# Talky

## Speech-Only and Simulated Gaze + Speech Selection Techniques

Maximilian Ruppert (maximilian.ruppert@hs-augsburg.de)

Interaction Engineering

Advisor: Prof. Dr. Michael Kipp

Wintersemester 2016/17

Hochschule
Augsburg University of
Applied Sciences

Augsburg University of Applied Sciences

## Abstract

*Talky* presents two novel approaches of selecting and clicking discrete targets, like hyperlinks. The core mechanic is based on *Actigaze*, a gaze-based input method, which makes it possible to select and click a discrete target by only dwelling over it and its corresponding button, even when the target is small or closely surrounded by other targets. This is done by color-coding the targets when the user dwells over them. *Actigaze* therefore relies on gaze-tracking techniques. *Talky* alters this approach by presenting a *speech-only* and *simulated gaze + speech* variant. In the *speech-only* scenario the content is overlayed with a grid and the user has to say the box coordinates of his target's location to color-code all discrete targets inside that box. He then only has to say the color in which his target is now highlighted to trigger the click. In the *simulated gaze + speech* variant the grid is no longer need because the user dwells over a target group with his eye and mouse cursor (this is the simulation part) to color code the discrete targets. Compared to *Actigaze* both variants of *Talky* are slower, but still offers unique use-cases.

## Introduction

On today's computers the main input devices are mouse and keyboard, as they have been for decades. With the launch of the iPhone another method of input, the touchscreen, became mainstream roughly ten years ago. Since then nothing really changed in the way we interact with computers on a daily basis. Only recently newer, more sophisticated interaction techniques, like eye-tracking and voice commands become available for the average consumer, thanks to progress in technology and the fast rise of machine learning algorithms. This opens up a whole new way of possibilities on how humans can interact with machines. Ways that are way more natural, like language and gaze, then moving a cursor to a certain position with an external device. With these technologies it finally becomes possible for people to use computers when they are restricted from using traditional input devices, may it be in scenarios were it is impossible due to hygienic reasons or the user just needs both hands else were, for example in medical applications or simply when cooking. It also opens up new ways for humans with physical disabilities to interact with computers and makes it possible for them to take part in the modern information age in a more inclusive way.

In this work I present two different approaches of interacting with everyday information in the form of webpages in a system called *Talky*. Both usage scenarios focus on ways of accomplishing one of the key aspects that made the internet what is today: selecting and clicking a hyperlink, to access content. The main challenge to achieve this objective is to find a way make it easy for the user to select a certain link even if it is surrounded by other targets or simply very small to be selected with gaze pointing alone and trigger the click without using any other input then voice or sight. The first approach to select a target is speech-only solution and the second a combination of gaze-tracking and speech. Both scenarios heavily rely on the work of a team of researchers that tried to explore the use of gaze-pointing in order to achieve the same goal this report tries to tackle.

## Related Work

In the paper "Gaze vs. Mouse: A Fast and Accurate Gaze-Only Click Alternative" Christof Lutteroth, Moiz Penkar, Gerald Weber from the University of Auckland developed a system called *Actigaze*, that allows a user to select a discrete target and click it using gaze-pointing. After evaluating previous direct- and indirect-click alternatives based on gaze tracking, they found one that worked as the foundation of their own system called *Multiple Confirm*. Here the user gazes over a certain are on the screen and all discrete targets in that dwell area get recognized by the system and individually displayed in a column next to the main content. In this sidebar each target gets its own distinct and relatively big click activation button and next to it the text of the target. The user can now gaze over one of these buttons to trigger the actual click on the target. One problem with *Multiple Confirm* is instability of the button order when dwelling over a discrete target. If a user dwells over a target from left to right, the order is different to when he gazed over it from right to left, causing confusion for the user and a drop in execution time. *Actigaze* addresses this problem by implementing stable, color coded confirm buttons. Each button has its own, very unique color and the order of those buttons is never changed, no matter what the dwell direction might be. This helps the user to learn and remember where each button is located in the sidebar and helps him increase his overall click time. To make use of the colors in the content area the team developed two different variants of usage, *dynamic coloring* and *static coloring*. Both share the same core principal of usage. First the user dwells over a certain area containing discrete targets. The system stores information about which targets are in the dwell area and assigns each target to one of the sidebar buttons in order of appearance. The user then looks over to the sidebar and gazes over the button that is dyed in the color of the link he intended to click. Finally he gazes over the selected button that represents the selected link for a small amount of time to trigger the click.

