We informed the participants that they could end their participation for any reason at any time during a session without penalty. In addition, if we determined while conducting a session that ending the session would be in the best interest of the participant, we could terminate the session at that time. However, none of the participants chose to end their sessions early.

We excluded volunteers from participating if they had a medical condition that prevented them from using a telephone headset for 90 minutes. In addition, we also excluded volunteers who required special equipment (e.g., a hearing aid) to use a telephone.

2.2 Facilities and Personnel

We conducted this study at the Research Development and Human Factors Laboratory (RDHFL) located at the FAA WJHTC. Engineering Research Psychologists from the Human Factors Research and Engineering Group, Human Factors Team–Atlantic City and its contract support personnel conducted this study. Engineers from the FAA Laboratory Future Development Team and its contractor configured the VoIP and other communications equipment.

2.3 Equipment

2.3.1 Routers and Voice Communications System

For this study, we used two Cisco 2811 Integrated Services routers. These devices encoded the voice signal at the source and decoded the signal at the playback position. The routers were capable of the following speech codecs: G.711, G.723.1, G.726, and G.729. The routers also allowed us to set the bitrate for each codec.

The RDHFL uses a Telex Adam 20 Hz/20 kHz intercom system that routes audio throughout the experiment rooms and laboratories. The VoIP routers passed the audio signals first to the Telex, and then to the participants' Telex PH5-R5 headsets.

2.3.2 The Modified Rhyme Test and the Message Completion Task

The Modified Rhyme Test (MRT) is an intelligibility test that consists of six lists of 50 items (American National Standards Institute [ANSI], 1996). Each item contains six rhyming monosyllabic words that differ only in their initial or final phoneme (e.g., pus, pub, pun, puff, puck, and pup; or hold, cold, told, fold, sold, and gold). The purpose of the task is to evaluate whether participants can identify a spoken word and distinguish it from a set of phonemically similar words.

The Message Completion Task (MCT) is an intelligibility test that consists of six versions of five ATC phrases. Each phrase had some details (e.g., altitudes, headings, call signs) removed. The purpose of this task is to examine whether controllers can fill in the missing details after listening to a recording of the phrase.

2.4 Procedure

Each participant was tested individually. Sessions for non-ATC personnel lasted about 60 minutes per participant. This included time for instructions, paperwork, training, data collection, and debriefing. Sessions for ATC participants lasted about 75 minutes.

At the start of the session, we briefed the participants regarding the goals of the project, the procedures, the data collection techniques, and their rights as participants. This included their rights to informed consent, anonymity, confidentiality, and termination without penalty. Each participant signed an Informed Consent Form (see Appendix A) and then completed a Background Questionnaire (see Appendix B).

After completing the background questionnaire, all participants completed the MRT. For the MRT, the participant was seated in front of a computer screen next to the researcher (see Figure 1). Using the telex headset, the participant listened to a prerecorded Waveform audio format (.wav) file that corresponded to a word (e.g., pun) selected from a set of six foils (e.g., pus, pub, pun, puff, puck, and pup). The computer displayed six choices to the participant (see Figure 2), and the participant's task was to select the correct word from the six alternatives. After making a selection, the participant repeated the word aloud, and the experimenter recorded (on a second computer) whether the verbal response was correct. All participants received five training trials to familiarize themselves with the procedure.

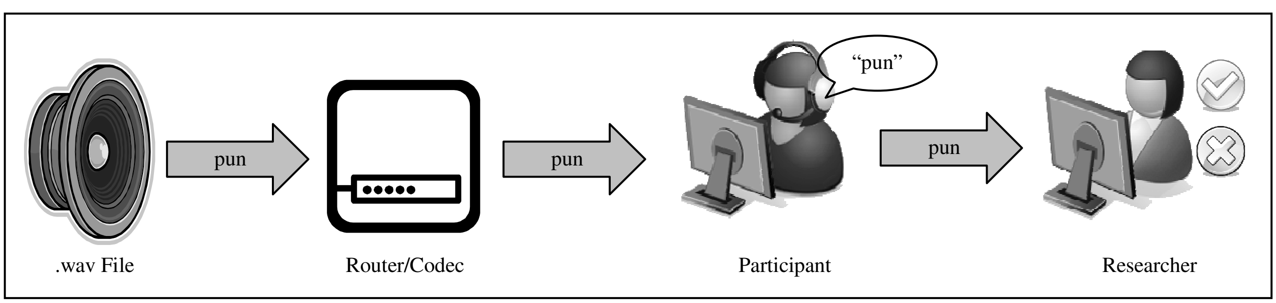

Figure 1. Experimental set-up.

