aerospace

Article

Faster Command Input Using the Multimodal Controller Working Position "TriControl"

Oliver Ohneiser [1,*], Malte Jauer [1], Jonathan R. Rein [2] and Matt Wallace [3]

[1] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany; Malte-Levin.Jauer@DLR.de

[2] Federal Aviation Administration (FAA), William J. Hughes Technical Center, Atlantic City International Airport, Egg Harbor Township, NJ 08405, USA; Jonathan.Rein@FAA.gov

[3] Deutsche Flugsicherung GmbH, Academy, Am DFS-Campus, 63225 Langen, Germany; Matthew.Wallace@DFS.de

* Correspondence: Oliver.Ohneiser@DLR.de; Tel.: +49-531-295-2566

Received: 6 April 2018; Accepted: 4 May 2018; Published: 8 May 2018

Abstract: TriControl is a controller working position (CWP) prototype developed by German Aerospace Center (DLR) to enable more natural, efficient, and faster command inputs. The prototype integrates three input modalities: speech recognition, eye tracking, and multi-touch sensing. Air traffic controllers may use all three modalities simultaneously to build commands that will be forwarded to the pilot and to the air traffic management (ATM) system. This paper evaluates possible speed improvements of TriControl compared to conventional systems involving voice transmission and manual data entry. 26 air traffic controllers participated in one of two air traffic control simulation sub-studies, one with each input system. Results show potential of a 15% speed gain for multimodal controller command input in contrast to conventional inputs. Thus, the use and combination of modern human machine interface (HMI) technologies at the CWP can increase controller productivity.

Keywords: air traffic controller; human machine interaction; human computer interaction; multimodality; eye tracking; automatic speech recognition; multi-touch gestures; controller command; speed gain

1. Introduction

Multimodal human-computer interaction (HCI) may enable more efficient [1,2] and especially natural "communication" because "natural conversation" is a complex interaction of different modalities [3]. It can also be seen as a "future HCI paradigm" [4]. The term "multimodal" can be defined as an "adjective that indicates that at least one of the directions of a two-way communication uses two sensory modalities (vision, touch, hearing, olfaction, speech, gestures, etc.)" [5]. Multimodal interaction (MMI) is interpreted as combining "natural input modes such as speech, pen, touch, hand gestures, eye gaze, and head and body movements" [6]. In addition, different types of cooperation between modalities in such systems can be used e.g., inputs from different channels might be redundant or need to be merged [7].

One advantage of multimodal systems is the flexibility due to alternative input modes, which also avoids overexertion and reduces errors [6]. Another benefit lies in the support of different types of users and tasks [6]. MMI also promise to be easy to learn and use, as well as being more transparent than unimodal interaction [8]. Furthermore, the operator load as a whole can be shared across all individual modalities [9]. Thus, the fusion of data with their origin in different input modalities is one essential factor for an effective multimodal system [10].

Even if some of the different modalities that are combined to a multimodal system are error-prone, the multimodal system normally is more robust against recognition errors [6]. In Oviatt's study, this is due to the users' intelligent selection of the best input mode for the current situation [6].

For many domains, the typical modalities to be used multimodally are speech recognition, eye gaze detection, and gestures [11], which are also applicable for the air traffic management (ATM) domain. In the course of SESAR (Single European Sky ATM Research Programme) it is necessary to integrate new technologies such as speech-, gaze-, and touch-inputs for an improved interaction between controllers and their system [12]. The possible speed and efficiency gain with multimodal interaction working with our CWP prototype TriControl is the key topic of this paper.

Section 2 outlines related work on multimodal systems. Our multimodal CWP prototype is described in Section 3. The study setup, methods, and participant data are presented in Section 4. The results of our usability study are shown in Section 5 and discussed in Section 6. Finally, Section 7 summarizes, draws conclusions, and sketches future work.

2. Related Work on Multimodal Human Computer Interaction

Different domains investigated the benefits and drawbacks of multimodal systems in the past. This section outlines multimodal prototypes and some important results on human performance when working multimodally.

2.1. Examples of Multimodal Interaction Prototypes

In recent years a variety of multimodal interfaces have been developed. MMI can support education for disabled people using gestures and sound [13]. Using MMI in a car, the driver may choose his/her preferred modality from speech, gaze, and gestures, and can combine the respective system input with different modalities [14]. Another MMI system connected to the steering wheel of a car enables input via speech and gestures [15]. A further example is the multimodal combination of gestures and voice to place items on a screen [16].

There are even a lot of examples in the air traffic domain. In a part-task flight simulator, the MMI allows for function control via eye gazes in combination with speech recognition [17]. Another air traffic control multimodal prototype incorporates pen and touch interaction as well as physical paper for flight strips [18]. The users of this system were able to get along with the MMI very quickly and did not feel overstrained. Eye tracking in combination with a mouse can be used for modification of air traffic control (ATC) radar display settings by the user as well [19,20].

In previous DLR developments, the use of eye tracking as an input device was also conceptually enhanced with speech recognition and multi-touch sensing for use in ATC [21,22]. DLR successfully evaluated the underlying unimodal prototypes for speech recognition [23], multi-touch [24], and eye tracking [25]. More MMIs related to air traffic management such as flight strip manipulation or an en-route interface can be found in [9].

2.2. Findings of Earlier Multimodal Interaction Studies

In findings of Oviatt, users were more likely to work multimodally if commands dealt with numbers or orientation of objects [26]. These aspects are to a certain extent also true when controlling aircraft on a radar display. However, the amount of multimodal overlap between manual input by hand and spoken utterances does vary heavily depending on the individual [6].

In a study on driving, the use of a buttons-only system was faster than the combination of gestures and speech interaction, but less visual demand and a comparable performance were reported for the multimodal system [15]. In a study concerning interactive maps, roughly 95% of users working with spatial data preferred interacting multimodally and a speed gain of 10% for system inputs was shown [27]. A preferred multimodal interaction was also found for a map-based military task, with a 3.5-fold improvement in error handling times [28].

3. Description of TriControl Prototype

In everyday communication with each other, we use multiple ways of transferring information, as for example speech, eye contact, and manual gestures. The same principle is the underlying idea of the DLR multimodal CWP prototype "TriControl". Therefore, it combines speech recognition, eye tracking, and multi-touch sensing, with the goal of enabling a faster and more intuitive interaction especially for approach controllers. The following section will outline the multimodal interaction philosophy of the prototype. A more detailed description including technical details can be found in [29].

The main task of the target user is the transfer of controller commands to the pilot and the read-back check [30,31]. Currently, these commands are mostly transmitted via radiotelephony (R/T) using standard phraseologies according to ICAO (International Civil Aviation Organization, Montreal, QC, Canada) specifications [32]. Furthermore, controllers are required to log the given commands into the aircraft radar labels respectively electronic flight strips of the CWP. This input is normally performed manually using keyboard and mouse. Regarding the structure of the commands, they are usually composed of "callsign–command type–value", e.g., "BER167–descend–FL60". Using TriControl, these three components are distributed over three input modalities, with each component assigned to a modality based on its suitability of transporting the corresponding piece of information.

As a result, the current two unimodal communication channels are replaced by one multimodal interaction with the CWP as visible in Figure 1. Each of the used input modalities of TriControl is presented in the following. When the multimodal input is completed, the combined controller command is logged by the system (no additional flight strip compiling required) and will then be transmitted from the CWP to the pilot via data-link. In case of failure in the submission process, the R/T communication may be used as a fallback solution. Although the TriControl concept is designed to enable a fast and efficient input into the CWP (intended for data-link transmission), the usage of text-to-speech could be a viable alternative for the transmission of commands entered into the system to aircraft in the traffic mix without data-link capability. As the concept focuses on the input method of the controller commands, future investigations are required regarding the integration of the pilots' read-back into the concept. Exemplarily, a digital acknowledge of the pilot sent via data-link could be visualized on the radar screen, or the usual read-back via R/T could be applied.