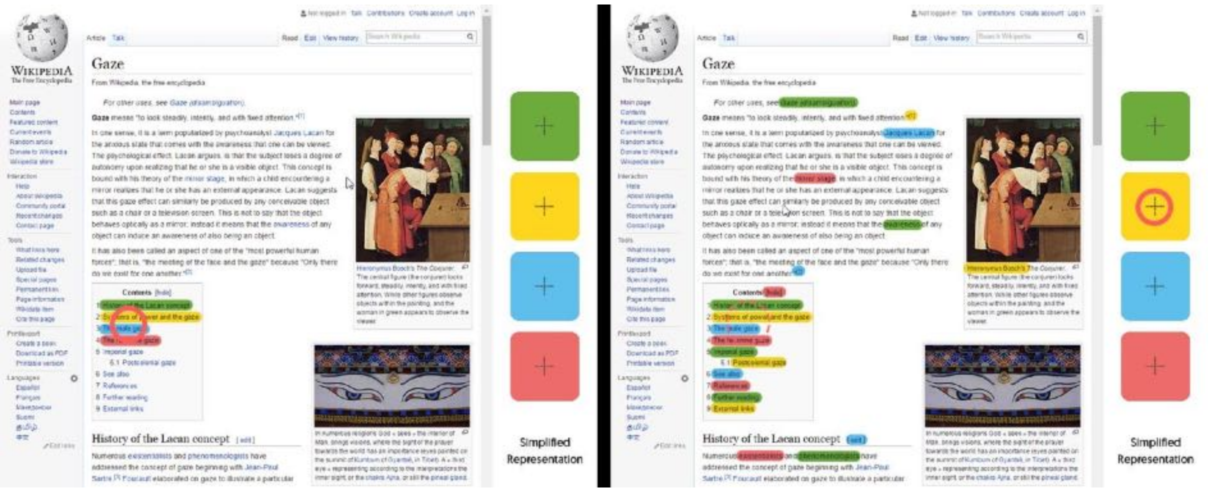

Wikipedia
The Free Encyclopedia
Gaze
History of the Lacan concept
Simplified Representation
Wikipedia
The Free Encyclopedia
Gaze
History of the Lacan concept
Simplified Representation

Figure 1. Actigaze dynamic coloring and static coloring variant

The two variants differ in the way they use color in the selection process. Using the *dynamic coloring* version of *Actigaze* initially all targets keep their original color, e.g. hyperlinks are in a certain shade of blue. Only when the user dwells over them the targets inside the dwell radius get color coded and keep their color until the user triggers a click or dwells over another area for a certain time. The advantage of this method is that content is not altered too much and keeps it original look and feel, but the dynamic change of target colors can lead to some distraction. In the *static coloring* variant all targets are color coded from the beginning. The user only has to dwell over a certain area and the system stores the information of what targets the users had been looking at internally without visual feedback. In this variant the content is heavily altered, but this can help the user in keeping track what color each target has. To evaluate their variants compared to *Multiple Confirm* and traditional mouse clicks they conducted a within-subjects test, were every test-user had to find, select and click a highlighted hyperlink on a Wikipedia page.

## Prototype

All this knowledge about *Actigaze* and its evaluation method is used as a foundation for the development and evaluation of *Talky*. The first idea was to replicate the *dynamic coloring* variant of *Actigaze* with one major difference: the sidebar with the color boxes should be replaced by a speech base input system. So in order to click a target, the user would only have to dwell over a certain area to get all the targets in that area color coded. He then says the color he wants to select to trigger the click. One extra requirement to this prototype was to achieve all that using only commonly available hardware. Were *Actigaze* used an expensive and hard to calibrate eye-tracking device, Talky should only rely on a high definition webcam and a microphone, to be as affordable as possible. Due to the requirement of making discrete targets on web pages selectable and clickable web technologies such as HTML 5, CSS and Java Script were used in creating the prototype, so the whole application could run in a modern web browser. The targets inside the dwell area are color coded in order of their appearance from left to right and top to bottom in the following colors: red, blue, yellow, green, purple, orange, grey, brown.