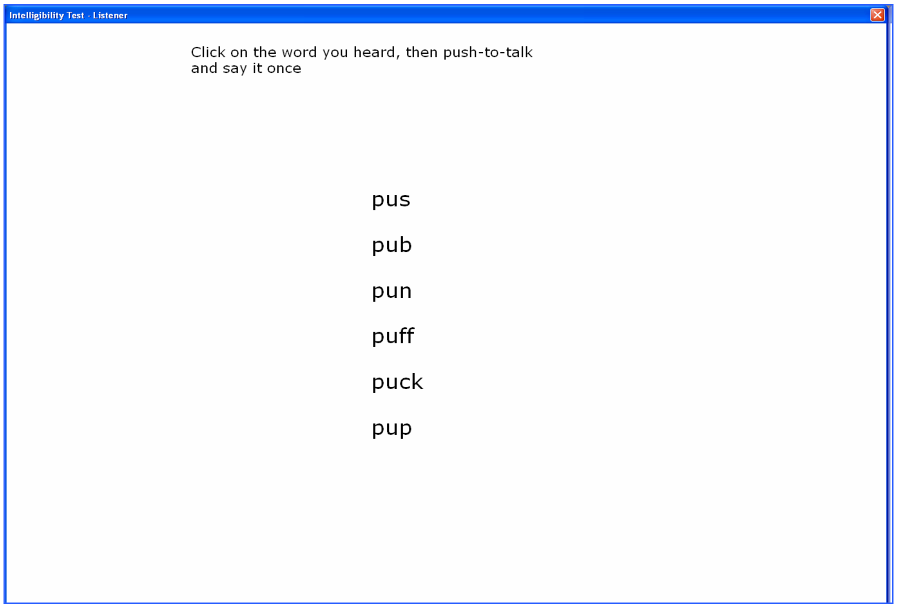

Figure 2. Screen shot of an MRT trial.

There were six lists of 50 items, but each participant saw only one list. For each participant, a list was paired with a baseline condition (uncompressed audio) and five codecs (see Table 1). The uncompressed audio condition was of a much higher voice quality than typically found in a Public Switched Telephone Network (PSTN). The NVS Program Office selected all of the codecs and sampling rates. Each word within a set served as the correct answer for a single condition. Consequently, even though a participant saw the same list six times, the set of correct answers was different for each condition. The order of list items was also randomized.

Table 1. Experimental Conditions

Condition	Codec and Parameter Settings	Bit Rate
Baseline	Uncompressed audio	N/A
VoIP 1	G.711	64
VoIP 4	G.726	32
VoIP 5	G.729	8
VoIP 2	G.723.1	53
VoIP 3	G.726	16

The controllers also completed a second task, the MCT, in which we examined the intelligibility of standard ATC words and phraseology. For this task, the controllers listened to five ATC phrases and filled in 13 missing details (e.g., altitudes, headings, call signs) on a worksheet (see Appendix C). There were six different answer sets for the five phrases. Each participant heard the same five phrases for all six codecs, but the answer set was different for each codec. Answer sets and codecs were counterbalanced across the participants.

After completing each set of 50 MRT trials and each set of phrases for the MCT, the participants completed a Feedback Questionnaire (see Appendix D). The participants rated the intelligibility and acceptability of the speech signal for each codec and provided comments. For the MRT, the researchers offered the participants a 2-minute break after each set of 50 trials and a 10-minute break after the third set.

3. DATA ANALYSIS

3.1 Background Questionnaire

Of the 24 participants in this study, 18 were noncontrollers. Of the noncontrollers, 12 were male and 6 were female. Six participants were current or former controllers; all six controllers were male. The median age of the noncontrollers was 43.5, and the median age of the controllers was 51. The ages of the noncontrollers ranged from 24 to 59, and the ages of the controllers ranged from 42 to 70. None of the participants had a medical or physical condition that would make it difficult for them to use a telephone for 90 minutes or that would require the use of a hearing aid or other assistive device when using a telephone.

3.2 Modified Rhyme Test

3.2.1 MRT Performance

We compared the speech intelligibility for the six conditions as measured by performance on the MRT (see Figure 3). Overall accuracy was quite high in all conditions. The overall model indicated that codec type did have an effect on intelligibility, Wald $\chi^2(5) = 99.79, p < .001$.[2] Further examination of pairwise comparisons indicated that performance in the uncompressed audio condition was the best ($p \leq .001$ for all comparisons). This was followed by codecs g711r64 and g726r32, which had virtually identical performance. The performance of codec g729r8 was in the middle, with slightly lower performance than codecs g711r64 ($p = .10$) and g726r32 ($p = .12$), similar performance to codec g723r53, and slightly higher performance than codec g726r16 ($p = .07$). The two lowest performing codecs g723r53 and g726r16 had similar performance.

---

[2] We used a Generalized Estimating Equation (GEE) to account for the nonnormality of proportions (Hanley, Negassa, Edwardes, & Forrester, 2003). The GEE is more flexible than a standard Generalized Linear Model (GLM) because it allows researchers to specify both the distribution and correlational structure of the data (Ballinger, 2004).

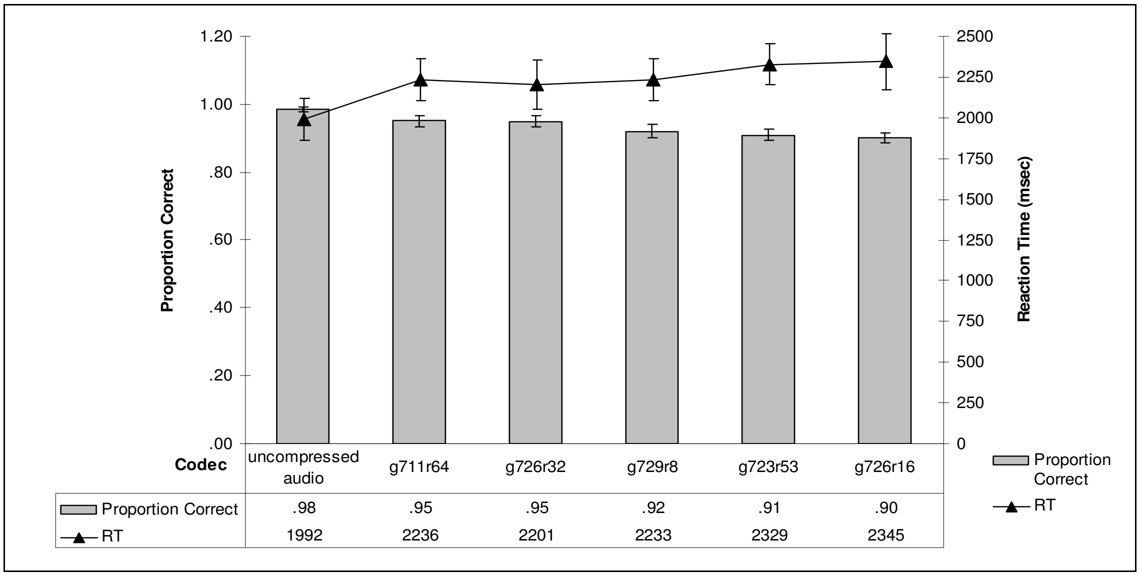

Figure 3. MRT proportion correct and mean reaction time by codec – error bars are equal to 2 standard errors.