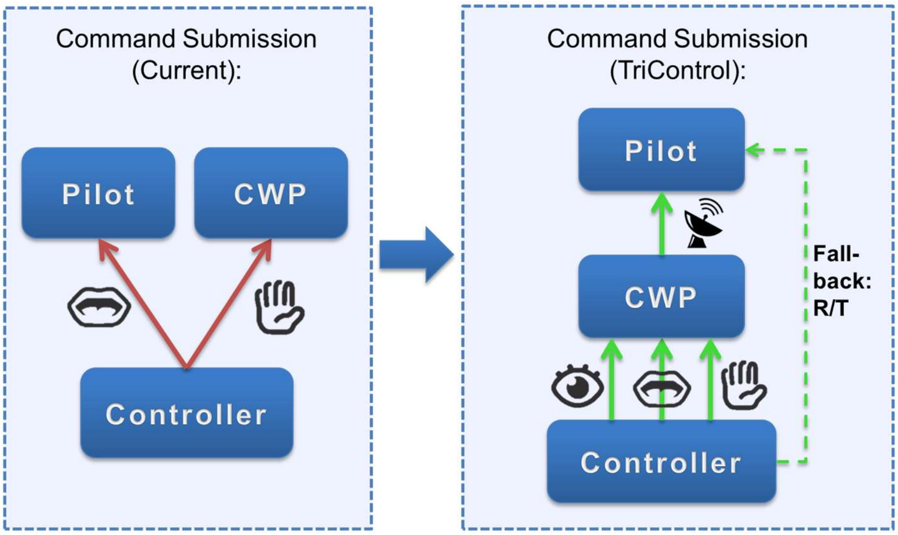
Command Submission (Current):
Pilot
CWP
Controller
Command Submission (TriControl):
Pilot
CWP
Controller
Fall-back: R/T

Figure 1. Difference between current controller command submission and TriControl command submission.

3.1. Eye-Tracking

TriControl's integrated eye tracking device determines the user's visual focus in order to detect the aircraft targeted by a controller command. In the same way as we address our conversational partners by eye contact in daily life, TriControl users can address aircraft on the radar screen just by looking at the radar labels or head symbols.

Therefore, the first part of the controller command—the callsign—is selected using the visual modality. To emphasize the selection of an aircraft, the corresponding label is highlighted by a white box around it as shown on the radar display in Figure 2 [33].

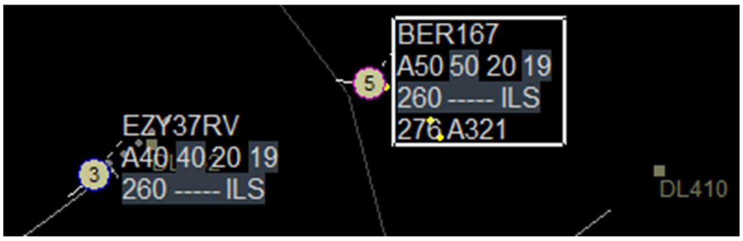

Figure 2. White box around the currently selected aircraft as shown in the radar display.

3.2. Gesture Recognition

If someone asks us how to navigate to a location, we automatically use hand gestures to describe directions. These coarse-grained pieces of information (e.g., "left", "right", "straight ahead", etc.) resemble the command types of a controller command like "climb", "descend" and so on. Thus, TriControl uses gesture recognition on a multi-touch device for the insertion of the command type. In the current prototype, the following gestures are implemented to insert the corresponding command types: swipe left/right for reduce/increase of speed; swipe up/down for a climb/descend in altitude; rotate two fingers for a heading command, and lastly long-press one finger for a direct/handover/cleared-ILS command, where the final type depends on the value of the complete command (e.g., a value of "two three right" would correspond to a runway, therefore the type would be interpreted as cleared-ILS).

Additionally, we implemented a few convenience functions for human machine interface (HMI) manipulations. Thus, the controller is able to change the zoom factor and visible area of the radar display by a 5-finger gesture: movement of all fingers moves the visible area and a spreading/contraction of the fingers zooms in/out the map section. TriControl also offers the display of distances in nautical miles between two aircraft that are selectable by multi-touch, i.e., by moving two fingers on the multi-touch screen to position two cursors on the main screen at one aircraft each.

3.3. Speech Recognition

Regarding discrete values, a natural choice to transmit this kind of information is by voice. Normally, when trying to insert specific values using different modalities, e.g., mouse, gestures, or eye gaze, the solution would most likely require some kind of menus that can be slow to search through. Automatic speech recognition is capable of converting spoken text or numbers into digital values, especially when the search space is limited—in this case limited to relevant values in the ICAO standard phraseology (e.g., flight levels, speed values, degrees, or waypoint names). TriControl therefore accepts spoken values from recorded utterances initiated by using a push-to-talk foot switch to complete the third element of a controller command in a natural and convenient way.

3.4. Controller Command Insertion

As an example, the insertion of the command "DLH271 reduce speed 180 knots" into the system is shown in Figure 3.

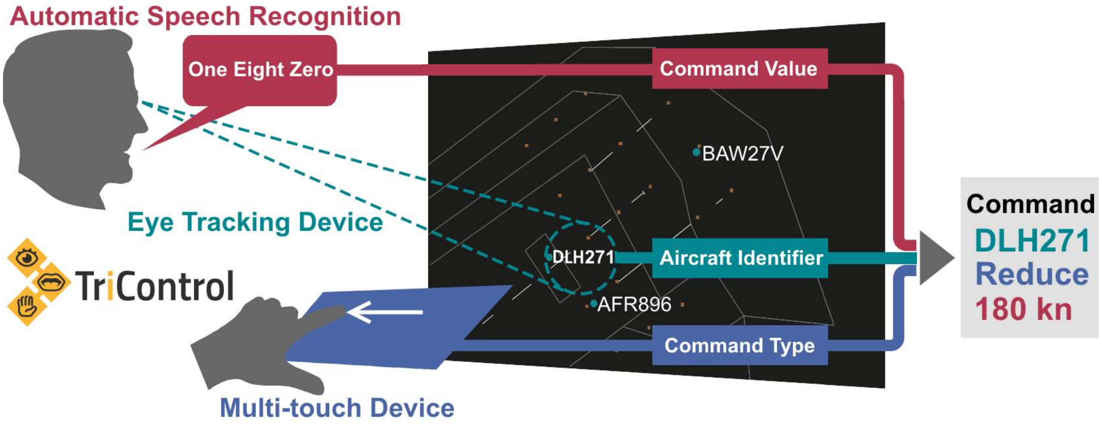

Figure 3. Interaction concept of TriControl with fusion of three modalities adapted from [34].

For this task, the controller firstly focused the corresponding label with the eyes. Then, the command type was inserted via swipe-left gesture and the value "one eight zero" was spoken. The order of uttering and swiping did not matter. It was even more efficient to perform it in parallel. When the controller began gesturing or speaking, the selected callsign locked so that the controller was already free to look somewhere else other than the label while completing the command with the other modalities. When all parts of the controller command were inserted into the system, a single tap on the touch screen confirmed the new clearance, which was also highlighted in the label. If command type or value was recognized or inserted incorrectly, it could be overridden by a new insertion before the confirmation. Alternatively, the whole command could be refused at any time using a cancel button on the multi-touch device as a measure of safety against unwanted insertion of clearances.

In contrast to the traditional method of inserting all command parts via voice when using radiotelephony or speech recognition of the whole command [23], this splitting of command parts to three modalities was envisaged to enable a timely overlap of the separate inputs. Because of this potential for concurrent input of the controller command parts and the marginal amount of time needed for the callsign selection (as the controller would look at the label anyway), the input method was designed to enable a faster, more natural and efficient insertion of commands into the system, while each modality was well suited for input of the respective type. A quantitative analysis of how much speed can be gained will be presented in the next section.