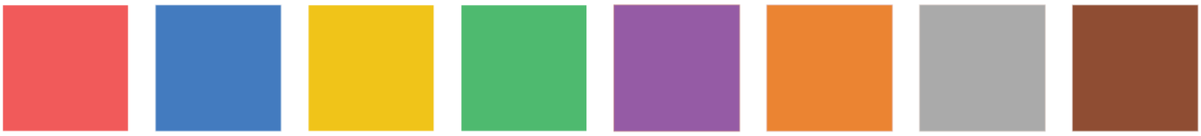
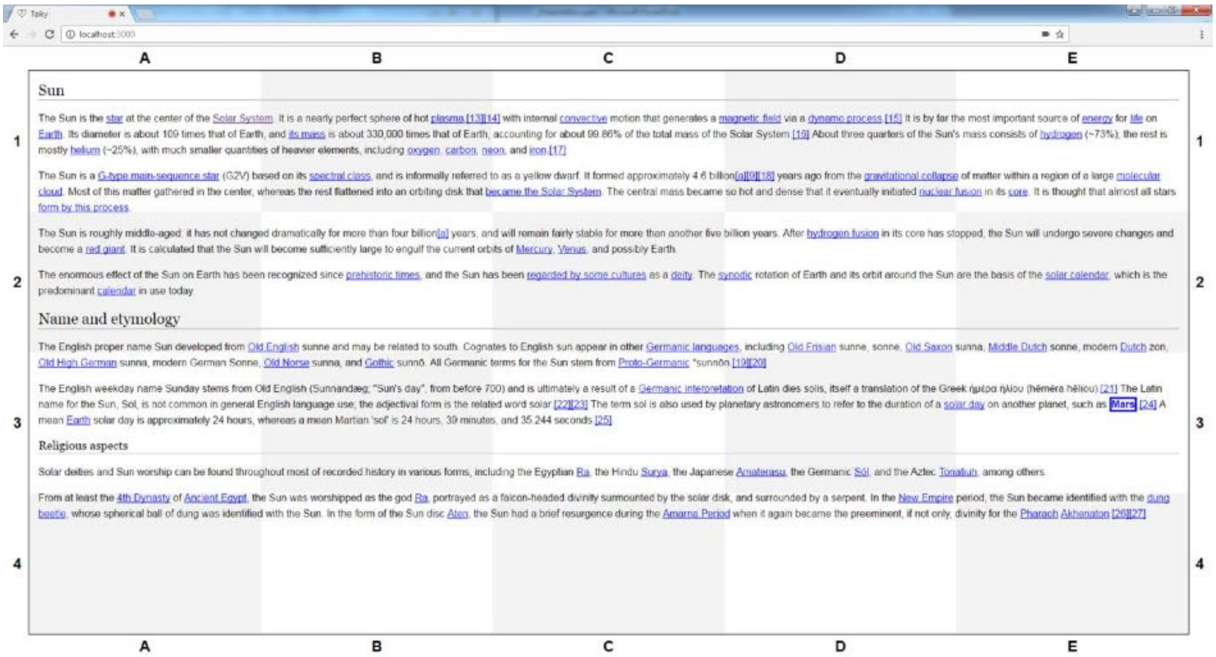


Figure 2. Talky colors

This order ensured that colors neighboring each other differ in contrast and are easily identifiable by the user. Because all of these colors have a very distinct name and sound to them, the speech computation framework can differentiate them with a higher precision.