3.2.2 MRT Reaction Time

Operationally, it is optimal to select codecs that have both high intelligibility and short reaction times, so we examined the MRT reaction time data to determine whether the best performing codecs also had correspondingly short reaction times. Because incorrect answers often include extremely long outliers, we analyzed reaction times for correct items only.

As illustrated in Figure 3, the reaction time data matched the accuracy data. We performed pairwise comparisons and found that reaction times were fastest for the uncompressed audio when contrasted with the five codec conditions ($p \leq .001$ for all comparisons). Reaction times for codec g711r64 were similar to reaction times for g726r32; both of those codecs had similar reaction times to codec g729r8. Reaction times for codec g711r64 were only marginally faster than codecs g723r53 ($p = .13$) and g726r16 ($p = .09$), but reaction times for codec g726r32 were faster than both codec g723r53 ($p = .01$) and codec g726r16 ($p = .006$). Reaction times for codec g729r8 were marginally faster than reaction times for codec g723r53 ($p = .06$) and g726r16 ($p = .05$). Reaction times for codec g723r53 were similar to reaction times for codec g236r16.

3.2.3 MRT Ratings and Comments

We examined the intelligibility and acceptability rating for each codec to determine whether the subjective ratings matched the performance data and reaction time data. Ratings used a 7-point scale that ranged from 1 (*poor intelligibility* or *low acceptability*) to 7 (*high intelligibility* or *high acceptability*). The median ratings for all of the codecs were fairly high, with the lowest being a 5. Because of the lack of variability in the uncompressed audio condition, we did not conduct statistical tests on the ratings. However, the median ratings were in line with the performance and reaction time data (see Figure 4). Uncompressed audio received the highest median ratings, followed by g711r64, g726r32, and g729r8. Codecs g723r53 received similar intelligibility ratings but lower acceptability ratings than codecs g711r64, g726r32, and g729r8. Codec g726r16 received the lowest median ratings for both intelligibility and acceptability.

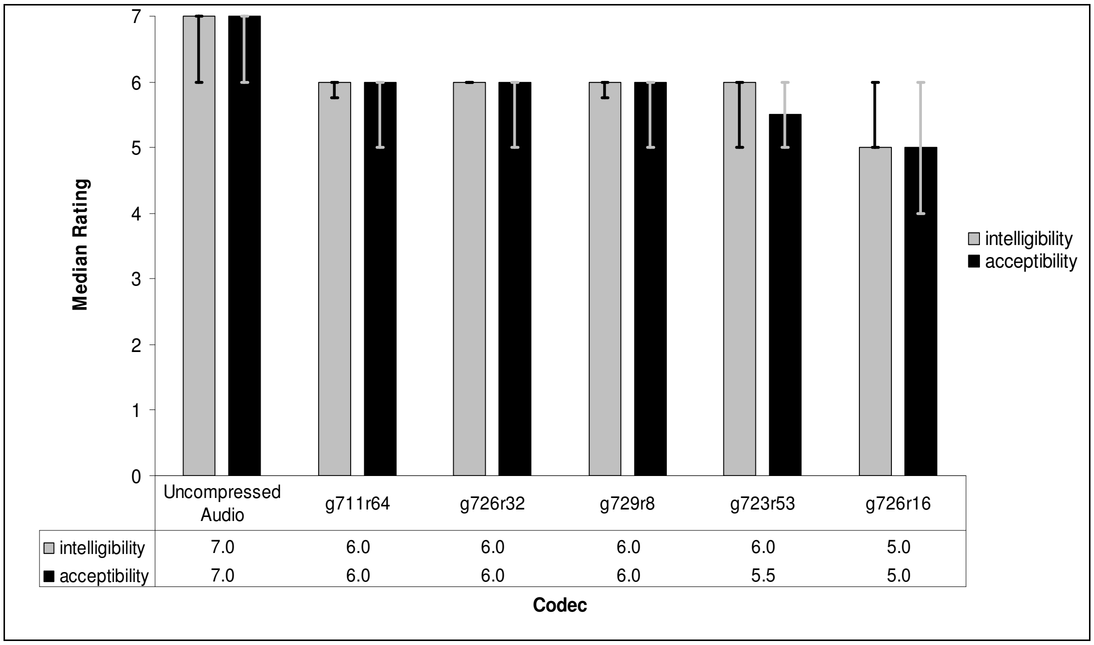

Figure 4. MRT median ratings for intelligibility and acceptability by codec type – bars indicate interquartile range.

We examined the subjective feedback and found that it was consistent with the ratings given by the participants, and it was also consistent with the performance and reaction time data (see Appendix E, Table E1). Almost all of the comments for the uncompressed audio were positive. Participants used words like *clear*, *easy to understand*, and *distinct* to describe the uncompressed audio. Codecs g711r64, g726r32, and g729r8 had both positive and negative comments, but there were only a small number of negative comments. Most of the negative comments referred to clipping of the initial or final consonants or to difficulty distinguishing exemplars of certain classes of phonemes like fricatives. Codecs g723r53 and g726r16 also received both positive and negative comments, but the negative comments were somewhat more critical than the comments for the other codecs. In addition to referring to problems related to distinguishing initial and final consonants, the negative comments also referred to problems related to distorted words or degraded, fragmented, and garbled audio.