4. Usability Study with Controllers at ANSP Site

For a quantitative evaluation on how controllers use TriControl, we prepared the generation process of log files from different parts of the system. The log files capture:

- Positions of aircraft icons, aircraft radar labels, and waypoints as shown on the radar display;
- Eye gaze data with timeticks and fixation positions;
- Begin, update, and end of a touch gesture, with timeticks including confirmation gesture or rejection button press and release times;
- Press and release times of the push-to-talk foot switch (related to length of controller utterance);
- Timeticks for appearance of callsign, command type, and command value on the radar display.

4.1. Recorded Timeticks during TriControl Interaction

Eleven points in time (timeticks) were really relevant to judge speed differences during interaction: E1/E2/E3 (Eye Fixation), G1/G2/G3 (Gesture), S1/S2/S3 (Speech), and C1/C2 (Confirmation) as shown in Figure 4.

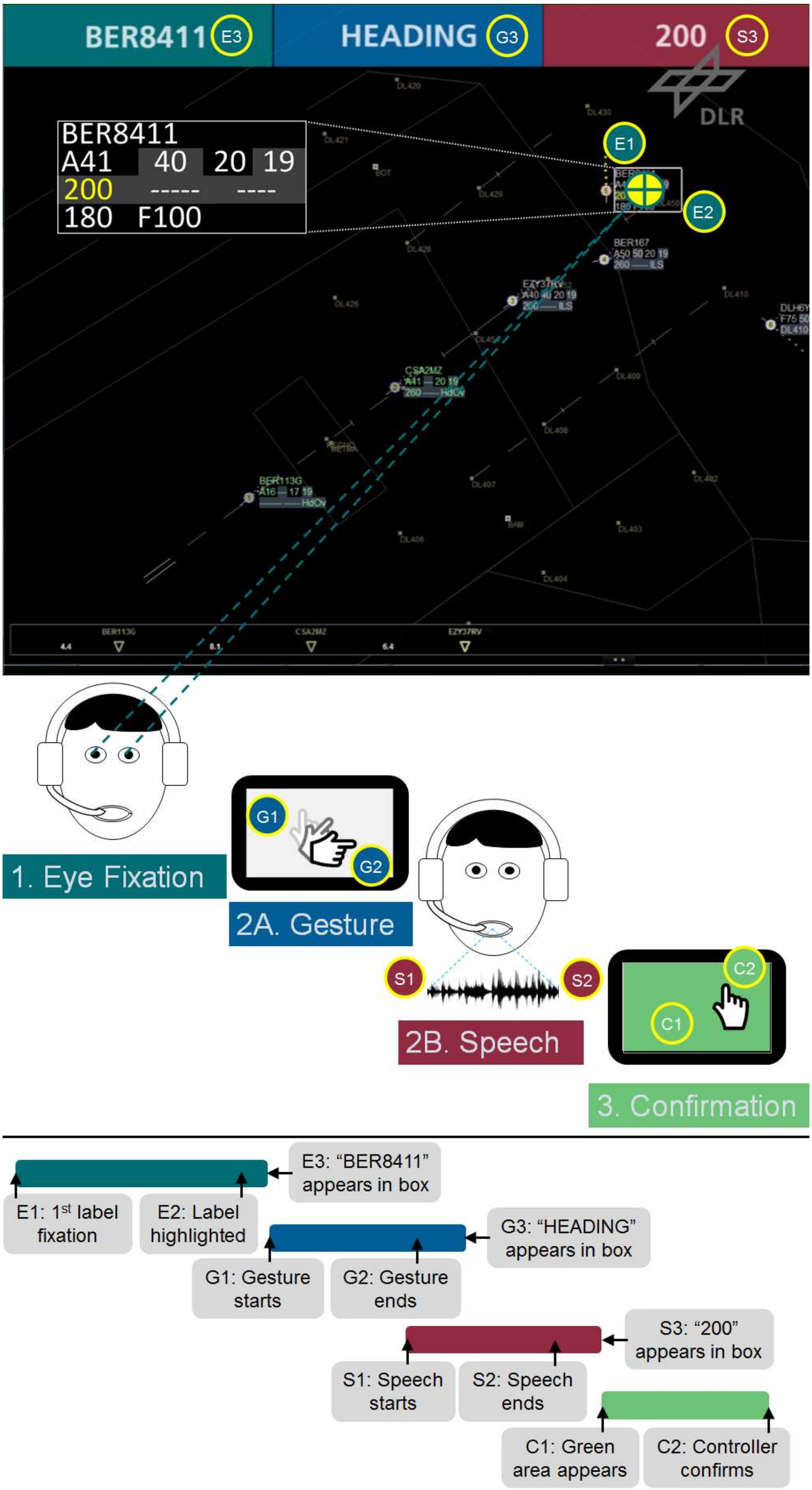

Figure 4. Measuring modality interaction and confirmation times when using multimodal controller working position (CWP) prototype TriControl.

The command input starts with the first eye fixation (1) of an aircraft icon or radar label. In the example of Figure 4, the label of "BER8411" was fixed by the participant's gaze (E1). If the captured gaze points are located within an area of around 0.2 percent of the screen for at least 20 ms, the system highlights the label with a white frame (E2) and presents the callsign in the dark green upper left box (E3).

Afterwards, the gesture phase (2A), the speech phase (2B), or both phases in parallel may start. In our example, the participant touched the multi-touch device (G1), performed a rotating "HEADING" gesture, and lifted his fingers again (G2). The gesture-recognizer module evaluated the gesture type and visualized it in the blue upper middle box (G3). However, the gestures may have already been recognized during the gesture update phase if further finger movements did not change the gesture type anymore. Hence, G3 could also happen before G2. The later timetick was valid for our calculations.

In our example, the participant pressed the foot switch to activate the speech recognition (S1) before ending up the gesture. After having spoken the value "Two Hundred" or "Two Zero Zero" (the system supports ICAO phraseology conform inputs as well as occasionally used variations), the controller had to release the foot switch (S2). The duration between S1 and S2 was also compared to the automatically evaluated length of the corresponding audio files in milliseconds. In the current prototype version, the utterance "200" was only analyzed after the foot switch release so that the recognition result appearance time (S3) in the red right upper box was always after S2.

As soon as all three command elements (callsign, command type, and command value) were available and presented in the three colored top boxes, two other elements appeared. First, a yellow "200" was displayed in the direction cell of the corresponding aircraft radar label (first cell in line 3). Second, the whole touchable area of the multi-touch device turned green to ask for confirmation of the generated command (C1). As soon as the single touch to confirm the command was recognized (C2), the command was entered into the system and the yellow "200" in the radar label turned white. The controller always had the opportunity to cancel his inputs by pressing a hardware button on the right side of the multi-touch device. This interaction was not considered as a completed command. After the confirmation event C2, the command insertion was considered as finished.

4.2. Multimodal Interaction Study Setup

From 4–6 April 2017 we conducted an interaction study using the multimodal CWP prototype TriControl at the German Air Navigation Service Provider (ANSP) DFS Deutsche Flugsicherung GmbH in Langen (Germany). 14 air traffic controllers from DFS took part in the study. The average age was 47 years (standard deviation = 10 years; with a range from 29 to 62 years). The ATC experience after finishing apprenticeship was an average of 21 years (standard deviation = 12 years; with a range from 7 to 42 years). Some of the older controllers were already retired. The current controller positions were Approach (7xAPP), Area Control Center (5xACC), Upper Area Center (2xUAC), Tower (4xTWR), and one generic instructor. Several controllers had experience in multiple positions. Eight participants wore glasses, two contact lenses, and four took part without vision correction. Two of those fourteen controllers were left-handed. Depending on the modality, some participants had previous experience with eye tracking (5), gesture-based interaction (3), or speech recognition (10). Controllers' native languages were German (9), English (3), Hindi (1), and Hungarian (1).