In an early version of the prototype the framework *WebGazer.js* was utilized to gain access over the user's webcam and track the movement of his eyes. *Annyang.js*, a framework for speech recognition, which functions as a wrapper for the Web Speech API, was used to compute the voice commands of the user. After a short test period the initial plan of how *Talky* should have worked had to be updated. This change had to be made because of some technical problems with *WebGazer.js*. This framework needed mouse movement as an additional parameter to track the user's eye, which could not be simulated in the code, and in combination with speech recognition was so computational intensive that it slowed down the whole system. Therefor the use of a webcam as an eye-tracking device was not feasible anymore. Apart from the problems with eye-tracking the first version of the prototype promoted another shortcoming. The speed in which words were recognized by *Annyang.js* was too slow for real usage, even so voice recognition worked in general. This led to a change in frameworks from *Annyang.js* to *Artyom.js*, which was much faster in presenting a result from speech input.



Sun

The Sun is the star at the center of the Solar System. It is a nearly perfect sphere of hot plasma,[13][14] with internal convective motion that generates a magnetic field via a dynamo process.[15] It is by far the most important source of energy for life on Earth. Its diameter is about 109 times that of Earth, and its mass is about 330,000 times that of Earth, accounting for about 99.86% of the total mass of the Solar System.[16] About three quarters of the Sun's mass consists of hydrogen (~73%); the rest is mostly helium (~25%), with much smaller quantities of heavier elements, including oxygen, carbon, neon, and iron.[17]

The Sun is a G-type main-sequence star (G2V) based on its spectral class, and is informally referred to as a yellow dwarf. It formed approximately 4.6 billion[a][9][18] years ago from the gravitational collapse of matter within a region of a large molecular cloud. Most of this matter gathered in the center, whereas the rest flattened into an orbiting disk that became the Solar System. The central mass became so hot and dense that it eventually initiated nuclear fusion in its core. It is thought that almost all stars form by this process.

The Sun is roughly middle-aged: it has not changed dramatically for more than four billion[a] years, and will remain fairly stable for more than another five billion years. After hydrogen fusion in its core has stopped, the Sun will undergo severe changes and become a red giant. It is calculated that the Sun will become sufficiently large to engulf the current orbits of Mercury, Venus, and possibly Earth.

The enormous effect of the Sun on Earth has been recognized since prehistoric times, and the Sun has been regarded by some cultures as a deity. The synodic rotation of Earth and its orbit around the Sun are the basis of the solar calendar, which is the predominant calendar in use today.

Name and etymology

The English proper name Sun developed from Old English sunne and may be related to south. Cognates to English sun appear in other Germanic languages, including Old Frisian sunne, sonne, Old Saxon sunna, Middle Dutch sonne, modern Dutch zon, Old High German sunna, modern German Sonne, Old Norse sunna, and Gothic sunnō. All Germanic terms for the Sun stem from Proto-Germanic *sunnōn.[19][20]

The English weekday name Sunday stems from Old English (Sunnandæg; "Sun's day", from before 700) and is ultimately a result of a Germanic interpretation of Latin dies solis, itself a translation of the Greek ἡμέρα ἡλίου (hêméra hēlíou).[21] The Latin name for the Sun, Sol, is not common in general English language use; the adjectival form is the related word solar.[22][23] The term sol is also used by planetary astronomers to refer to the duration of a solar day on another planet, such as Mars.[24] A mean Earth solar day is approximately 24 hours, whereas a mean Martian 'sol' is 24 hours, 39 minutes, and 35.244 seconds.[25]

Religious aspects

Solar deities and Sun worship can be found throughout most of recorded history in various forms, including the Egyptian Ra, the Hindu Surya, the Japanese Amaterasu, the Germanic Sól, and the Aztec Tonatiuh, among others.