3.3 Message Completion Task

3.3.1 MCT Performance

For the MCT, the average percentage correct was very high for each codec (see Figure 5). Uncompressed audio had the lowest percentage correct, followed by g729r8 and g723r53. However, due to the small number of participants and small number of errors, we were unable to discern any clear statistical pattern.

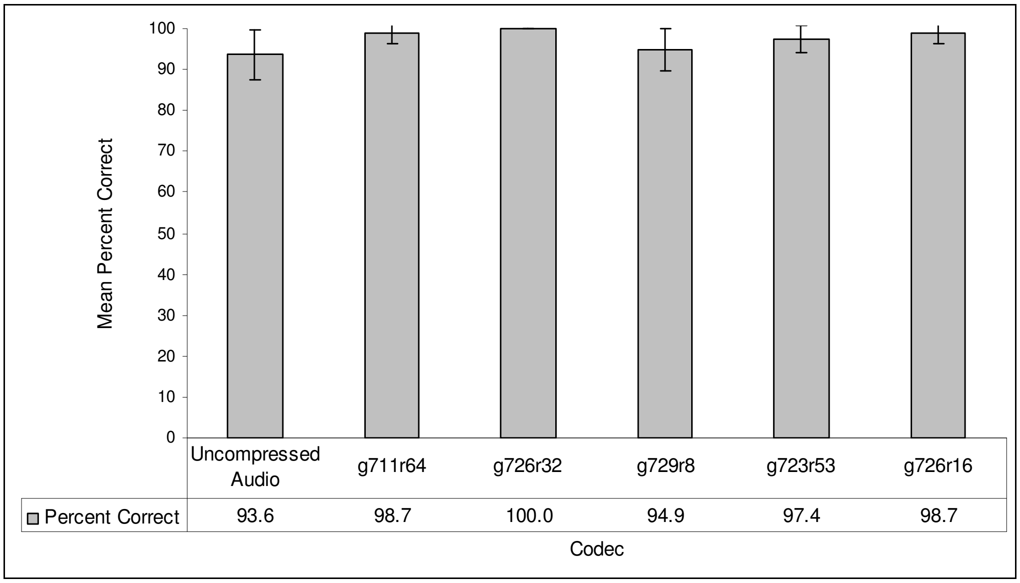

Figure 5. MCT average percent correct by codec type – error bars are equal to 2 standard errors.

3.3.2 MCT Ratings and Comments

Although uncompressed audio received the highest median rating, all of the median ratings for the MCT were 6 or higher, indicating high perceived intelligibility and acceptability for standard controller speech for all of the codecs (see Figure 6). For g726r16, one person gave it an acceptability rating of 4. No other ratings were lower than a 5.

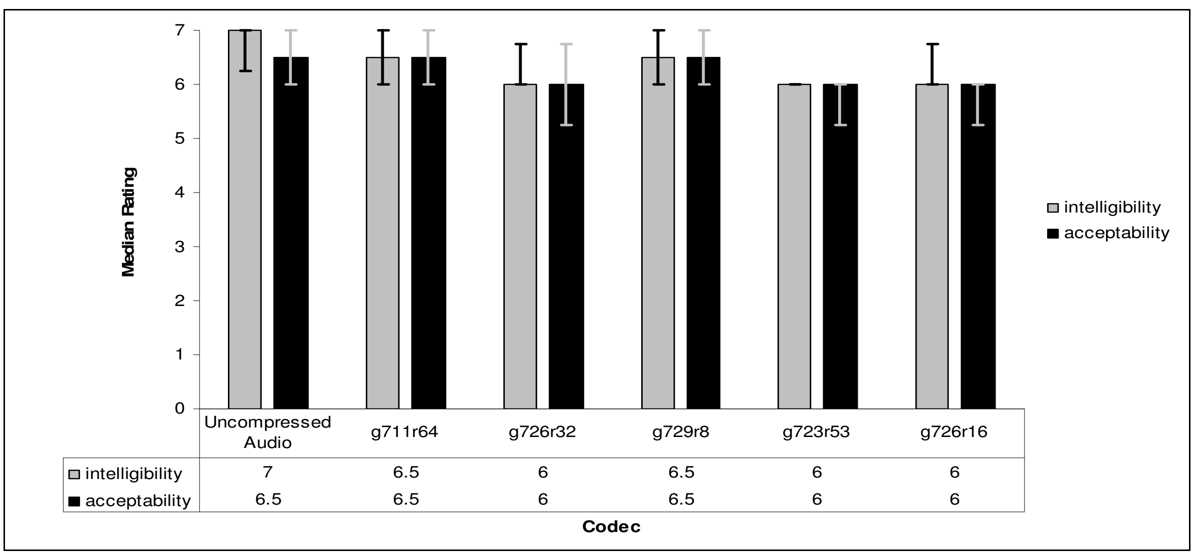

Figure 6. MCT median ratings for intelligibility and acceptability by codec type – bars indicate interquartile range.

The controllers did not provide much subjective feedback for the MCT. Uncompressed audio and codec g711r64 received only positive comments, whereas codec g726r32 received no comments. Codec g729r8 received both positive and negative comments, whereas codecs g723r53 and g726r16 received only negative comments (see Appendix E, Table E2).

