Speech-based Natural Language Interface for UAV Trajectory Generation

Erica L. Meszaros[1], Meghan Chandarana[2], Anna Trujillo[3], and B. Danette Allen[3]

Abstract— **In recent years, natural language machine interfaces have become increasingly common. These interfaces allow for more intuitive communication with machines, reducing the complexity of interacting with these systems and enabling their use by non-expert users. Most of these natural language interfaces rely on speech, including such well-known devices as the iPhone's Siri application, Cortana, Amazon's Alexa and Echo devices, and others. Given their intuitive functionality, natural language interfaces have also been investigated as a method for controlling unmanned aerial vehicles (UAVs), allowing non-subject matter experts to use these tools in their scientific pursuits. This paper examines a speech-based natural language interface for defining UAV trajectories. To determine the efficacy of this interface, a user study is also presented that examines how users perform with this interface compared to a traditional mouse-based interface. The results of the user study are described in order to show how accurately users were able to define trajectories as well as user preference for using the speech-based system both before and after participating in the user study. Additional data are presented on whether users had previous experience with speech-based interfaces and how long they spent training with the interface before participating in the study. The user study demonstrates the potential of speech-based interfaces for UAV trajectory generation and suggests methods for future improvement and incorporation of natural language interfaces for UAV pilots.**

I. Introduction

Unmanned aerial vehicles (UAVs) are becoming increasingly ubiquitous devices in modern society. With recent advancements in UAV design and increasing availability, these systems have been given new roles as delivery mechanisms, hobby devices, and even tools for carrying out scientific investigations [1]. The power of UAVs makes them an appealing tool for carrying out repetitive but demanding tasks, while their ability to operate in enclosed spaces makes them significant assets to warehouse management and even product delivery [2]. Furthermore, UAVs enable researchers to place instruments, gather data, and interact with environments that would otherwise be prevented. This includes accessing previously inaccessible locations, such as beneath the forest canopy or high up in the Earth's atmosphere, as well as potentially dangerous locations, such as near active volcanos [1]. While UAVs themselves have become more popular and capable, the interfaces used to operate them have remained complex, often precluding their use by any but the highly trained. In fact, the extensive training required to be able to successfully operate UAVs has restricted their use to subject matter experts with many hours spent training and operating these systems [3]. Reducing the complexity of UAV interfaces would enable greater use of these powerful tools by a wider subset of the population and for an increased number of tasks. Specifically, a reduction in the complexity of the interface would enable UAVs to be used more extensively in scientific fields.

One way to reduce the complexity of UAV interfaces is to rely on intuitive methods of interaction between the user and the machine such as natural language. Natural language interfaces make use of human speech and gesture patterns to interact with non-human systems. Because users are accustomed to interacting using vocal commands and information sharing, utilizing this already familiar tool to allow users to interact with UAVs reduces the complexity of interaction [4]. By reducing this complexity, UAVs become more accessible and therefore more widely applicable as tools.

This study presents an initial speech-based natural language interface specifically designed for inputting sections of a UAV flight path (Section III). In order to determine whether this speech-based interface is accepted by users, it is compared to a traditional mouse-based interface. The set-up of this experiment and design of the mouse-based interface is presented in Section IV. The results of this comparative study are presented in Section V, focusing on overall success, training time, input time, and user preference. A discussion of these results follows in Section VI, as well as conclusion and discussion of future work in Section VII. An initial analysis of this interface appears in [5], however this paper expands upon the speech-based interface and includes additional analysis of the effects of previous speech interface use, training time, and success.

II. Related Work

The most common speech-based natural language interfaces are currently found in smartphones and other smart home devices. Recent research such as Ruan et al.'s suggests that speech interfaces are not just novel or convenient but more efficient for text entry and operation of smartphones [6]. Kojima et al.'s research indicates that speech recognition interfaces in cars result in increased usability and satisfaction as well [7]. Given the widespread use and success of such speech interfaces, they have also been investigated for use

[1] This author is with the Department of Social Science, University of Chicago, Chicago, Illinois, USA. Email: elmeszaros@uchicago.edu.

[2] This author is with the Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Email: mchandar@cmu.edu.

[3] These authors are with National Aeronautics and Space Administration's Langley Research Center, Hampton, Virgina, USA. Email: {a.c.trujillo, danette.allen}@nasa.gov.

in human-robot interaction. In a meta-analysis of speech interfaces for swarm control, Hocraffer and Nam indicate that a speech interface can help to reduce the workload of the human operator and increase situation awareness [8]. Novitzky et al. examine how a speech interface can be utilized in a marine robot to improve team dynamics and performance [9]. Some research has even explicitely looked into using speech interfaces to control UAVs, including Peshkkova et al. [4], Ferreiros et al. [10], and Williamson et al. [11]. However, these studies focus on how to replicate expert control systems that are currently in use. Limited research has been carried out on utilizing simple speech-based natural language interfaces to extend UAV usability beyond its traditional scope.

Understanding current research and limitations of natural language processing is an integral part of incorporating speech interaction in UAV interfaces. Current trends in deep learning [12] and neural networks [13] have propelled the success of speech recognition and enabled speech-based natural language interfaces. Such powerful language processing comes at a cost; researchers have noted potential problems with memory, time and energy consumption, and necessary power that reduce the usability of interfaces reliant on these techniques [14]. Ensuring that a speech-based natural language interface is accurate while minimizing its set of power, time, energy, and memory needs may be crucial to allow such interfaces to be accepted. The speech interface examined in this paper is designed to be simple to reduce such needs, while still powerful and accurate enough to allow non-expert users to interact with UAVs to generate flight paths.

III. Speech Interface Design

The speech interface presented in this study allows users to create flight path trajectories for UAVs using vocal commands. Users are presented with a list of twelve available trajectory segments (Figure 1), developed by Chandarana et al. [1]. These segments are presented in the form of a dropdown menu (Figure 2). Speaking the name of any trajectory segment selects that specified segment and adds it to the total flight path. This selection is confirmed by showing the user an image of the selected trajectory segment for three seconds, after which they are given a dialogue box asking if they would like to add an additional trajectory segment to the overall flight path. The user can then respond verbally with either a "yes," continuing to add segments until the desired flight path is completed, or a "no," completing the flight path (Figure 3). The speech interface presented here is a simple version that does not make use of many of the current areas of research focus in natural language processing, including deep learning and neural networks. However, the simplicity of the task favors speed and responsiveness of the overall interface over power in speech recognition; the limited vocabulary of the trajectory segment library and commands of the interface require little parsing effort in order to allow the system to understand.