From at least the 4th Dynasty of Ancient Egypt, the Sun was worshipped as the god Ra, portrayed as a falcon-headed divinity surmounted by the solar disk, and surrounded by a serpent. In the New Empire period, the Sun became identified with the dung beetle, whose spherical ball of dung was identified with the Sun. In the form of the Sun disc Aten, the Sun had a brief resurgence during the Amarna Period when it again became the preeminent, if not only, divinity for the Pharaoh Akhenaton.[26][27]

Figure 3. Speech-Only test setup

With speech recognition working two new scenarios for using *Talky* were implemented, one only relying on speech input and the second one using speech in combination with a simulated form of eye-tracking similar to *Actigaze*. The *simulated gaze + speech* was implemented after the first version of the prototype had been tested which only used the *speech-only* scenario. The second and final version of Talky was not only fitted with both usage scenarios, but had undergone some minor tweaks, like adapting the color order and optimizations to address speech recognition problems, for words like “green” were the algorithm failed to reliably identify it properly.

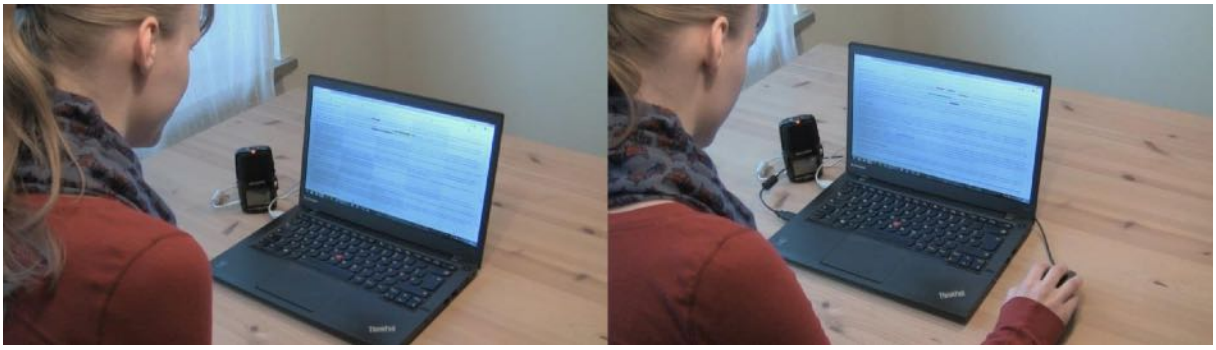


Figure 4. *Talky* variants in use

## Interaction Mechanisms

In the *speech-only* scenario the content gets overlayed by a grid, similar to a cheeseboard, with the coordinates A-E (columns) and 1-4 (rows). To select a target, the user first says the coordinates of the box he were the link is located. All targets, in this case hyperlinks, which are located inside the box get color coded. The targets in a box stay colored as long as user has not selected another box or said the special command “reset”, which resets all targets to their initial state. To trigger the click the user then has to say the color of target he wants to select.

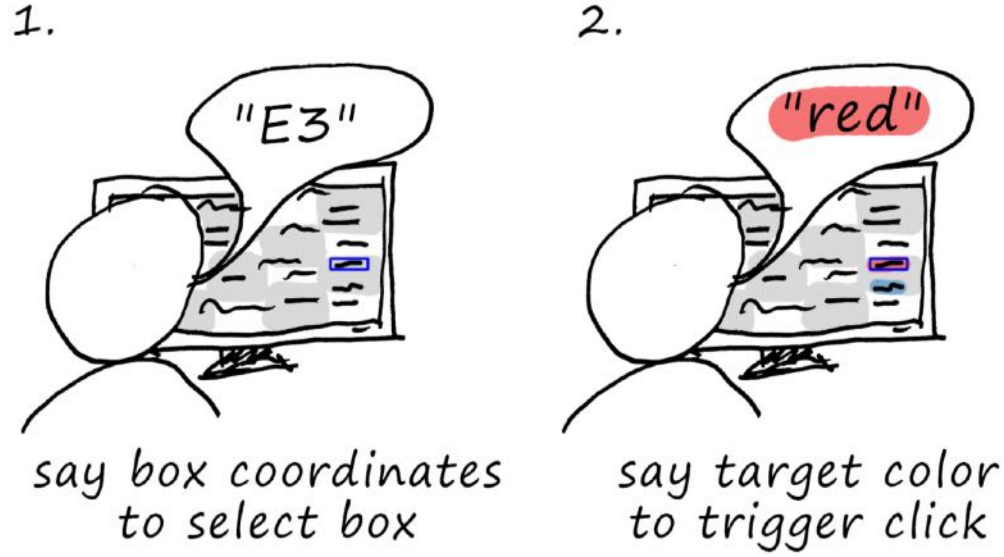

1.
"E3"
say box coordinates to select box
2.
"red"
say target color to trigger click