4. CONCLUSION AND RECOMMENDATIONS

On the basis of this assessment, we recommend that the NVS program consider a further evaluation of the highest three performing codecs for VoIP. These three codecs, g711r64, g726r32, and g729r8, delivered the best performance and the shortest reaction times, and received the most positive feedback. As a caveat, although we include codec g729r8 on our list for further evaluation, the intelligibility and acceptability of that codec was not as high as the intelligibility and acceptability of codecs g711r64 and g726r32. This difference should be considered in any future evaluation.

We also recommend that the program office make speech intelligibility a top priority when evaluating performance and cost reduction trade-offs related to lowering the bitrate. For example, wideband codecs (e.g., G.722 and its variants) are becoming more popular because they may provide substantial improvements in speech intelligibility (Valin, 2008). Although standard VoIP codecs and the typical PSTN are capable of reproducing frequencies from about 300 Hz to about 3.2 kHz, wideband codecs can currently reproduce frequencies from about 20 Hz to about 7 kHz, and in the future, codecs may be able to reproduce frequencies up to 20–22 kHz. This is important because humans use the information contained in the higher and lower frequencies. They use higher frequencies (e.g., 4−18 kHz) to differentiate certain phonemes like *s* and *f* or *m* and *n*. Lower frequencies add a sense of presence to a speaker's voice (Rodman, 2006, 2008). Additionally, wideband codecs, by reproducing a wider range of frequencies, can help compensate for the effect of noise on speech intelligibility, reduce listener fatigue, and improve listener concentration. Even though we might expect wideband codecs to require higher bitrates, we found that this is not necessarily true. A codec that extends into the 7 kHz range might only require a bitrate of 10, 24, or 64. Because of the safety-critical nature of ATC, the NVS Program Office may want to evaluate the benefits of these wideband codecs and compare them to narrowband codecs before making a final selection.

For the currently evaluated codecs, we believe that the MRT paradigm could represent a worst-case scenario and we might see even better performance in an operational environment. In the MRT, words are spoken in isolation and there is no semantic, syntactic, or phonemic context to help the listener. In an operational environment, ATC phraseology provides controllers with contextual cues that are intentionally designed to avoid phonological ambiguity and help the controller correctly identify words. However, there are other factors in the operational environment, such as noise, unclear speech, static, and background conversations, which could potentially have a negative impact on controller performance.

As the program office moves closer to selecting a codec for operational use, we suggest that they expand their investigation to determine how channel noise, signal delay, jitter, packet loss, ambient noise, and varying levels of network utilization affect intelligibility (Sollenberger, McAnulty, & Kerns, 2003; Zingale, McAnulty, & Kerns, 2003).[3] Although these are important factors that can affect intelligibility, we did not have an opportunity to evaluate their impact in this assessment.

We also recommend that the NVS Program Office investigate the suitability of using commercially available algorithms to measure speech quality. Examples of these types of algorithms include, the Mean Opinion Score (MOS), the Perceptual Analysis Measurement System (PAMS), the Perceptual Evaluation of Speech Quality (PESQ), and the Perceptual Speech Quality Measurement (PSQM/PSQM+). The benefit of using metrics like the MOS,[4] PAMS, PESQ,[5] and PSQM/PSQM+[6] is that they can be measured using Commercial-Off-The-Shelf (COTS) tools (e.g., the Opticom OPERA[7] voice/audio quality analyzer and the PEXQ voice quality measurement software suite). Our ordering of four of the five codecs (i.e., g711r64, g726r32, g729r8, and g723r53) matches the ordering of the MOS scores found in other studies (Bhatia, Davidson, Kalidindi, Mukherjee, & Peters, 2006).

Because speech intelligibility and speech quality are not identical (Hicks, 2003; Rao, 2002; Storm, 2007), we recommend investigating the components of intelligibility that are not captured by objective speech quality measurements. These factors include signal tonality, signal annoyance, perceived signal noisiness, signal interference or masking, location of signal interference in the speech signal, signal distortion, and signal-to-noise ratio. Some objective metrics that exist for measuring speech intelligibility include the Speech Interference Level (L_{SIL} and SIL), the Articulation Index (AI), the Speech Intelligibility Index (SII), and the Speech Transmission Index (STI); see Hicks (2003) for a more thorough discussion of these metrics.

If an evaluation demonstrated the benefits of these COTS products, speech quality metrics, and speech intelligibility metrics, they would provide the FAA with a set of tools for measuring speech quality and intelligibility in a simple but reliable way. Field personnel could compare new codecs to old codecs or one bitrate setting to another without performing an MRT for every potential codec change. These products would also provide vendors with tools to use to evaluate compliance with intelligibility requirements. Vendors could perform most intelligibility requirement testing in house, without much equipment, and without performing more complicated and costly experimental testing. However, before we could recommend the use of these tools, we recommend performing a human factors assessment to determine their benefits and limitations and to compare their ratings to the ratings from an actual MRT.

---

[3] We recommend using a product such as iTrigeny Network Emulator to simulate factors like signal delay, jitter, and packet loss.

[4] See the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) P.800 and ITU-T P.800.1 for recommendations regarding the MOS.

[5] See ITU-T P.862 for recommendations regarding the PESQ.

[6] See ITU-T P.861 for recommendations regarding the PSQM and PSQM+.

[7] Additional information about the Opticom OPERA analyzer can be found at http://www.opticom.de

References

American National Standards Institute. (1996). *Method for measuring the intelligibility of speech over communication systems* (ANSI S3.2-1989). New York: ANSI.

Ballinger, G. A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods, 7*, 127-150.