The speech interface tested in this study relies on the CMU-Sphinx speech-to-text software [15]. A product of more than 20 years of continuous improvement, CMU-Sphinx is an open source tool produced at Carnegie Mellon University. It utilizes a pre-fabricated and standard dictionary of phones and lexicon mapping of phone-groups to words. In this case, the English lexicon was used, mapping phones to English phonemes and groups of these phonemes into English words. The full English dictionary available to the CMU-Sphinx application contains thousands of words, but searching through a dictionary of that size is costly in processing power and time. In order to speed up the speech-recognition, an application-specific dictionary was created for this user study that instructs the system to listen only for the specific words users would use in order to define UAV flight paths. This small dictionary consisted of a little less than 100 words, corresponding directly to the possible flight path trajectories. Each of these words is stored in traditional English orthography, so the word "up" is stored spelled with a "u" followed by a "p." The traditional orthography is then mapped to a CMU-Sphinx specific pronunciation orthography that defines the sounds that compose the word. For the example word "up," CMU-Sphinx would map it to "AH P", with "AH" representing the sound that the "u" makes. In the event that a word has multiple pronunciations, CMU-Sphinx can store multiple pronunciations, accounting for variability in vowel sounds, pronounced consonants, and even extra syllables. By storing multiple possible pronunciations for each word in the dictionary, CMU-Sphinx allows for more successful word identification even with accent variances, providing a powerful tool for speech identification.

In addition to the application specific dictionary, an application specific grammar was created. This defined word combinations that were allowed, specifically the compound diagonal trajectories "forward right", "forward left", "backward right", and "backward left" (Figure 1). It also indicated that variations on "yes" such as "yeah" and "yup" should be interpreted as "yes," and variations on "no" (including "nope" or "nah") should be interpreted as "no." This further specifies what type of language the CMU Sphinx system should be looking for, and improves accuracy in speech identification as a result. Furthermore, the speech system is trained to ignore non-speech sounds, such as a cough into the microphone, or non-content bearing words, such as "um" or "uh."

Users accessed this speech interface using an Audio Technica microphone headset [16]. This microphone sat on the temples, wrapping around the back of the head, and provided no audio output or any part that would cover the ears. The microphone itself extended downward toward the mouth (Figure 4), and users were given time and instruction in how to adjust the angle for comfort.

IV. Methods

The experimental setup utilized here has been previously described [5]. The speech interface is compared to a simple yet familiar mouse interface in order to determine usability

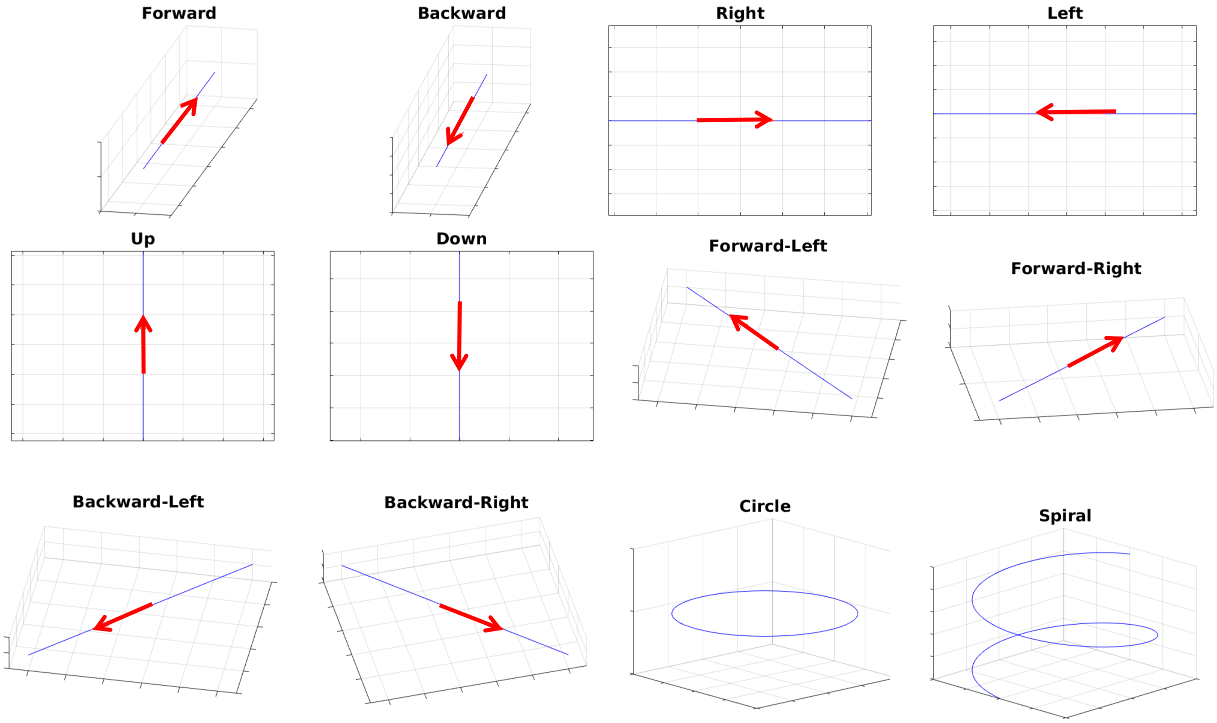


Fig. 1: Gesture library of 12 trajectory segments developed by Chandarana et al. [1].

and user preference. A broad level description of how this mouse interface works is outlined in this section. Additionally, the methodology used to conduct the user studies is outlined as well.

A. Mouse Interface

The mouse interface is deliberately designed to look like the speech interface, such that the only major difference between the two is not in appearance but in how they are operated. As with the speech interface, the mouse interface presents users with a drop-down menu listing all 12 trajectory options (Figure 2). Users select an option from the drop-down menu and their selection is confirmed by displaying an image of their selected trajectory segment. Afterwards, they are presented with a dialog box asking if they would like to add another segment (Figure 3). Users select "yes" to add additional segments to the flight path, or "no" to stop adding segments and generate the final flight path. If a user selects "yes," they return to the drop-down menu and can select another segment. If they select "no," they are shown the final flight path. Users were presented with a standard mouse to operate this interface: the mouse was attached to the computer via USB cord and contained two buttons and one wheel. Users were instructed to operate the mouse with their right hand in order to ensure standardization across the study.