Figure 5. Speech-Only

In the *simulated gaze + speech* scenario the user can select targets in a certain dwell area by looking at them. In this case the grid is not overlayed. Because of the previously mentioned technical shortcomings of this prototype the gaze point has to be marked by moving the mouse cursor in form of a crosshair to the position the user is looking at. This aims to simulate the gaze aspect. 50 milliseconds after the cursor came to a halt all targets inside an invisible square with the size of 150 pixels with the cursor positon as the center get color coded. The user then has to say the color of the target he likes to select and click, like in the previous scenario.

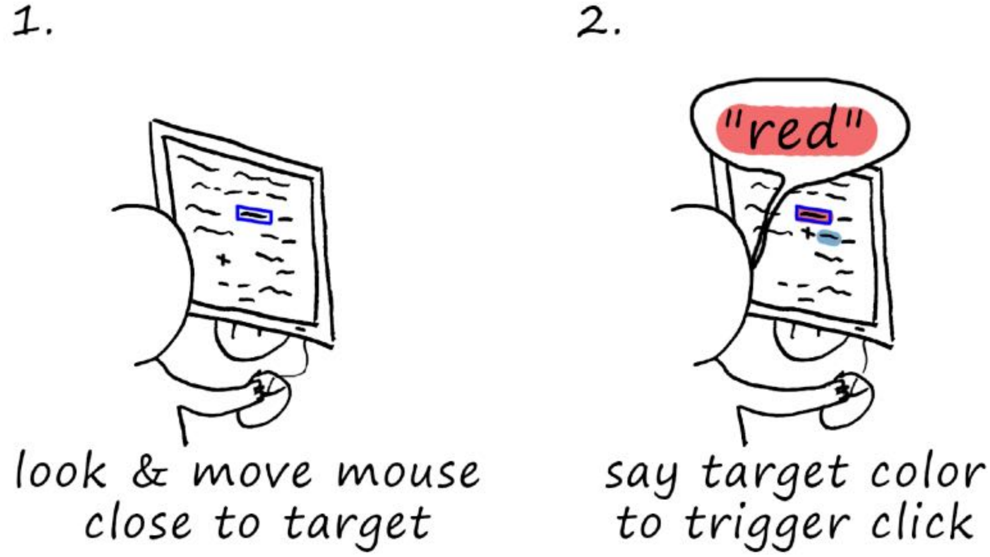
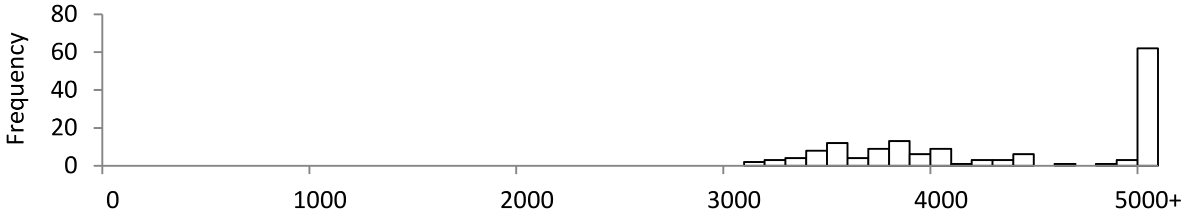

1.
look & move mouse close to target
2.
"red"
say target color to trigger click

Figure 6. Simulated Gaze + Speech

## Evaluation

To evaluate the success of these two new scenarios and to compare them to *Actigaze* a user study was conducted. The microphone used for this test was a *Zoom H2n* with a surround sound characteristic. As mentioned earlier the prototype was tested twice, in both cases with five people. The first version, tested by people between the age of 21-27, one of them with red-green-colorblindness, one with a problem of identifying shade of color and one English native speaker, only featured the *speech-only* scenario. Every participant did 5 training tasks before doing 30 timed tasks. The targets in this test were laid out in way that every box had to be selected at least once.  Before every task started a countdown from 3 to 0 was displayed, the same way it was done in *Actigaze*. Once the user had selected and clicked the target hyperlink, marked by a big blue border, a new countdown would begin. All test results were stored in a CSV-file for further evaluation. In this file the time it took the user to complete a whole task as well as the name of the box and the target color were saved.