The simulation setup comprised a Düsseldorf Approach scenario using only runway 23R. There were 38 aircraft in a one-hour scenario without departures. Seven of them belonged to the weight category "Heavy", all others to "Medium". Each participant had to work as a "Complete Approach" controller (means combined pickup/feeder controller in Europe and combined feeder/final controller in the US, respectively). After approximately 15 min of training using a training run with less traffic density, a 30-min human-in-the-loop simulation run was conducted (see Figure 5).

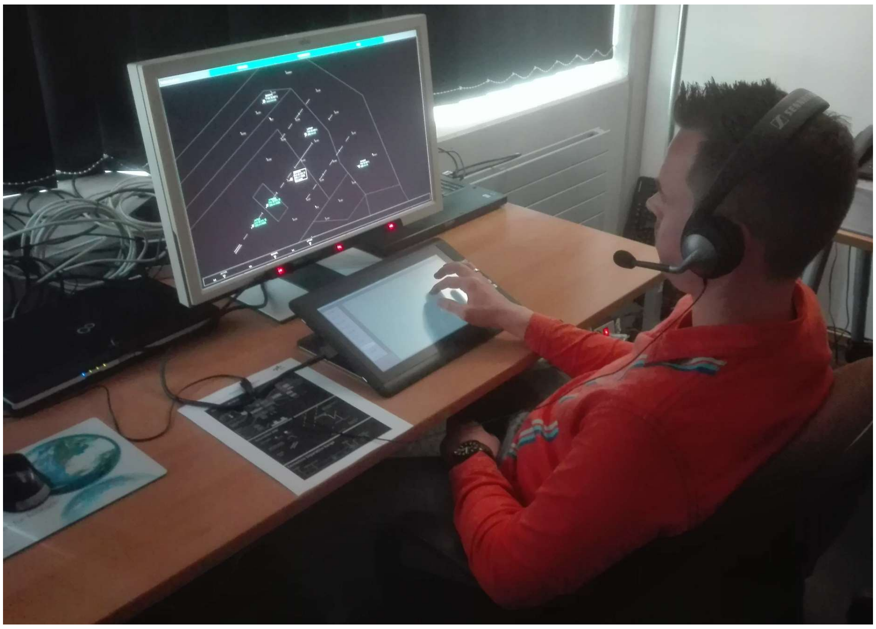

Figure 5. Participant during TriControl trials before command confirmation and after uttering a value, performing a multi-touch gesture. His gaze is being recognized at an aircraft almost on final (white radar label box).

The TriControl multimodal interaction results were compared against Baseline data gathered during a study with a conventional input system in November–December 2015 and January 2017. During these simulations, controllers had to give commands via voice to simulation pilots and had to enter those commands manually via mouse into the CWP. Both of these necessary controller tasks could be performed sequentially or in parallel. Twelve radar approach controllers took part in this earlier study. Four controllers were sent from DFS, eight controllers from COOPANS (consortium of air navigation service providers of Austria (Austro Control, Vienna, Austria, four controllers), Croatia (Croatia Control, Velika Gorica, Croatia, one controller), Denmark (Naviair, Copenhagen, Denmark, one controller), Ireland (Irish Aviation Authority, Dublin, Ireland, one controller), and Sweden (LFV, Norrköping, Sweden, one controller)).

The average age was 39 years (standard deviation = 11 years; within a range from 22 to 56), their professional work experience 17 years (standard deviation = 11 years; within a range from 1 to 34). The number of aircraft and weight category mix in the traffic scenario were the same as for the TriControl study. However, training and simulation runs lasted some minutes longer in this study (at least 45 min each). This aspect should not affect the duration of single controller commands during the run time except for some training effects.

5. Results Regarding Controller Command Input Duration and Efficiency

One of the primary goals of TriControl is to enable faster controller commands, as compared to Baseline. We defined the TriControl command input duration as the amount of time between the controller's initiation of the multi-touch gesture (or the pressing of the voice transmission pedal, whichever came first) and the confirmation of the command on the multi-touch display. In Figure 4, this is the difference between G1 and C2. For Baseline, we defined the command duration as the amount of time between the controller's pressing of the voice transmission pedal (or entering a value in the data label, whichever came first) and the confirmation of the command in the data label (or the release of the voice pedal, whichever came later).

The time for visually acquiring the radar label information and consideration (roughly E3–E1 in Figure 4 for TriControl) should be similar to the "time to think" in Baseline. Therefore, the "time to think" is not part of the command input duration time in both conditions. Response times are generally not normally distributed, due to the occasional outlier response that is significantly longer than average. Therefore, for both measures (command input duration of TriControl and Baseline), we computed the median duration for each controller, which is a much more stable estimate of individual controllers' response times. We compute the mean of those median response times and conduct *t*-tests comparing the means of Baseline and TriControl, reporting the *t*-statistic and *p*-values of individual comparisons. Although raw response times are not normally distributed, the controller medians are consistent with the normality assumptions of the *t*-test. When we presented data for individual command types (e.g., altitude, speed), we only included data from controllers with at least five of those commands.

5.1. Command Input Duration

Figure 6 shows the command duration results for the full set of commands and for the most common command types. Overall, TriControl commands (mean = 5.4 s) were slower than Baseline (mean = 4.7 s), though this was only marginally statistically significant: $t(24) = 1.77$; $p = 0.09$.

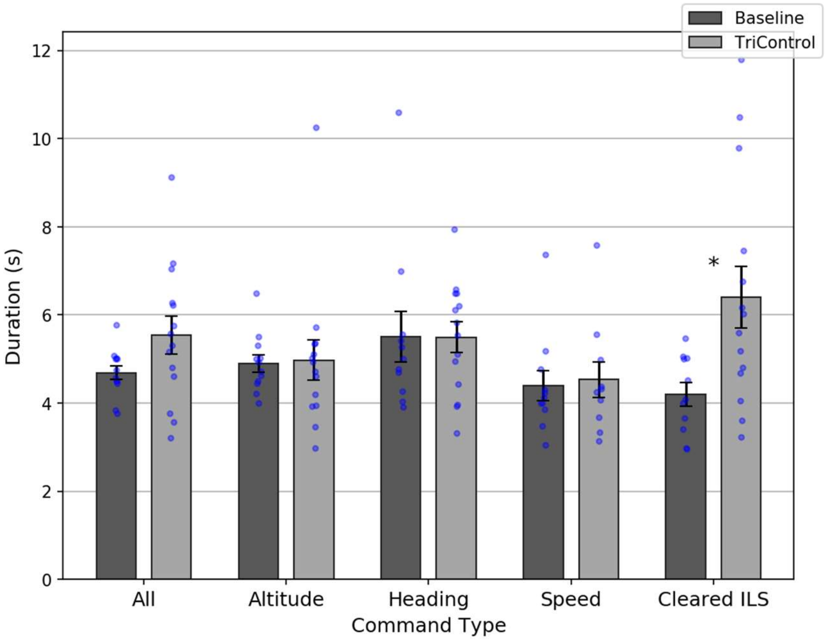

Figure 6. Baseline vs. TriControl command duration. Error bars represent ±1 standard error of the mean. Blue dots represent individual controllers.

Although TriControl was slower overall, this likely does not reflect the system's true performance potential. With only 15 min training, participants were still learning the system during the main scenario. Comparing the first 15 min to the second, there was an 11% drop in the number of gesture or voice corrections per command (0.62 to 0.55). There was also a 0.4 s decrease in the command duration for commands that did not require a correction (4.8 to 4.4 s).

In addition to training, there are also known limitations with TriControl that we plan to address in the future. At present, the speech recognition module is tuned to English language with a German accent. The German-speaking controllers made 50% fewer voice corrections and their command durations were 1.7 s shorter. Also, TriControl does not presently support chained commands (e.g., an altitude and speed command in a single transmission). In the Baseline condition, unchained command durations were 0.3 s longer than chained commands, on a per-command basis (mean = 4.7 s).