B. Experimental Setup

Fourteen users participated in this study. All of the users were asked to generate flight paths with both the mouse and speech interfaces, but the order in which they used each interface was randomized and counterbalanced across all subjects. Users generated all three flight paths (Figure 5), but were asked to create them in a randomly assigned order that was counterbalanced across all subjects. For each subject, the same order was used for flight path generation in both interfaces.

Before beginning trials, subjcts were asked to read and sign the Privacy Act Notice and Informed Consent Forms, after which the researcher(s) would outline the purpose of the user study and describe the participation process. Users then completed the background questionnaire and began training on the first interface. Users were given a maximum of ten minutes to train on the interface, and were provided with a printout of the library of 12 trajectory segments (Figure 1). Users were allowed to retain this printout during the actual data collection runs. The amount of time that the user chose to train on each interface was recorded. After the user indicated that they were ready to proceed to trial runs, they were presented with a printout defining the flight path they would be asked to generate (Figure 5). These printouts specified an entire flight path consisting of three unique trajectory segments that the user must define in order. Each segment was presented visually and also included a numbered label to ensure correct interpretation of the flight path by the user. Users were able to study the flight path for five seconds before the researcher(s) began the test run, but were allowed to retain this printout and view it during the entire test run. While the user completed three test runs with three different flight paths, data were collected on

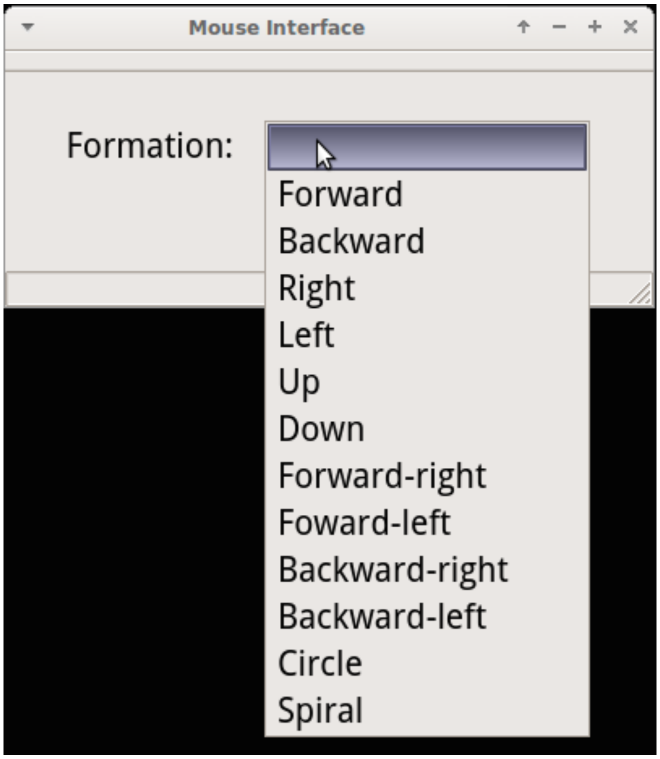
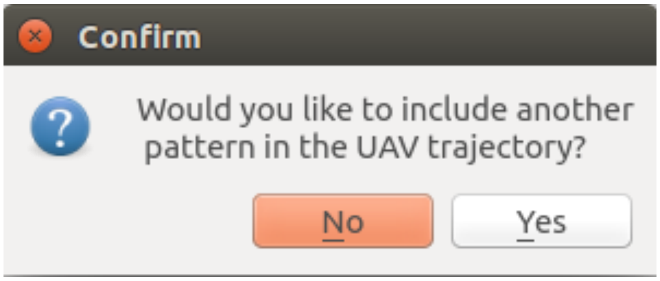


Fig. 2: Interface drop-down menu listing trajectory segments. The same drop-down menu was used for the speech and mouse interfaces.



Would you like to include another pattern in the UAV trajectory?

Fig. 3: Interface dialog box for adding an extra segment. This same dialog box was used in both the speech and mouse interfaces.

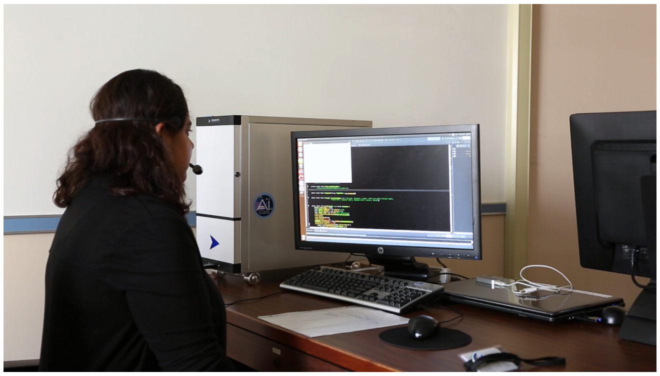


Fig. 4: Speech interface set-up.

correctness of each flight path segment, overall correctness of the entire flight path, and the time it took to complete the generation of each flight path. The user was also asked to complete a NASA TLX workload assessment and subjective questionnaire after completing all three flight paths. The user was then presented with the second interface, repeating the process of training and generating three flight paths. The same data were collected for each interface.

An entire flight path was classified as correct if each of the three contained segments was correct and no additional segments were added. Five different error types were recorded throughout the study, described below.

1) System Misinterpretation: A user provided the correct information to the system, but the system misinterpreted the information and produced the wrong trajectory segment.
2) Extra Segment: A user provided an extra segment to the entire flight path, resulting in a flight path of four or more trajectory segments instead of one of three segments.
3) Human Error: The user provided the wrong information to the system.
4) System Misinterpretation + Human Error: The user provided the wrong information to the system and the system misinterpreted the information that the user provided (type 1 + type 3).
5) System Misinterpretation + Extra Segment: The user provided an extra segment to the entire flight path and the system misinterpreted the information that the user provided (type 1 + type 2).

Error types 2 and 5 only appear for overall flight path correctness and do not appear at the level of individual trajectory segments. By tracing not only the occurence rate of errors but the type of errors, we can see how well the system is performing and how well the user is interacting with the system.