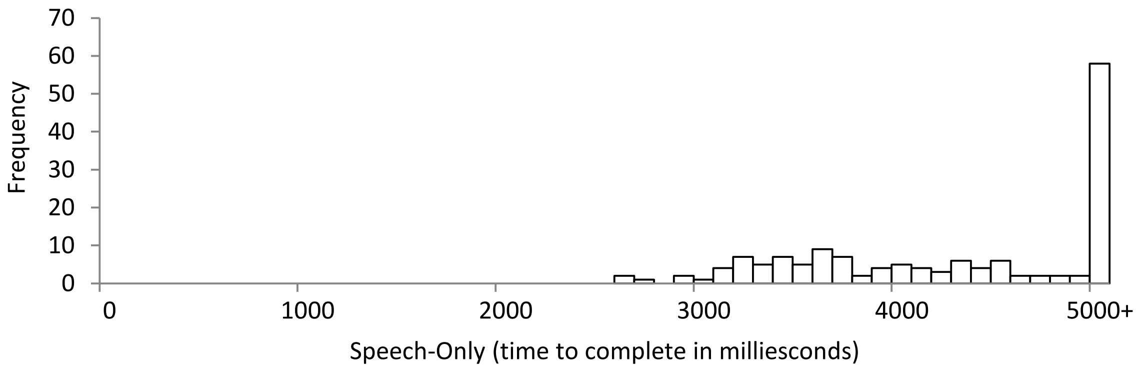
With a median task completion time of 6.7 seconds *Talky* was very slow compared to all other methods mentioned in the paper of the *Actigaze* team.

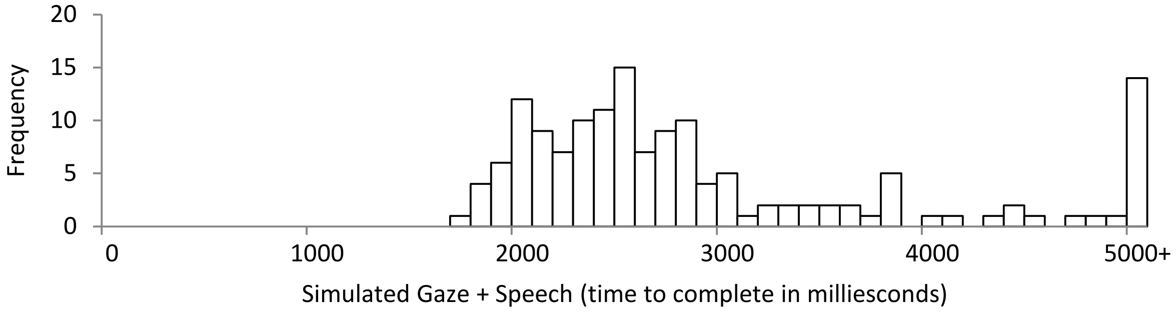


Speech-Only Talky Version 1 (time to complete in milliesconds)

Besides the slow execution time, the prototype suffered from problems in understanding the word green, the two color blind people sometimes had a little bit of trouble identifying the color green, the prototype sometimes misunderstood the characters of the coordinates, like the word B4 as before or B as E and had some problems with the German pronunciation of the word three. Only one of the users noticed that the color order was always the same.

The second, overhauled version of *Talky* was then also tested by five people between the age of 21-30. One of the users is an English native speaker and one has a red-green-colorblindness. This time both *speech-only* and *simulated gaze + speech* were tested.  The *speech-only* variant benefited a lot from the later improvements of version two and the median time to complete a tasked drop from 6.7 sec to 5.44 sec, which is still slow compared to Actigaze but an 18 % increase in speed overall. Using the *simulated gaze + speech* variant it took users only a median time of 3,01 seconds to complete a task.