A best-case comparison would reflect the faster, low-error performance that we would expect from additional training and use, speech recognition that is trained on more accents, and an implementation of chained commands. With the current data set, we compared the Baseline unchained commands to the TriControl commands issued by native German speakers in the second half of the simulation that did not require corrections.

This comparison was reasonable, as controllers in the Baseline were very skilled and trained due to the similarity of the test system with current operational systems. Therefore, all following analyses encompass only the second half of the exercise. Results of the best-case comparison are shown in Figure 7.

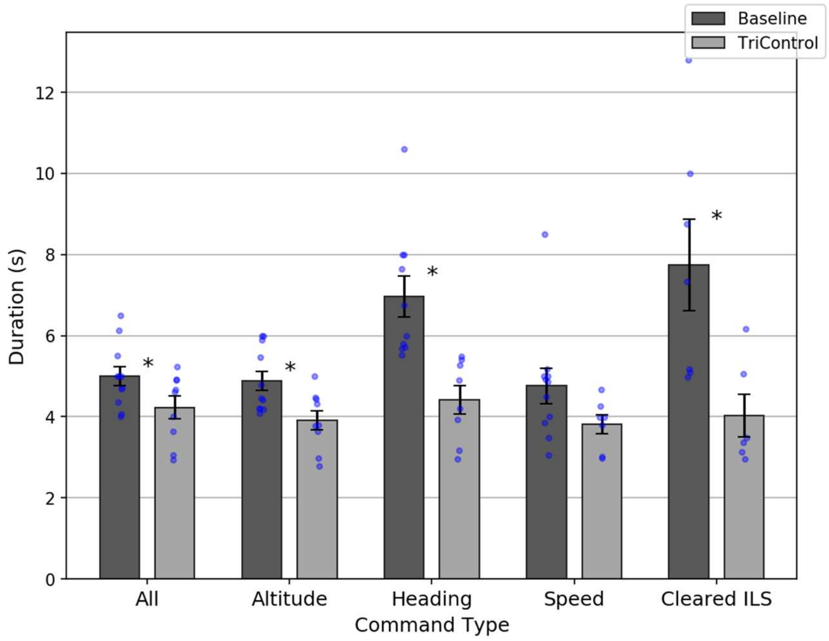

Figure 7. Baseline unchained commands vs. TriControl second half uncorrected German-speaker command duration. Error bars represent ±1 standard error of the mean. Blue dots represent individual controllers.

Overall, TriControl command durations (mean = 4.2 s) were shorter than Baseline (mean = 5.0 s), which is statistically significant: $t(18) = 2.11$; $p = 0.049$. This is a drop that exceeds 15%. There was also a significant speed improvement for altitude (3.9 vs. 4.9 s, $t(18) = 2.82$; $p = 0.01$), heading (4.4 vs. 7.0 s, $t(16) = 3.9$; $p = 0.001$), and ILS (4.0 vs. 7.7 s, $t(11) = 2.82$; $p = 0.02$) commands.

In addition to shorter median command durations in this best-case comparison, TriControl also has a smaller number of long-duration commands, excluding those that required input corrections. Figure 8 shows the cumulative distributions of command durations for Baseline and the uncorrected commands issued during the second 15 min of the TriControl simulation. Although the shortest 50% of commands have similar durations, the 75th percentile TriControl command is approximately one second shorter than Baseline, and the 90th percentile is two seconds shorter.

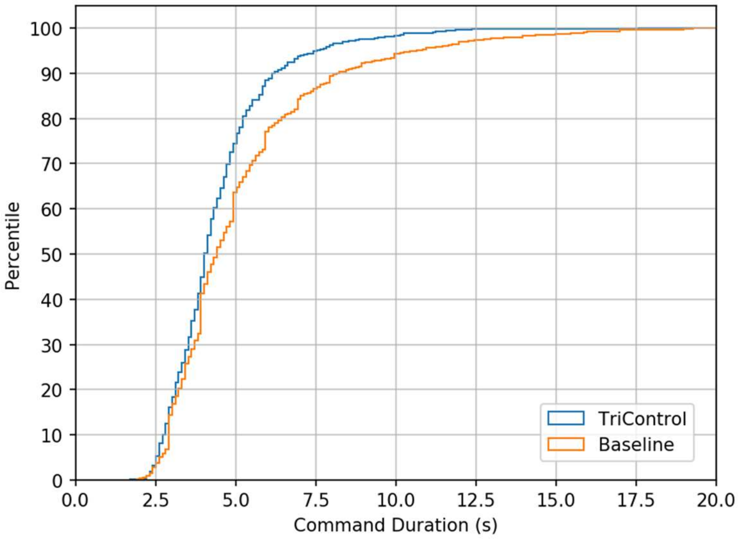

Figure 8. Cumulative distribution of Baseline vs. TriControl second half uncorrected command durations.

The 75th percentile comparison is also shown in Figure 9. Command durations were shorter for TriControl overall ($t(24) = 2.69$; $p = 0.01$), and for altitude ($t(23) = 5.51$; $p < 0.001$), heading ($t(21) = 2.41$; $p = 0.03$), and speed ($t(24) = 2.56$; $p = 0.02$) commands.

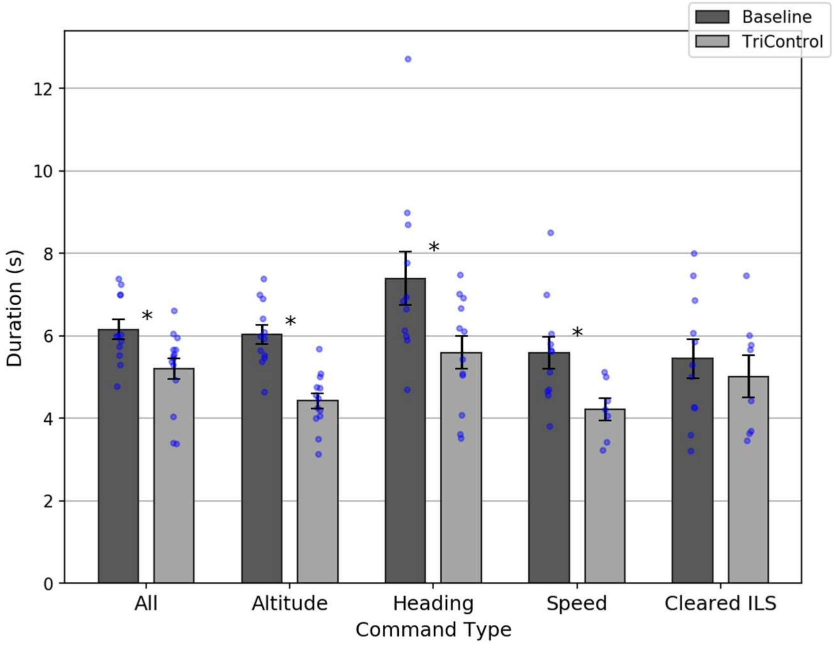

Figure 9. Baseline vs. TriControl second half uncorrected 75th percentile command duration. Error bars represent ±1 standard error of the mean. Blue dots represent individual controllers.

5.2. Command Input Efficiency

Finally, we investigated how efficiently controllers used the voice and manual modalities with each system. In particular, we looked at the portions of the command that can be parallelized. In the Baseline condition, this is the voice transmission of the command and the manual data entry into the radar label. In TriControl, this is the command type gesture and the command value vocalization. We defined the modality consecutiveness of the command as the ratio of (1) the total time to complete the voice and manual steps to (2) the larger of the two steps' durations. For example, consider a command with a manual duration of one second and a voicing duration of two seconds. If the controller starts and completes the manual component while they are performing the voicing component—completely in parallel—then that command has a modality consecutiveness of 1.0. If the steps are done sequentially with no delay, that is a modality consecutiveness of 1.5. A delay of half a second would correspond to a consecutiveness of 2.0.