V. Results

Analyses were conducted using IBM SPSS version 24, focusing on an analysis of variance (ANOVA). Independent variables included input interface (speech vs. mouse), previous experience with UAVs, previous experience with speech-based interfaces, training time, and flight path. The results shown here reflect the impact of these independent variables on the number and type of error segments, the overall accuracy of the flight path, the input time for each flight path, as well as a number of subjective workload assessment variables measured by the NASA TLX, including mental demand, physical demand, temporal demand, performance, effort, and frustration. While no results are significant, the trends are presented here. This lack of statistical significance suggests that the speech interface worked as well as the baseline mouse interface. Where the relationship between independent and dependent variables were continuous and related, linear regression analysis was applied to demonstrate the direction and rate of correlation. Where appropriate, error

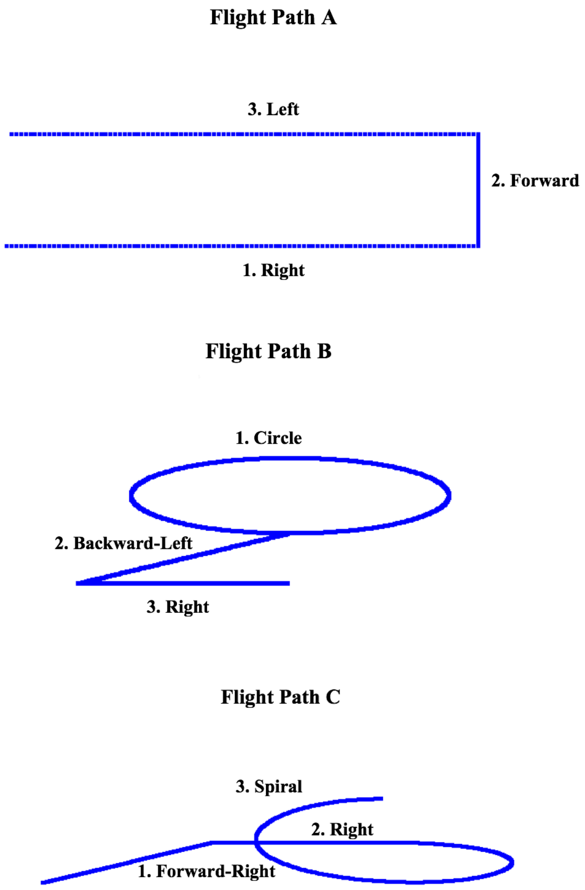


Fig. 5: Three flight paths that users were asked to create using speech and mouse interfaces. Each flight path consists of three segments, and each includes one "right" segment for comparison [5].

bars indicating standard error and mean are provided with graphs.

Of the 14 participants, only one user had any previously experience flying UAVs, and they had been doing so for roughly 4 years. Ten of the participants had previously used speech-based interfaces, predominately interacting with their smart phones and applications in their cars. Nine of these were satisfied with their interactions with these current interfaces.

A. Flight Path Accuracy

Overall, users met with success in creating flight paths using the speech interface. With both the mouse and the speech interfaces, users created flight paths with a 90.5% correctness rate (Table I). Most importantly, users were equally successful with the novel speech interface as they were with the familiar mouse interface. Two out of the 14 users made mistakes on any portion of the flight paths using the mouse based interface, while 10 out of 14 made mistakes using the speech interface (Figure 6). However, the speech system misinterpreted very few commands, and the majority of errors occurred on only flight path B (Figure 7). This likely occurred due to the inclusion of the compound segment "Forward Back" in Flight path B, and indeed most of the mistakes occurred on this segment (Figure 8). Interestingly, while this segment proved problematic for the speech interface, Flight Path B led to the highest performance for the mouse interface (Figure 9). Otherwise, performance was relatively standard between users and across flight paths. All but three users met with a greater than 80% success rate using the speech interface.

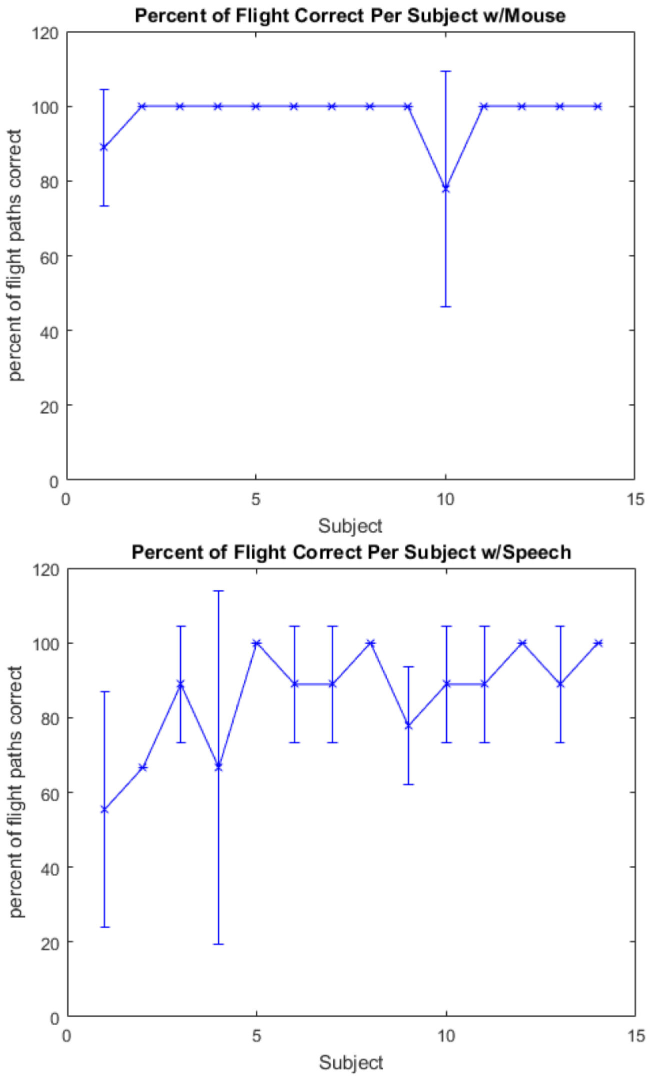


Fig. 6: Percent of correct flight paths based on subject for speech and mouse interfaces.

Interestingly, users performed the best with the mouse interface on the same Flight Path B that resulted in the misinterpretation errors with the speech interface (Figure 9). With this mouse interface, users performed the worst with Flight Path A, which met with the highest accuracy from the speech interface.