Bhatia, M., Davidson, J., Kalidindi, S., Mukherjee, S., & Peters, J. (2006). *VoIP: An in-depth analysis*. In Voice over IP Fundamentals (2nd ed.). Indianapolis, IN: Cisco Press.

GAIN Working Group E: Flight Ops/ATC Ops Safety Information Sharing. (2004). *The other end of the radio: Identifying and overcoming common pilot/controller misconceptions*. Retrieved February 6, 2009, from http://www.flightsafety.org/gain/pilot_controller_miscon.pdf

Hanley, J. A., Negassa, A., Edwardes, M. D., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology, 157,* 364-375.

Hicks, D. A. (2003). *A comparison of speech intelligibility between the Call sign Acquisition Test and the Modified Rhyme Test*. Unpublished Master's thesis, Michigan Technological University.

International Telecommunication Union. (1996). *Methods for subjective determination of transmission quality* (ITU-T Rec. P .800).

International Telecommunication Union. (1998). *Objective quality measurement of telephone band (300-3400 Hz) speech codecs* (ITU-T Rec. P .861).

International Telecommunication Union. (2001). *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs* (ITU-T Rec. P .862).

International Telecommunication Union. (2006). *Mean opinion score (MOS) terminology* (ITU-T Rec. P .800.1).

La Due, J. L., Sollenberger, R. L., Belanger, B., & Heinze, A. (1997). *Human factors evaluation of vocoders for air traffic control environments. Phase I: Field evaluation* (DOT/FAA/CT-TN97/11). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

Rao, M. (2002). *Speech intelligibility of the Call-sign Acquisition Test (CAT) for Army communication systems* (TCN 02063). Aberdeen, MD: U.S. Army Research Laboratory, Human Research & Engineering Directorate.

Rodman, J. (2006). *The effect of bandwidth on speech intelligibility* [White paper]. Retrieved March 5, 2009, from http://www.polycom.com/global/documents/whitepapers/effect_of_bandwidth_on_speech_intelligibility_2.pdf

Rodman, J. (2008). *VoIP to 20kHz: Codec choices for high definition voice telephony* [White paper]. Retrieved March 5, 2009, from http://www.polycom.com/global/documents/whitepapers/codecs_white_paper.pdf

Sollenberger, R. L., La Due, J. L., Carver, B., & Heinze, A. (1997). *Human factors evaluation of vocoders for air traffic control (ATC) environments. Phase II: ATC simulation* (DOT/FAA/CT-TN97/25). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

Sollenberger, R. L., McAnulty, D. M., & Kerns, K. (2003). *The effect of voice communications latency in high density, communications-intensive airspace* (DOT/FAA/CT-TN03/04). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

Storm, A. (2007). *Speech quality investigation using PESQ in a simulated climax system for ATM*. Unpublished Master's thesis, Lulea University of Technology, Lulea, Sweden. Retrieved February 6, 2009, from http://epubl.luth.se/1402-1617/2007/260/LTU-EX-07260-SE.pdf

Valin, J. M. (2008). *The codec world widens* [Online editorial]. Retrieved January 13, 2009, from http://compactpci-systems.com/articles/id/?3697

Zingale, C. M., McAnulty, D. M., & Kerns, K. (2003). *The effect of voice communications latency in high density, communications-intensive airspace. Phase II: Flight deck perspective and comparison of analog and digital systems* (DOT/FAA/CT-TN04/02). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

Acronyms

ATC	Air Traffic Control
COTS	Commercial-Off-The-Shelf
CPC	Certified Professional Controller
DataComm	Data Communications
FAA	Federal Aviation Administration
FLM	Front Line Manager
GEE	Generalized Estimating Equation
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
MCT	Message Completion Task
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
NAS	National Airspace System
NextGen	Next Generation Air Transportation System
NVS	NAS Voice Switch
PAMS	Perceptual Analysis Measurement System
PESQ	Perceptual Evaluation of Speech Quality
PSQM/PSQM+	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
RDHFL	Research Development and Human Factors Laboratory
VoIP	Voice over Internet Protocol
WJHTC	William J. Hughes Technical Center

Appendix A

Statement of Ethics & Informed Consent

Voice Over Internet Protocol (VoIP) Speech Intelligibility Assessment

Statement of Ethics & Informed Consent

Nature and Purpose of Activity

Thank you for volunteering to participate in the Assessment of Speech Intelligibility of Voice over Internet Protocol (VoIP). This study will examine possible human factors effects of VoIP techniques that may be used in future air traffic control voice communications systems.

Who is Eligible to Participate?

Any FAA employee who is physically able to engage in voice communications over a telephone is eligible to participate.

Experimental Procedures

We will explain the procedures and your rights and responsibilities before you begin. The session will last 60 minutes. At the beginning of the study we will ask you to complete a background questionnaire. During the communication tasks we will ask you to listen to words and select the spoken word from a list on a visual display, or to speak words. We will ask you to complete questionnaires about your opinions regarding the quality of the voice signal. Researchers will observe all activities.

Discomforts and Risks

You will be able to adjust the headset so that it is comfortable for you. Even so, you may feel some minor discomfort on your ear or head from wearing the headset. If you feel you must remove or adjust the headset during the session, please inform the experimenter so that the session can be paused.

The audio volume will approximate the volume of a typical telephone. Before the session begins, you will be able to adjust the audio volume to your comfort level. Once data collection begins, however, you will not be able to adjust the audio volume further. If the volume of the audio becomes uncomfortable for you during the session, please inform the experimenter as soon as possible so that the session can be paused and the problem addressed.