Figure 10 shows a scatterplot of each controller's median command duration and mean command modality consecutiveness of Baseline and for the uncorrected commands in the second half of TriControl. The data points for Baseline were fairly clustered, with modality consecutiveness ranging from 1.09 to 1.47, and most durations falling between 3.75 to 5 s. In contrast, the TriControl modality consecutiveness ranged from 1.0 to 2.04, and duration ranged from 2.9 to 5.2 s. A third of TriControl controllers were issuing their commands with more parallelism than the most parallel controller in the Baseline condition. This suggests that TriControl does support more parallelism than Baseline, and that as controllers become more comfortable performing the command components in parallel, they will produce commands more efficiently.

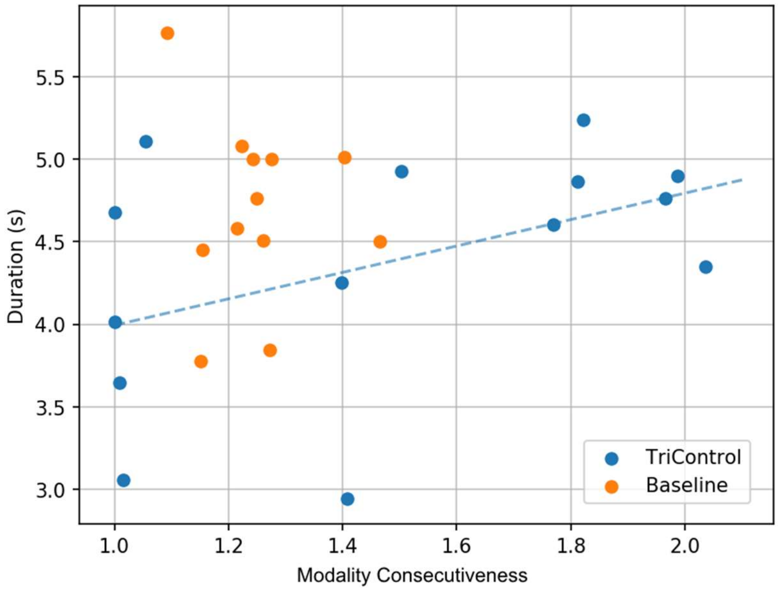

Figure 10. Command duration by modality consecutiveness for Baseline and TriControl second half uncorrected.

6. Discussion of Study Results and the TriControl Prototype

This section discusses the presented results with respect to performance, outlines some qualitative feedback and then considers safety aspects of the TriControl system.

6.1. Performance Considerations Derived from Results

On the one hand, for the majority of command types in Figure 6, the best performance was achieved from TriControl controllers. On the other hand, this is also true for the worst performance. However, this indicated a lack of training. More training could help the worst TriControl performers to also achieve results comparable to the best TriControl performers and thus better than the Baseline average. This assumption is already supported by the analyzed data subset of Figure 7.

Some TriControl enhancements, especially for the value recognition of different accents and more training, might lead to faster command inputs of up to two seconds per command, depending on the command type. The misrecognition of the uttered word "tower" of some native respectively non-German accent speakers to initiate an aircraft handover to the next controller position is one main aspect that needs to be improved. There was of course, only a limited set of modeled accents for this first study. Enhancements are also necessary regarding a number of misrecognized confirmation touch taps on the multi-touch device.

The fastest commands of TriControl and Baseline were comparable (Figure 8). However, TriControl enabled faster medium command duration times that might again even improve with more training. This can result in roughly one second saved per command input over all analyzed command types (Figure 9).

The assumed improvement due to more than 15 min of training (and 30 min simulation run) with the multimodal system can also be seen in Figure 10. When analyzing the best participants working with TriControl (near *x*- and *y*-axis in Figure 10) they were more efficient and faster than the best Baseline participants. As the Baseline environment is quite equal to today's controller working positions, it can be assumed that the potential for improvement of performance (orange dots in Figure 10) is low. This is also supported by the clustering effect of those orange dots for Baseline participants. As there might not be anticipated a performance decrease of the best TriControl participants (lower left blue dots), the bad performers will probably improve with training and reach or overcome the Baseline average (blue dots in the upper right area of the plot might move more to the lower left). The familiarization to the new parallel multimodal input was reported as a main aspect of training. Many subjects already improved in the short amount of time during the study. This is also interesting from a pedagogical perspective: an increase in performance over the Baseline with only a short 15 min period of training raises the prospect of what results could be expected with the inclusion of a more comprehensive training and preparation program for the test subjects.

6.2. Qualitative Information about Test Subjects' Opinion

The feedback of test subjects gathered in a debriefing session after the simulation run concerned various positive aspects and suggestions for improvements. A lot of statements were related to the learnability such as:

- "I found it, surprisingly, simple after a period of practice.",
- " … easy to spot—and after a short time as well easy to correct.",
- "I had no problems at all.",
- "It's quite interesting on a conceptual level and easy to use.",
- "The system is easy to learn.".

However, there were some flaws that need to be taken into account in the further development according to:

- "Conditional Clearances and Combined Calls are needed for daily work; at this time the clearance given is too simple to reflect the demands.",
- "The system often reacts too slow; e.g., with respect to the eye tracking dwell time; too long 'staring' necessary to 'activate' the square.",
- "Instead of 'watching the traffic' I needed to 'watch if the system complies'.",
- "TriControl focuses on just one aircraft. That might be a reason for attentional/change blindness.",
- "Uttering the whole command might be better for the situational awareness.",
- "There is potential for confusion and errors.",
- " … for use in ATM systems many more variables need to be incorporated and tested for safety".

Nevertheless there were many encouraging comments to further follow the multimodal approach:

- "A useful tool.",
- "This leads to less misunderstandings.",
- "I think it's worth to think about systems like TriControl, but it is really hard to state now if I would really work live-traffic with it … ; the use is fun!",
- "I would prefer TriControl over mouse or screen inputs.",

- “As an On-the-job Training Instructor (OJTI) teaching controllers to be instructors, this would be a good system to easily see what the controller is thinking/doing. I liked the system. Naturally, it would require practice and exercise to improve the skills required. However, once done, I believe it would be a good aid to the controller.”,
- “After adequate training I expect significantly more performance.”.

The level of comments from the participants broadly corresponded to the age of the participants and their success with the system—the younger participants often had greater success (best case scenarios) with making command submissions/inputs simultaneously, and as a result, enjoyed a higher success rate with the evaluation. This lead to more positive anecdotal comments regarding the potential of the system compared to many of the older test subjects. Older subjects found the multi-modality more challenging to coordinate and commented that they were frustrated by the system. The success rate in ATM training generally has been recognized to be better with younger trainees within a certain age range; it is assumed that within this range they can develop cognitive skills more readily.

To sum comments and observations up, the TriControl prototype in the current stage has—as expected—still far to go until it can be considered for operationalization. However, the underlying concept seems to have great potential for benefits, and for being used in future ATC environments with controllers who are “native” with modern HMI technologies.

6.3. Safety Considerations

In terms of safety, the general ATC tasks and procedures of the controllers did not change using TriControl compared to the traditional operation. Although the current state of the TriControl prototype has a limited amount of available commands, the radiotelephony channel is always intended to serve as a safe fallback solution for exceptional cases. Also today there is the potential for a misunderstanding between controller and pilot in the voice communication. This safety factor is to some extent comparable to an accidental insertion of erroneous commands into the system. To prevent this issue, the confirmation step for the command inserted via the three modalities was introduced. However, during the trials, a number of commands mistakenly input into the system were recognized. As stated earlier, this is most probably a result of the voice recognition not modeled for the respective accent, and misrecognized confirmations. Those issues are expected to decrease due to two reasons. First, increased training of the controllers with the system will lead to less erroneous inputs. Second, further development of the prototype foresees an elaborated plausibility check with respect to all three parts of a command and the command as a whole with respect to the current air traffic situation.