TABLE I: Segment definition errors by type for both mouse and speech interfaces

	Mouse	Speech
Misinterpret	0%	2.38%
Extra Segment	4.76%	0%
Human Error	4.76%	7.14%
Human + Misinterpret	0%	0%
Extra + Misinterpret	0%	0%
Overall Correct	90.48%	90.48%

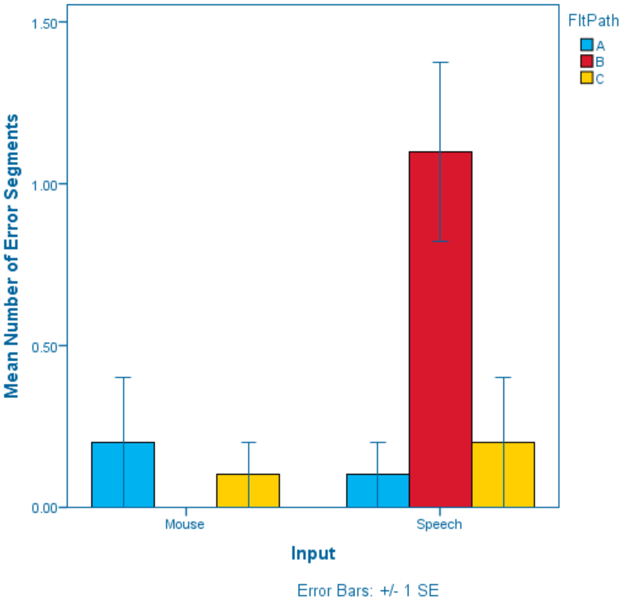


Fig. 7: Mean number of error segments per flight path segment for both speech and mouse interfaces.

B. *Performance Based on Training Time*

Training time was comparable for both speech and mouse interfaces. While users were offered the opportunity to take up to ten minutes to train with each interface, users took on average just over 130 seconds, or slightly more than two minutes of the allotted ten minutes of training time. This was similar in length to the time users needed to train with the mouse interface, an interface with which every user admitted familiarity. This indicates that no significant extension of training time is necessary when switching to a speech-based interface.

Performance improved as training time increased for both speech and mouse interfaces (Figure 10). The increased rate of improvement seen in the speech interface suggests that users improve at a faster rate than on the traditional mouse interface, indicating that with continued practice and more frequent use of this novel interface type users would become even more proficient.

Interestingly, there is a slight downward trend in overall performance, measured by an increase in the number of error segments, as training time increases for both the mouse and speech-based interface (Figure 12). This may suggest a type of burn-out, with more time spent with each interface leading to the production of more errors, or it may suggest that those who felt the least confident with each interface desired more training time and produced more errors. Because the effect was measurable in both the speech and mouse interfaces, however, it seems unlikely that this would lead users to select one interface over another.

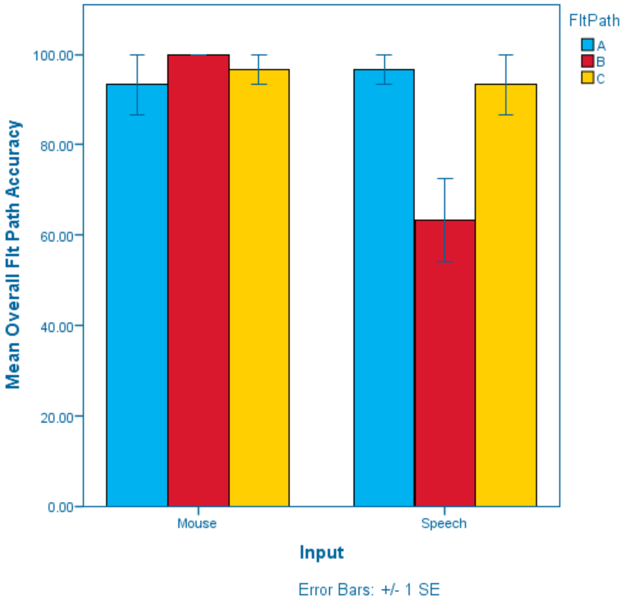


Fig. 8: Mean overall flightpath accuracy based on flightpath segment for both speech and mouse interfaces.

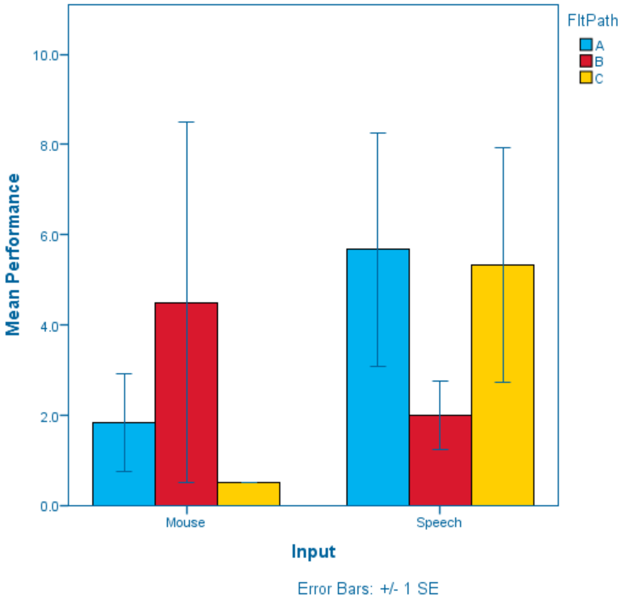


Fig. 9: Mean performance per flight path segment for both speech and mouse interfaces

Users also seemed to require increased time to input their flight paths as the amount of time they spent training increased (Figure 11). This pattern is visible for both the speech and mouse interface, but is substantially heightened for users of the mouse interface. This suggests that there is