Benefits

Participation in this study provides no direct benefit to you. Your participation benefits the FAA and its workforce in that the data will be used to develop requirements for the National Airspace System (NVS) Voice Switch and related systems.

Participant's Responsibilities

The results of this effort depend greatly upon your attention to the required tasks, and upon forthright responses to the questionnaire. If there is something you do not understand, please ask a researcher. In addition, to avoid biasing the results, please do not discuss the study with other potential participants until the study is completed in about 30 days.

<u>Participant's Assurances</u>

Researchers from the Human Factors Team–Atlantic City maintain strict standards regarding participant confidentiality and informed consent in all our activities. Our standards are based on the Ethical Principles in the Conduct of Research with Human Participants by the American Psychological Association and FAA Order 9500/25. The standards are structured around four main principles:

- Your participation in this study is completely voluntary. You may withdraw from this assessment at any time without consequence. If you feel you must withdraw at any time, for whatever reason, please inform a researcher immediately. There is no penalty whatsoever for ending the session. In addition, the researchers may terminate your participation if they believe this to be in your best interest.
- Your responsibilities will be clear. The researchers will explain everything to you and answer all your questions.
- Your identity will be kept anonymous. Your responses will be identified by a code known only to you and the researchers conducting the assessment. Your identity will be kept separate from the data you provide. To facilitate this, please do not write your name or any other identifying marks on the questionnaires. Please do not share your participant code with anyone other than the researchers. Your name will not be associated with any data contained in any report or briefing.
- The data you provide will be kept confidential. The *raw* data collected in this assessment will become the property of the Human Factors Team–Atlantic City. The raw data will be analyzed by specialists from this organization and its contractor employees. The raw data will not be made available to other organizations without your permission. The *aggregate* data from this assessment will be presented in briefings and reports made by the group. These data will take the form of averages, standard deviations, and other similar statistics.
- Results – The Study Report will be provided upon request.

If you any have questions about this study or need to report any adverse conditions you may contact the researchers conducting the study, the Primary Investigators Kenneth Allendoerfer or Dr. Ferne Friedman-Berg. You may also contact Dr. Earl Stein (609) 485-6389, the Human Factors Team – Atlantic City Manager, at any time with questions or concerns.

I have read this consent document. I understand its contents, and I freely consent to participate in this study under the conditions described. I have received a copy of this consent form.

Research Participant: ______________________ Date: _____________

Investigator: ______________________________ Date: _____________

Witness: _________________________________ Date: _____________

Appendix B

Background Questionnaire

VOIP study

Background Questionnaire

Participant ID	SS24	Setup
Experimenter ID	FF	
Study day	1	

Instructions: The purpose of this questionnaire is to gather background information about the individuals who participate in this research. This questionnaire complies with the rules outlined in the Statement of Ethics and Informed Consent.

1. Do you have any medical or physical condition that might make it difficult for you to use a telephone for 90 minutes? Yes ○ No ○

2. Do you have any medical or physical condition where you use a hearing aid or other assistive device when using a telephone? Yes ○ No ○

3. What is your gender? Male ○ Female ○

4. What is your age? ______ years

5. How many years have you worked as a Certified Professional Controller (or equivalent) or Front Line Manager (or equivalent) in the following air traffic control domains:

   En Route: ______

   Approach control: ______

   Tower: ______

   TMU: ______

   ATC System Command Center: ______

6. What is your current job title? ______

7. Is there anything else we should know about your participation in this study? ______

Appendix C

Participant Feedback Survey

VOIP study

Feedback Questionnaire

Participant ID	SS24	Setup
Experimenter ID	ff	
Codec	4	

Instructions: For the condition you just finished, complete the following rating scales. ***Intelligibility*** refers to how well you were able to understand what was said in this condition. ***Acceptability*** refers to how appropriate the audio in this condition would be for you to use for multiple hours each day as part of your job.

	1	2	3	4	5	6	7
1 INTELLIGIBILITY	I could understand ***nothing*** that was said	I could understand ***nearly nothing*** that was said	I could understand ***a little*** of what was said	I could understand ***about half*** of was said	I could understand ***most*** of what was said	I could understand ***nearly everything*** that was said	I could understand ***everything*** that was said
	O	O	O	O	O	O	O

2 ACCEPTABILITY	In ***all*** foreseeable situation, the audio would be ***unsatisfactory***	In ***nearly all*** foreseeable situations, the audio would be ***unsatisfactory***	In ***most*** foreseeable situations, the audio would be ***unsatisfactory***	In ***about half*** of foreseeable situations, the audio would be ***satisfactory***	In ***most*** foreseeable situations, the audio would be ***satisfactory***	In ***nearly all*** foreseeable situations, the audio would be ***satisfactory***	In ***all*** foreseeable situation, the audio would be ***satisfactory***
	O	O	O	O	O	O	O

3 Please describe any POSITIVE characteristics of the audio that you experienced during this condition.

4 Please describe any NEGATIVE characteristics of the audio that you experienced during this condition.

Done

Appendix D

Message Completion Task

Voice Over Internet Protocol (VoIP) Speech Intelligibility Assessment

Message Completion Task

Participant Code: _____   C: 1 2 3 4 5 6

Instructions: You will hear a phrase number followed by a short voice message either from a terminal or en route environment. For advisories to airmen, you are asked to play the role of the pilot in command. Each audio message you hear will correspond to a numbered phrase. The phrases will not be presented in the order in which they appear on the page. Using the spoken phrase number, you are asked to go to that phrase and print legibly in the space provided the missing parts which will make the statement complete. You may use any standard controller abbreviations. If you cannot understand a part of the message either guess or leave that space in the statement blank.