In comparison with *Actigaze* and mouse clicks, *Talky* was considerably slower than the there other method and at this point cannot be considered a real alternative to any of those methods.

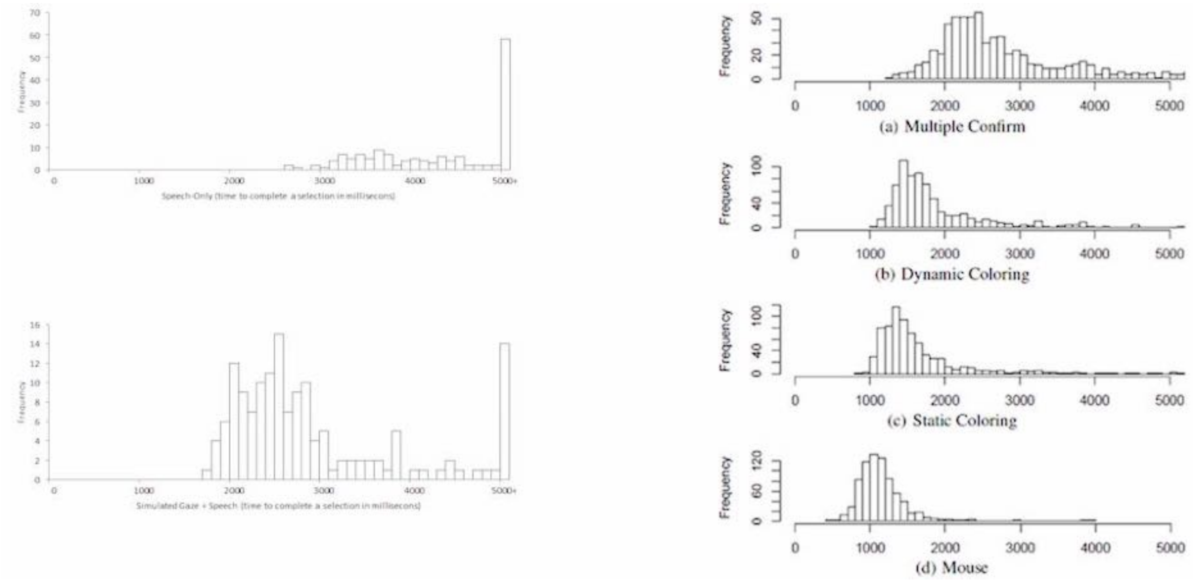


Figure 7. Talky vs. Actigaze vs. Mouse

| Talky | | Actigaze | | |
|---|---|---|---|---|
| Speech-Only | Simulated Gaze + Speech | Dynamic Coloring | Static Coloring | Mouse |
| 5,44 s | 3,01 s | 1,67 s | 1,46 s | 1,10 s |

Figure 8. Median times comparison

Besides the improvements from the previous versions both variants still suffered from technical shortcomings in sense of microphone disconnects, voice recognition errors and sometimes slow voice analyzation speeds. Another major problem was the open-mic and continuous evaluation of sound. As soon as a user said something different than coordinates and colors the speech recognizer tried to match it against his target words and often chained words together instead of restarting his evaluation process, which led to some of the slow execution times.

One positive aspect that users of both teste mentioned was, besides all problems, the ease of use and fun they had using it when it worked properly.

## Conclusion

*Talky* has still a long way to go in its development of becoming a new input alternative. In situations where ambient noises are controlled or relatively mute, eye-tracking is not an option or execution time is not of the essence *Talky* could still be a viable alternative to *Actigaze*. It would be interesting to see the *simulated gaze + speech* variant of *Talky* with actual gaze-tracking equipment and once again compare it to *Actigaze*. With speech recognition getting better and better it is only a question of time when a system like *Talky* can be used in a fast and reliable way and maybe become the input device of the future.