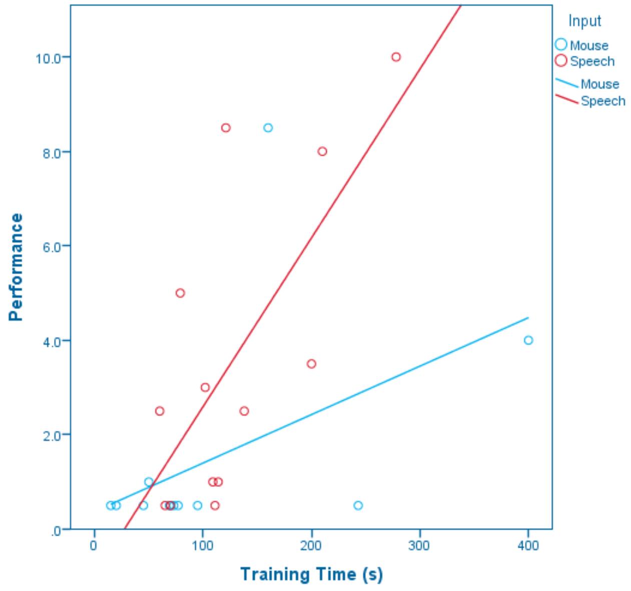


Fig. 10: Average performance success based on training time for both mouse and speech interfaces

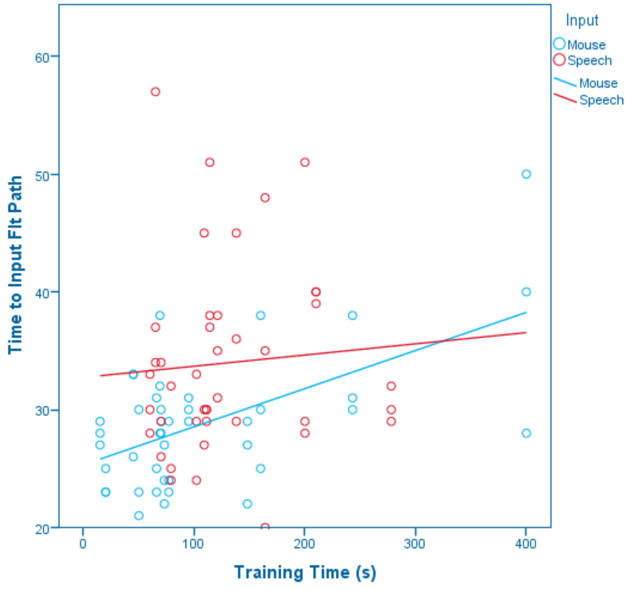


Fig. 11: Time to input flight path based on training time for both speech and mouse interfaces, including regression to average.

a correlation between the amount of time taken to train on an interface and the amount of time taken to input the flight path, perhaps out of worry of making mistakes or care taken to input the correct information.

C. Previous Experience and Overall Opinion

It is worthwhile to note that many participants had previous experience using speech interfaces. Thanks to the ubiquity of interfaces such as Siri and Cortana, and recent advances in aids such as the Amazon Echo and Google Home devices, many users are familiar with how speech interfaces operate. The effect that previous exposure has on user acceptance of this type of interface is therefore important to examine. To do so, the relationship between previous experience, as noted in the pre-test questionnaire, and overall satisfaction, as noted in the NASA-TLX form, are examined. In addition, user comments from both these forms are included.

The subjective questionnaire that users completed after using each interface asked them to rate the overall difficulty and responsiveness, indicate how likely they would be to use the interface again, how sufficient the training time was, and how sufficient the time to view the flight path before beginning the trial was. Users rated these values on a likert scale from 1 to 5, were 1 indicated that the interface was easy, too fast, not likely to use the interface again, too little training time, and too little viewing time (respectively). A 5 indicates that the interface was difficult, too slow, users were very likely to use it again, too much training time, and too much viewing time (respectively). After using both interfaces and completing both subjective questionnaires, users were asked whether they preferred the speech based interface or the mouse based interface overall. A rating of 1 indicated a preference for the mouse based interface and a rating of 5 indicated a preference for the speech based interface.

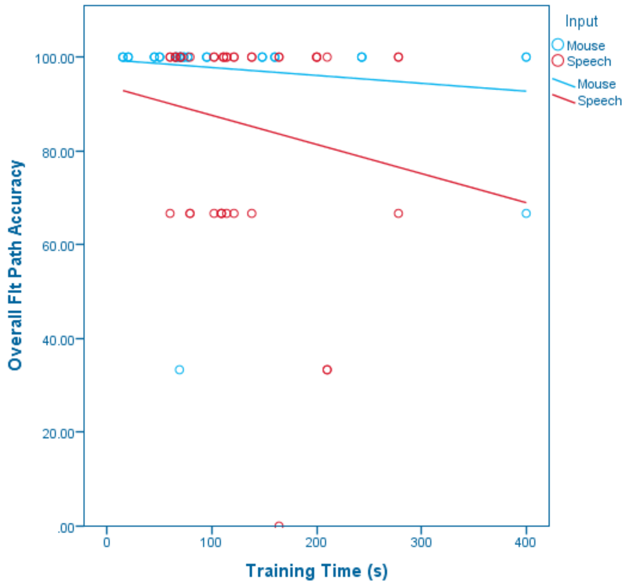


Fig. 12: Overall flight path accuracy based on training time for both speech and mouse interfaces, including regression to average.

Table II shows that users indicated a lower overall difficulty score for the mouse interface than for the speech interface. Both interfaces, however, were rated within one full point of each other and below the median score of 3. Both speech and mouse interfaces were rated similarly close in responsiveness as well. Both interfaces were rated slightly below the median score, indicating that both were viewed by users as slightly too slow. Users indicated that they would

TABLE II: Subjective Questionnaire Values for Speech and Mouse Interfaces

	Speech Total	Mouse Total
Difficulty	2.11	1.43
Responsiveness	2.36	2.79
Likely Use Again	3.43	4.14
Sufficient Practice Time	3.07	3.07
Sufficient Flight Path Study Time	2.92	2.93

be willing to use both interfaces again, with each being rated above the median line. Users did indicate, however, that they would be more likely to use the mouse interface than the speech interface. Users rated both interfaces identically when considering whether they were given enough time to practice with each interface, and nearly identically when asked if they were given enough time to study the flight path before being asked to create it.

The NASA TLX form also allowed for the collection of data on user's workload. Users were asked to rate the mental, physical, and temporal demands that both the speech and mouse interfaced required. Users rated these demands on a scale from 0 to 10, with low numbers corresponding to a low demand and high numbers to a high demand. Users also rated their overall performance on the interface from 0 to 10, with a low score indicating good performance and a high score indicating poor performance. Finally, users rated the amount of effort they put in to operate the interface and their level of frustration on a scale from 0 to 10. Lower measures corresponded to lower levels of effort and frustration, while high measures indicated higher levels. Table III shows the TLX measures for the speech interface. It also shows the TLX measures for those users who had previous experience with speech based interfaces and those users who were satisfied with previous speech based interfaces.