Hence, this plausibility check even goes beyond the two-way command-read-back-check of controllers and pilots today. With TriControl it will not be possible to assign commands to aircraft callsigns that do not exist in the airspace as sometimes happens nowadays. Furthermore, the TriControl system might immediately execute a what-if-probing and warn the controller before confirmation and thus issuing of the command, in case of potential negative implications to the air traffic situation. Besides, the ambiguity of human verbal communication can be reduced as TriControl uses a clearly defined set of commands to be issued.

Nevertheless, it is obvious that low error rates regarding eye tracking, speech, and gesture recognition should exist. Those rates are achievable e.g., using Assistance Based Speech Recognition (ABSR) as already partly used by TriControl. When analyzing complete commands with all three command parts uttered as usual, Command Error Rates of 1.7% can be achieved [23]. At the same time, wrong and forgotten inputs into the aircraft radar labels can be reduced with the help of an electronic support system compared to just using the mouse for all manual label inputs [23]. An intelligent aircraft radar label deconflicting is important to avoid selecting wrong aircraft via eye tracking if some aircraft are near each other. Malfunction of gesture recognition to generate unintended command types have hardly being pointed out by controllers during the study. As low error rates with respect to the "automatic recognition" of generated commands can be expected in the future TriControl version and the additional confirmation step exists, no basic ATC safety showstopper is currently expected.

The aspect of fatigue can also be related to safety. It needs to be analyzed if controllers might experience fatigue earlier using the three modalities as implemented in the current stage of the TriControl prototype compared to just using voice communication and mouse inputs. However, if a controller is able to—at a later stage of TriControl—choose the modality that is the most convenient for the current situation, this should not lead to earlier fatigue than in traditional CWP HMIs.

As this study was intended to get early insights on the potential speed and efficiency benefit of the novel interaction concept, future versions of the prototype will incorporate measures to maintain or increase the safety introduced by TriControl compared to traditional CWPs.

7. Summary and Outlook

TriControl is DLR's multimodal CWP prototype to enable faster and more natural interaction for air traffic controllers. The comparison of TriControl (controllers used this setup for the first time ever) and Baseline (controllers were very familiar with this kind of setup) study results primarily reveals the potential to improve speed and efficiency of air traffic controller command input under certain conditions. The short training duration during the TriControl study and thus the familiarization with the multimodal system most probably even worsened the TriControl in contrast to Baseline results.

Hence, in the pure analysis of the data as a whole, TriControl does not seem to enable faster and more efficient interaction. However, a closer look into specific parts of the data unveils relevant benefits. When investigating the second 15 min of each TriControl simulation result, the data of unchained and uncorrected commands and of participants with a German accent, for which TriControl was designed in its current version, a speed gain of 15% can be seen. An improvement in speed and efficiency might lead to less controller workload, or could in the long run help to increase the number of aircraft handled per controller.

The multimodal TriControl system is of interest to research and educational specialists, for example within the DFS, because it combines several developing technologies, which are either being researched, in operational use in ATC systems, or already used in training. Touch input devices are a mature and widely used ATC interface. The ability to 'email' clearances to aircraft via Controller Pilot Data Link (CPDLC) rather than by voice is used extensively in ATC communication systems. Research is being undertaken in eye tracking and its' use in controller support tools. Voice recognition and response (VRR) systems are used in various aspects of training, and the technology is improving.

Trials of VRR in ATC simulation within the DFS Academy have suggested similar results to those in the TriControl trial: general VRR recognition results are often overshadowed by higher misrecognition rates of the verbal instructions of a small number of users, sometimes because of an unusual accent for which the VRR has not been tuned. This limits its use as a training tool, and can also lead to low expectations and poor acceptance levels by users. Improving the system's recognition ability can remove this potential roadblock to its implementation.

Many aspects like controllers' experience with the involved interaction technologies, the parallelism of input, their own controller working position design and responsibility area as well as age might have influenced the TriControl results. Nevertheless, some controllers performed really well with the multimodal system that has only been planned for future CWPs. Hence, there will probably be a new generation of trainable controllers in the future to benefit from the potential speed gain in command input. There is a list of reasonable enhancements for this early version of a multimodal CWP prototype that we gained from the evaluations.

In a next step, TriControl will for example, use a context-based fusion of input modality data for plausibility checking of information, as DLR has already shown that it is reliable in speech recognition [35]. This step aims to reduce recognition errors by using the information from all three modalities to cooperatively construct a reasonable controller command. This could also increase the safety as the pre-check of complementary command parts will be performed before sending the information to a pilot via data-link or standardized text-to-speech.

Other differences to current interaction are the lack of chain commands, conditional clearances, or traffic information. Therefore, we hope to find reasonable ways to include possibilities to cover a great amount of the controllers' interaction spectrum by incorporating their feedback as early as possible. Accordingly, the effects on safety using the TriControl system have to be analyzed as well, when the maturity of the prototype has increased. The limited time available to each participant did not allow familiarity with the system to be fully developed before the trial took place. With a longer period of time in a second trial, the individual tasks could be evaluated, as well as combined tasks, prior to the exercise. This data evaluation would be of interest, but would involve more intensive evaluation that was not possible in the first trial. Training could then be tailored to assist and overcome the main problems identified in a qualitative error analysis. A quantitative error analysis was not reasonable for the first trial due to the described different sources of errors and as it was not the scope of the first trial.

To furthermore investigate the effects of users' free choice of modalities, we will extend each of the involved modalities to enable inputs for as many command parts as possible. On the one hand, users will be able to choose the modality for a piece of information as it fits their needs or the situation. On the other hand, quantitative evaluations will be performed to assess the performance of different interaction modality combinations. This will also enable an assessment of the performance using the modality combination presented in this work compared to combinations more similar to the current work of the controller.

The multimodal interaction CWP prototype TriControl showed the possibility of fast and efficient input of most approach controller command types into the digital system already now. This will be essential in the context of digital ATC information distribution and future ATC tasks.

Very similar benefits are as well anticipated by the major European ANSPs and ATM system providers in the course of SESAR2020's solution "PJ.16-04 CWP HMI" investigating the effects of automatic speech recognition, multi-touch inputs and eye-tracking at the CWP on controller productivity. Thus, future activities will probably also use and combine the advantages of those innovative HMI technologies.

Author Contributions: O.O. and M.J. were responsible for development of the TriControl system. They also conceived and designed the experiments with great organizational support of M.W., O.O. was the main author of this article being supported by all three co-authors. J.R.R. was responsible for the preparation and conduction of the data analysis (mainly represented in results section).

Acknowledgments: We like to thank all air traffic controllers that participated in our human-in-the-loop studies during the last two years. Besides, we are grateful for the support of Konrad Hagemann (DFS Planning and Innovation) and Hartmut Helmke (DLR) in preparing the simulation trials in Langen (DFS) respectively Braunschweig (DLR). Thanks also to Axel Schmugler (Technical University Dresden) for assisting during the TriControl trials and conducting a feasibility analysis questionnaire (results to be published soon).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Oviatt, S. Multimodal Interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*; L. Erlbaum Associates Inc.: Hillsdale, NJ, USA, 2003; pp. 286–304.
2. Sharma, R.; Pavlovic, V.I.; Huang, T.S. Toward multimodal human-computer interface. *Proc. IEEE* **1998**, *86*, 853–869. [CrossRef]
3. Quek, F.; McNeill, D.; Bryll, R.; Kirbas, C.; Arslan, H.; McCullough, K.E.; Furuyama, N.; Ansari, R. Gesture, speech, and gaze cues for discourse segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No. PR00662), Hilton Head Island, SC, USA, 15 June 2000; Volume 2, pp. 247–254.