Phrase 1

SOUTHWEST 127, DESCEND AND MAINTAIN FLIGHT LEVEL ONE NINER ZERO, TRAFFIC ____________ O'CLOCK AND ____________ MILES SOUTHWEST BOUND, A BOEING 737, AT FLIGHT LEVEL ____________

Phrase 2

AMERICAN ____________, TURN LEFT HEADING ____________, CLIMB AND MAINTAIN ____________

Phrase 3

____________ 2341, SQUAWK ____________, EXPECT HIGHER IN ____________ MINUTES

Phrase 4

DENVER THIRTY SIX, LOS ANGELES ______, ON THE ONE LINE, APREQ ("ap-prek") DELTA THREE TWENTY FLIGHT LEVEL ________

Phrase 5

R SEVEN, R ______, REQUEST CONTROL __________ FOR RIGHT TURNS

Appendix E

Subjective Feedback

Table E1. Summary of MRT Subjective Feedback by Codec Condition: Pros and Cons

Codec	Summary of Subjective Feedback
Uncompressed Audio	PROS: All was clear. Nothing negative. Words easy to understand. Very clear. This session was much more clear and easy to understand than others. Clear as a whistle. No problems hearing or understanding any words. Very clear! Perfect! No distortions. Sound quality was good. Very understandable. Mostly clear and distinct.
g711r64	PROS: Mostly clear but not as clear as others. It was fairly easy to understand. CONS: Some words more difficult to understand than others, particularly the first consonant…Clipping? Some difficulty (distinguishing) between “s” and “f”. Some consonant (clusters) were too similar to others. On words not understood… it was the beginning of the word being clipped that caused confusion. Hard to distinguish between “m” and “n” at the beginning and end of words. Some beginning and ending sounds perceived as garbled or muffled.
g726r32	PROS: Slightly easier to distinguish initial/trailing consonants. No problems at all. Clearer than other trials. Very good codec. Easy listening. Transmissions were clear and distinct. CONS: Less effective in clearly compressing the audio than others. There were some words that were a little difficult to discern. A bit of clipping at the start of words. Some word-initial clipping. Some sounds (can't pinpoint which) were hard to differentiate. Some consonant (clusters) were too similar to others. Had some trouble distinguishing words that ended with “m” vs. “n”. Had slight trouble when the base of the word was the same such as boil, oil.
g729r8	PROS: Clearer than some, not as clear as others. Much of the spoken text was clear and easily understood. Surprisingly good intelligibility given the negatives. Best one so far I say 6.5 in intelligibility and acceptability. In a normal conversation, I think it would sound ok. CONS: Somewhat watery sound, some buzz. Some difficulty in hearing the leading consonant for a couple of words. Could not hear the beginning of a few words. Some words or consonant clusters sound sounded similar. Had trouble distinguishing between the “b” and “d” sound at the beginning of words. Had trouble with the “th” sound, sounded like an "s" sound. Some of the first consonants were muffled.

Codec	Summary of Subjective Feedback
g723r53	PROS: I think I understood all the words. Relatively clear for certain words. It was easy to understand what was being said. CONS: These words were slightly more difficult to understand. It seemed that part of the audio was clearer than other parts. Could not quite tell errors. The codec did degrade considerably on some words, however others were very clear. Not easy to understand. Transmissions of certain words were fragmented or garbled. This one was a little more difficult to understand, but still clear enough to understand nearly all of the spoken words. Slightly more difficult. Mostly the initial, but sometimes the trailing consonant was hard to understand. Words ending in “ig” were somewhat difficult. Intelligibility of some words ending in "ig". When the base of the word is the same (i.e., *ig*), it is often hard to pick out the first letter or sound to distinguish the word. Leading consonants sometimes hard to understand. Word-initial consonants were clipped (especially fricatives). “S” and “f” were difficult to differentiate. Consonants sounded very similar, static, etc. Some words clipped at beginning and hard to understand.
g726r16	PROS: Eighty percent was OK. Though I could not understand all of the words, I think the sound is acceptable to listen to all day. Reasonably clear. Most words I understood. Nothing distracting to the point that it would bother me. I could understand most of what was said. Quality was ok. CONS: This codec distorted all the words to some degree. Did not like that one. I had a hard time understanding some of the words. The audio signal had more interference than the other conditions. I had to guess at one or two. Seem to have trouble understanding some of the words. Some sounds getting more fragmented. It was quite hard to understand, and a lot of guessing was used. Some phonetically similar sounds merged together. “s” hard to hear. There was some clipping in the front of some of the words, which made the words unclear. The leading consonants seemed the hardest to hear. I had trouble discerning between “m” and “n”. It was hard to understand the sounds. Consonants sounded similar like “p”, “t”, etc. Most unsure words were due to last part of word being unclear, although a few unsure words were due to first part of word being clipped.

Table E2. Summary of MCT Subjective Feedback by Codec Condition: Pros and Cons

Codec	Summary of Subjective Feedback
Uncompressed Audio	PROS: This speech was very good and usable. Clearest scenario, but no side tone which most controllers like to hear (at least a light scratch).
g711r64	PROS: All speech was very clear. On one call sign I could understand the numbers but not the call sign. I don't know if it was pronunciation or a call sign I'm not familiar with. Very clear.
g726r32	N/A
g729r8	PROS: Very clear and 'bright' speech. CONS: Least intelligible of all scenarios.
g723r53	CONS: Some speech hard to understand, but the context helped to know what it was. Muffled and garbled with some phrases but the altitude, heading, and call sign were all generally clear. Scratchy background noise during the transmission.
g726r16	CONS: A lot of white noise imbedded in the speech. Audible and discernable but 'scratchy'.