Overall, users indicated very low mental, physical, and temporal demand, as well as low levels of effort and frustration for the speech interface. The physical demand of the speech interface was the lowest, with an average measure below 1, while the mental demand required was the highest, but still only slightly above 2.5. The amount of effort was similarly rated an average of just above 2.5, while users indicated a frustration level averaging just under 3.3. Generally, users indicated lower TLX ratings for the mouse interface than the speech interface. However, the low demand, effort, and frustration values given to the speech interface demonstrate that this novel interface was similarly acceptable to users. Users also indicated that they performed fairly well with the speech interface, rating their performance on average at 3.32. Interestingly, users with previous experience with a speech interface suggested that the interface required less demand and effort and produced less frustration, but indicated that their performance was slightly worse than the overall average with a score of 3.9 compared to the average of 3.32. Moreover, users who had previously used a speech interface and were satisfied with how it worked rated this speech interface as requiring less demand and effort and producing less frustration than the overall average, but also viewed their performance with the interface as better than the average at 3.22 compared to 3.32.

D. Subjective Information

Users were also able to provide general subjective feedback on the questionnaire. Generally users seemed to substantially favor the mouse interface to the speech interface. However, a majority of these responses listed the predominant reason for this preference as the "system delay." During the study, it became clear that the time taken to show the user the trajectory segment they selected before asking if they would like to input another segment was interpreted as a delay. While this delay was programmed in and intentional, lasting only three seconds, users viewed it (perhaps correctly) as a system flaw. This delay was specifically included in order to allow the results of this study to be compared with additional user interface studies that required a delay for timing. Future versions of this interface should certainly omit any unnecessary delays.

Users also suggested that while the speech interface was easy to use overall, creating full flight paths by selecting individual trajectory segments from a dropdown menu could become cumbersome no matter what interface style was used. Another common comment was that users would like some sort of error correction, a way to undo the selection of a trajectory segment. Addressing these user comments on artificial delay, how cumbersome the task was, and providing error correction could go a long way in improving the usability of the interface.

VI. Discussion

Post-study questionnaires indicate that users preferred the mouse interface to the speech interface. Comments suggested that this was predominately a matter of what users were already familiar with and therefore what they felt more comfortable and confident using. They also suggested that the built-in time delay, added to both the mouse and speech interfaces, proved more troublesome in the speech-based interface. Despite this preference, users met with sufficient success when using the speech interface (90.5% success rate for each).

Flight path B proved to be the most significant source of errors generated while using the speech interface (Figure 7). The predominate difference in this flight path was the inclusion of a diagonal segment that contained the word "backward." While users were prompted to say the word "backward" by the listing of flight path segments, and trained using this correct form, when presented with this flight path

TABLE III: NASA TLX Measures for Speech Interface

	Total	Previous Speech Interface Experience	Previous Speech Interface Satisfaction
Mental	2.54	2.35	2.28
Physical	0.93	0.7	0.72
Temporal	1.46	1.15	1.22
Performance	3.32	3.9	3.22
Effort	2.57	2.55	2.67
Frustration	3.25	2.9	2.44

during the study they often fell back on alternative forms (e.g., “back” or “backwards”.) Accounting for all options would provide a more robust and more successful interface.

Overall, users did not make use of the full allotted ten minutes of training time for either the speech or mouse interfaces. However, the amount of training time was largely correlated with overall success, and generally the amount of time taken to input the flight path. This may suggest an overall trend of carefulness – those users who took more time to ensure they were comfortable with the interface also took more time and care in the input of the flight path, thus meeting with higher success rates but also longer input time. Likely, experience also contributed to the limited training time necessary for both the speech and mouse interfaces. Because speech-based interfaces were not wholly new to most participants, little of the offered training time was ever used.

Another point that users made when selecting between the mouse and speech interface systems was that the mouse system was more familiar to them. This familiarity was expected, and understandably impacts levels of comfort and confidence experienced by users. The impact that familiarity with an interface has on the overall acceptance of that interface can be seen even within the data from the speech interface. Users who had previous experience with a speech-based interface rated this interface as less demanding, requiring less effort, and overall less frustrating than the overall average for these scores (Table III). Similarly, the longer that users spent training on the speech-based interface the more successful they were at using it to generate flight paths (Figure 10). This relationship suggests that the more time a user spends with the speech-based interface the more effective they are at using it to generate flight paths. As natural language interfaces become more ubiquitous and users become more familiar with their operation, acceptance of these interfaces should likewise increase.

Users also brought up the question of accent and language. For non-native English speakers, ensuring that each word was pronounced with accurate English accent increased workload significantly. Future systems should be prepared for multiple accents, as well as multiple languages, to ensure that a speech based interface can be used by as broad a spectrum of potential users as possible. Many participants switched to over-articulated and exact speech patterns when operating the speech-based interface. This is perhaps indicative of low expectations for machine performance – users expected the system to perform poorly, and as a result switched their speech patterns in an attempt to provide the system with exact language to digest. However, these shifting speech patterns had the opposite effect. The speech parsing system used in this study was built and trained upon colloquial speech data and expected users to pronounce and pace language as they would in everyday human-to-human communication. Over-enunciation of commands changed the way in which they were pronounced enough to cause potential trouble for the speech recognition system.

The over-enunciation problem encountered by some subjects has some interesting implications for the future of a speech-based interface. Should the interface be designed to accommodate such unusual speech patterns if they occur due to expected low capabilities of the system? Contrarily, should human users be trained to expect better of machines and communicate using the same basic speech patterns they would use for other humans? Due to the small size of the dictionary in operation for this study, the simplest solution would be to include alternate pronunciation options for each flight path segment, thereby allowing the system to anticipate and account for over-enunciated commands. However, future systems with more extensive vocabularies (e.g., the entire English language) may not have such luxury. It is also an interesting problem to consider whether human-to-machine communication relying on distinct pronunciation patterns may constitute its own dialect of sorts.