4. Caschera, M.; D'Ulizia, A.; Ferri, F.; Grifoni, P. Towards Evolutionary Multimodal Interaction. In *On the Move to Meaningful Internet Systems: OTM 2012 Workshops: Confederated International Workshops: OTM Academy, Industry Case Studies Program*; Herrero, P., Panetto, H., Meersman, R., Dillon, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 608–616.
5. ETSI. *Human Factors (HF); Multimodal Interaction, Communication and Navigation Guidelines*; ETSI EG 202 191 V1.1.1; ETSI: Valbonne, France, 2003; p. 7.
6. Oviatt, S. Ten myths of multimodal interaction. *Commun. ACM* **1999**, *42*, 74–81. [CrossRef]
7. Martin, J.C. Towards intelligent cooperation between modalities. The example of a system enabling multimodal interaction with a map. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97) Workshop on Intelligent Multimodal Systems, Nagoya, Japan, 24 August 1997.
8. Oviatt, S. Multimodal interfaces. In *Handbook of Human-Computer Interaction*; Jacko, J., Sears, A., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 2002.
9. Tavanti, M. *Multimodal Interfaces: A Brief Literature Review*; EEC Note No. 01/07; EUROCONTROL: Brussels, Belgium, 2007.
10. Dumas, B.; Lalanne, D.; Oviatt, S. Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In *Human Machine Interaction: Research Results of the MMI Program*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 3–26.
11. Koons, D.B.; Sparrell, C.J.; Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*; Maybury, M.T., Ed.; American Association for Artificial Intelligence: Menlo Park, CA, USA, 1993; pp. 257–276.
12. SESAR. *The Roadmap for Delivering High Performing Aviation for Europe—European ATM Master Plan*; EU Publications: Brussels, Belgium, 2015.
13. Czyzewski, A. New applications of multimodal human-computer interfaces. In Proceedings of the 2012 Joint Conference New Trends in Audio & Video and Signal Processing: Algorithms, Architectures, Arrangements and Applications (NTAV/SPA), Lodz, Poland, 27–29 September 2012; pp. 19–24.
14. Neßelrath, R.; Moniri, M.M.; Feld, M. Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. In Proceedings of the 12th International Conference on Intelligent Environments (IE), London, UK, 14–16 September 2016; pp. 190–193.
15. Pfleging, B.; Schneegass, S.; Schmidt, A. Multimodal interaction in the car: combining speech and gestures on the steering wheel. In Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'12), Portsmouth, NH, USA, 17–19 October 2012; ACM: New York, NY, USA, 2012; pp. 155–162.
16. Bolt, R.A. 'Put-that-there': Voice and gesture at the graphics interface. In Proceedings of the 7th annual Conference on Computer Graphics and Interactive Techniques, Seattle, WA, USA, 14–18 July 1980; pp. 262–270.
17. Merchant, S.; Schnell, T. Applying Eye Tracking as an Alternative Approach for Activation of Controls and Functions in Aircraft. In Proceedings of the 19th DASC, Philadelphia, PA, USA, 7–13 October 2000; pp. 5.A.5-1–5.A.5-9.
18. Savery, C.; Hurter, C.; Lesbordes, R.; Cordeil, M.; Graham, T. When Paper Meets Multi-touch: A Study of Multi-modal Interactions in Air Traffic Control. In Proceedings of the 14th International Conference on Human-Computer Interaction (INTERACT), Cape Town, South Africa, 2–6 September 2013; Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M., Eds.; Lecture Notes in Computer Science, LNCS-8119 (Part III), Human-Computer Interaction. Springer: Berlin/Heidelberg, Germany, 2013; pp. 196–213.
19. Traoré, M.; Hurter, C. Exploratory study with eye tracking devices to build interactive systems for air traffic controllers. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero'16), Paris, France, 14–16 September 2016; ACM: New York, NY, USA, 2016. Article 6.
20. Alonso, R.; Causse, M.; Vachon, F.; Robert, P.; Frédéric, D.; Terrier, P. Evaluation of Head-Free Eye Tracking as an Input Device for Air Traffic Control. In *Ergonomics*; Taylor & Francis Group: Abingdon, UK, 2012; Volume 56, pp. 246–255.
21. Seelmann, P.-E. Evaluation of an Eye Tracking and Multi-Touch Based Operational Concept for a Future Multimodal Approach Controller Working Position (Original German Title: Evaluierung eines Eyetracking und Multi-Touch basierten Bedienkonzeptes für einen zukünftigen multimodalen Anfluglotsenarbeitsplatz). Bachelor's Thesis, DLR-Interner Bericht, Braunschweig, Germany, 2015.

22. Jauer, M.-L. Multimodal Controller Working Position, Integration of Automatic Speech Recognition and Multi-Touch Technology (Original German Title: Multimodaler Fluglotsenarbeitsplatz, Integration von automatischer Spracherkennung und Multi-Touch-Technologie). Bachelor's Thesis, Duale Hochschule Baden-Württemberg Mannheim in Cooperation with DLR, Braunschweig, Germany, 2014.
23. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th DASC, Sacramento, CA, USA, 25–29 September 2016.
24. Uebbing-Rumke, M.; Gürlük, H.; Jauer, M.-L.; Hagemann, K.; Udovic, A. Usability evaluation of multi-touch displays for TMA controller working positions. In Proceedings of the 4th SESAR Innovation Days, Madrid, Spain, 25–27 November 2014.
25. Möhlenbrink, C.; Papenfuß, A. *Eye-Data Metrics to Characterize Tower Controllers' Visual Attention in a Multiple Remote Tower Exercise*; ICRAT: Istanbul, Turkey, 2014.
26. Oviatt, S.; DeAngeli, A.; Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'97), Atlanta, GA, USA, 22–27 March 1997; ACM: New York, NY, USA, 1997; pp. 415–422.
27. Oviatt, S. Multimodal interactive maps: Designing for human performance. *Hum. Comput. Interact.* **1997**, *12*, 93–129.
28. Cohen, P.; McGee, D.; Clow, J. The efficiency of multimodal interaction for a map-based task. In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLC'00), Association for Computational Linguistics, Seattle, WA, USA, 29 April–4 May 2000; pp. 331–338.
29. Ohneiser, O.; Jauer, M.-L.; Gürlük, H.; Uebbing-Rumke, M. TriControl—A Multimodal Air Traffic Controller Working Position. In Proceedings of the Sixth SESAR Innovation Days, Delft, The Netherlands, 8–10 November 2016.
30. McMillan, D. Miscommunications in Air Traffic Control. Master's Thesis, Queensland University of Technology, Brisbane, Australia, 1999.
31. Cardosi, K.M.; Brett, B.; Han, S. *An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications*; (DOT/FAA/AR-96/66); DOT FAA: Washington, DC, USA, 1996.
32. ICAO. *ATM (Air Traffic Management): Procedures for Air Navigation Services*; DOC 4444 ATM/501; International Civil Aviation Organization (ICAO): Montréal, QC, Canada, 2007.
33. Ohneiser, O. *RadarVision—Manual for Controllers (Original German Title: RadarVision—Benutzerhandbuch für Lotsen)*; Internal Report 112-2010/54; German Aerospace Center, Institute of Flight Guidance: Braunschweig, Germany, 2010.
34. DLR Institute of Flight Guidance, TriControl—Multimodal ATC Interaction. 2016. Available online: http://www.dlr.de/fl/Portaldata/14/Resources/dokumente/veroeffentlichungen/TriControl_web.pdf (accessed on 6 April 2018).
35. Helmke, H.; Rataj, J.; Mühlhausen, T.; Ohneiser, O.; Ehr, H.; Kleinert, M.; Oualil, Y.; Schulder, M. Assistant-Based Speech Recognition for ATM Applications. In Proceedings of the Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 23–26 June 2015.