Finally, the speech interface examined in this user study included a number of characteristics that were designed not to be optimal for performance or user acceptance, but rather to allow for comparison of results with additional user studies. As a result, design choices such as the drop-down menu, the inclusion of a delay in between selecting a trajectory segment and selecting whether to add an additional segment, and the lack of any error correction methods were all purposefully included. In comments on their subjective questionnaires, users often identified these areas as problematic and factored them into their overall opinion of the speech-based interface. These are all obvious areas of improvement for the next generation of speech interfaces, and should substantially improve overall user acceptance of the interface once accounted for.

VII. Conclusion and Future Work

This paper presented an initial design of a speech-based interface for defining flight path trajectories for UAVs. It also presented the results of a user study designed to test the acceptability of such an interface.

The compound flight path segments, such as "Backward-Left", resulted in the lowest success rate, with an average of 1.1 errors per user for the speech interface compared to the 0 errors averaged for the mouse interface. This indicates one specific area where this current speech interface can be improved. However, despite the lowered success rate on this segment, users in general demonstrated comparable success with the speech interface as they did with the mouse interface (90.5% for both), suggesting that even with minimal training a novel speech interface can prove as effective as other common interfaces. Training time was, however, directly correlated to overall success, both for the mouse interface and to an even larger extent for the speech interface.

Most users did have prior experience with speech interfaces (71% of all participants), and most were satisfied with the speech interfaces they were using (90% of all participants with previous experience with speech interfaces, including common interfaces such as Siri, Amazon's Alexa, and other GPS/car interfaces.) Users with previous experience with these interfaces indicated that this speech interface required less mental, physical, and temporal demand as well as effort from them, and produced less frustration. However, these same users suggested that their overall performance was worse than the overall average. This suggests that as users become more familiar with speech based natural language interfaces, their ability to effectively use these interfaces will increase.

In addition, a high success rate and a higher user preference rate could be reached if several areas identified this study were improved during future work. Improving the CMU Sphinx dictionary and grammar to better recognize compound trajectory segments and over-enunciated word pronunciations and reducing the imposed delay between entering trajectory segments. Continued research on speech-based natural language interfaces could also improve the overall accuracy of the system by making use of recent advances in deep learning. By increasing the level of intelligence of the speech recognition system, the UAV interface can work with a wider range of users at a high speed, and produce even higher accuracy. This study has identified areas of improvement for next generation speech interfaces for UAVs.

This study has shown that users are willing to accept a speech-based natural language interface for defining flight paths. Prior research has also shown that natural language interfaces allow for a wider user-base than more complex subject-matter specific interfaces [1]. Together, this indicates that utilizing a speech-based interface will allow for an increase in UAV usability as well as an increase in the type of tasks that UAVs are used for.

However, previous research has also indicated that speech-based interfaces carry their own limitations [5]. Speech-based interfaces work best when a system can be trained to a particular voice, let alone a particular accent or language. Moreover, speech interfaces may become problematic in noisy environments. Additionally, truly intuitive natural language communication relies on speech used in conjunction with other forms of natural language. In order to compensate for these limitations, providing a highly intuitive natural language interface that works in the broadest possible set of circumstances, a multimodal interface should be considered. Such an interface should incorporate not only a speech-based system but also a gesture-based system, thereby combining two different natural language modalities for enhanced performance and more intuitive operation.

Acknowledgment

The authors would like to acknowledge Kyle McQuarry (Analytical Mechanics Associates, Inc.) for helping to create the mouse interface.

References

[1] M. Chandarana, A. Trujillo, K. Shimada, and B. D. Allen, "A natural interaction interface for UAVs using intuitive gesture recognition," in *Advances in Human Factors in Robots and Unmanned Systems*. Springer, 2017, pp. 387–398.

[2] D. Bamburry, "Drones: Designed for product delivery," *Design Management Review*, vol. 26, no. 1, pp. 40–48, 2015.

[3] H. Chen, X.-m. Wang, and Y. Li, "A survey of autonomous control for uav," in *Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on*, vol. 2. IEEE, 2009, pp. 267–271.

[4] E. Peshkova, M. Hitz, and B. Kaufmann, "Natural interaction techniques for an unmanned aerial vehicle system," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 34–42, 2017.

[5] M. Chandarana, E. Meszaros, A. Trujillo, and B. Allen, "Fly like this: Natural language interfaces for uav mission planning," in *Proceedings of the 10th International Conference on Advances in Computer-Human Interaction*. ThinkMind, 2017.

[6] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng, and J. Landay, "Speech is 3x faster than typing for english and mandarin text entry on mobile devices," *arXiv preprint arXiv:1608.07323*, 2016.

[7] T. Kojima, A. Kaminuma, N. Isoyama, and L. Guillaume, "Evaluation of natural language understanding based speech dialog interface's effectiveness regarding car navigation system usability performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2966–2966, 2016.

[8] A. Hocraffer and C. S. Nam, "A meta-analysis of human-system interfaces in unmanned aerial vehicle (uav) swarm management," *Applied Ergonomics*, vol. 58, pp. 66–80, 2017.

[9] M. Novitzky, H. R. Dougherty, and M. R. Benjamin, "A human-robot speech interface for an autonomous marine teammate," in *International Conference on Social Robotics*. Springer, 2016, pp. 513–520.

[10] J. Ferreiros, R. San-Segundo, R. Barra, and V. Pérez, "Increasing robustness, reliability and ergonomics in speech interfaces for aerial control systems," *Aerospace Science and Technology*, vol. 13, no. 8, pp. 423–430, 2009.

[11] D. T. Williamson, M. H. Draper, G. L. Calhoun, and T. P. Barry, "Commercial speech recognition technology in the military domain: Results of two recent research efforts," *International Journal of Speech Technology*, vol. 8, no. 1, pp. 9–16, 2005.

[12] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[13] M. A. Nielsen, "Neural networks and deep learning," *URL: http://neuralnetworksanddeeplearning. com/.(visited: 01.11. 2014)*, 2015.

[14] P. Mittal and N. Singh, "Speech based command and control system for mobile phones: Issues and challenges," in *Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on*. IEEE, 2016, pp. 729–732.

[15] C. M. University, "Cmu sphinx4-5prealpha," 2016, retrieved: Jan. 2016. [Online]. Available: http://cmusphinx.sourceforge.net/

[16] Audio-Technica, "Pro 8hemw hypercardioid dynamic headworn microphone," 2016, retrieved: Nov. 2016. [Online]. Available: http://www.audio-technica.com/cms/accessories/b119dcfe66995ac5/index.html