Text input tools' complementarity in immersive virtual environments

Aurélien Agnès[1], Sylvain Fleury[2], Aristide Auzerais[3], Isaline Bisson[3], Eva Dulau[3], Stéphanie Buisine[4], Simon Richir[5]

[1] Institut Laval Arts et Métiers
Laval Virtual Center, rue Marie Curie
53810 Changé
aurelien.agnes@ensam.eu

[2] Institut Laval Arts et Métiers
Laval Virtual Center, rue Marie Curie
53810 Changé
sylvain.fleury@ensam.eu

[3] Institut Laval Arts et Métiers
Laval Virtual Center, rue Marie Curie
53810 Changé

[4] CESI LINEACT
93 boulevard de la Seine
92000 Nanterre
sbuisine@cesi.fr

[1] Institut Laval Arts et Métiers
Laval Virtual Center, rue Marie Curie
53810 Changé
simon.richir@ensam.eu

ABSTRACT. *This study presents a user test in order to ascertain the advantages and disadvantages of three different text input methods in immersive virtual environment: individual Speech-to-Text, collective Speech-to-Text and a virtual keyboard named Drum-Like Keyboard. We measured participants' user experience, especially related to usability and utility, in order to offer relevant recommendations to people seeking to integrate text input in virtual reality. Our results show that Speech-to-Text and the virtual keyboard have complementary qualities, which can be used together for optimal results and experience.*

RÉSUMÉ. *Cette étude présente un test utilisateur afin de déterminer quels sont les avantages et inconvénients de différents modes de saisie de texte en environnement virtuel immersif : la reconnaissance vocale individuelle, la reconnaissance vocale collective et le clavier virtuel surnommé Drum-Like Keyboard. Nous avons mesuré l'expérience utilisateur des participants notamment selon l'utilisabilité et l'utilité afin de pouvoir proposer des recommandations adéquates aux personnes cherchant à intégrer la saisie de texte en réalité virtuelle. Nos résultats montrent que la reconnaissance vocale et le clavier virtuel ont des qualités complémentaires, qui peuvent être utilisées de concert pour obtenir des résultats et une expérience optimale.*

KEYWORDS: *virtual reality, text input, innovation, Drum-Like Keyboard, Speech-to-Text*

***MOTS-CLÉS**: réalité virtuelle, saisie de texte, innovation, Drum-Like Keyboard, Speech-to-Text*

1. Introduction

According to recent studies, we know that virtual reality (VR) is more efficient to create new objects concepts than pen and paper [Yang et al., 2018] or computer aided design [Feeman et al., 2018] for a sole person. However, creativity sessions are traditionally based upon collective methods, such as the brainstorming. During the latter, ideas are usually collected in written form, be it Post-it® notes, mind maps or reports. In VR, this need of note taking, meaning conveying to other people what happened in a creativity session and keeping a written evidence, remains difficult to meet. Indeed, if VR can help us improve performances for creativity tasks, written evidence may still remain the best way of communicating to stakeholders all retained ideas. That is how we came to wonder about text input from the virtual world to the physical world, integrated inside the VR users' immersive virtual environment (IVE). Indeed, besides avoiding asking additional efforts to write a report after all participants are out of the IVE, such integration would allow to actively involve all participants. We did not want for another individual, exterior to the session, to be present only to write the report either.

Yet, according to another study [Jimenez, 2017], text input in VR is a problem for which no conventional solution has really been accepted. Moreover, VR users being isolated from the outside world when wearing headsets, it would be uncomfortable to try and use a physical keyboard to take notes or write ideas as it is possible in the physical world. We will thus see that there are two frequently used solutions: Speech-to-Text (STT) and the use of a virtual keyboard. The need for studying both is stressed by the fact that previous studies did not use recent technologies available to us today. To this end, we conducted a user test to highlight the main qualities of the aforementioned solutions.

2. Text input in IVE: a necessary update?

2.1. The drawbacks of Speech-to-Text

In 2002, Bowman and his colleagues [Bowman et al., 2002] compared four input text techniques to ascertain which allowed to be fast, make fewer mistakes, provide comfort and satisfaction, but also which were easy to learn. Those four techniques are:

- The STT
- The "Pinch Keyboard", using a data glove
- The "Chord Keyboard", where each subject had half a keyboard in either hand
- A physical tablet with its virtual equivalent in the IVE

For the latter, results showed that it led to arm fatigue despite its good performances. This comforts us in our decision to create a tool entirely integrated in VR. The STT was introduced using the wizard of Oz technique[1], and participants had to spell out words instead of saying them. Meanwhile, another study studied the use of a real STT software compared to a keyboard input [Karat et al., 1999]. The result was that the STT was far behind in terms of precision and typing speed. Moreover, the inherent lack of precision made it very difficult to correct mistakes using only the software.

[1] Meaning participants did not know there was no real Speech-to-Text software, and that the input was made by an experimenter

According to GlobalWebIndex, 33% of respondents had recently used voice recognition, during an Internet survey in 2017[2]. If a third of the population has already recently used this technology, this shows some interest, or at least a consideration for it. Did its performance improve over the last few years?

Today, voice recognition has been improved thanks to the integration of more effective deep learning softwares [Amodei et al., 2015]: their error rates were smaller than those of keyboard input. Another interesting value in favor of the STT against the keyboard is the potential speed of text input: speech is approximately 3 times faster than keyboard input for the English language [Ruan et al., 2016]. We thereby have more reasons to be interested in this technology.

2.2. The virtual keyboards

On the other hand, we have virtual keyboards. A recent thesis in 2018 [Kongsvik, 2018] compared several text input tools in IVE, and was able to ascertain the most efficient and appreciated-by-users tool. Between selecting letters with your gaze, pretending to play battery on a keyboard, pointing the keyboard with a virtual laser and having half a keyboard per controller, the battery won in almost all criteria. Performances (word per minute written, error ratios) and user experience (usability, flow, tension, etc) were measured. The immersive dimension of the technologies was very appreciated from users. In the end, we chose to use the “Drum-Like Keyboard” (DLK), in order to have the best tool to date.

In conclusion, we know that the voice recognition, as well as the DLK, are two technologies that users appreciate and that offer satisfactory performances. However, it does not help decide which one to use. Moreover, they have not been compared, especially in terms of user experience, and we want to define which technology is the most powerful for text input in VR.

3. The experiment

3.1. Protocol

In order to evaluate the performances and preferences for those three tools, a user test was designed to gather users’ preferences, according to three conditions of text input:

- In the first one, participants used the virtual keyboard DLK (see figure 1);
- In the second one, voice recognition was only used individually for the input;
- In the last one, voice recognition was used collectively: as soon as it was activated, everything said by a member was transcribed (see figure 2 for the two types of voice recognition).

The voice recognition software used is derived from Windows’ dictation tool.

[2] https://www.globalwebindex.com/reports/trends-18

Figure 1. *Drum-Like Keyboard, as developed in the experiment's application*

Figure 2. *Both tools of voice recognition: the collective one on the left, and the individual on the right, with a participant talking in the microphone.*

In order to implement them in a situation, participants were invited to take the exercise of the "defence" by groups of three while wearing a VR headset. This exercise entails to choose a well-known person, be it fictional or not. After this choice, participants were given a context for the exercise: during an apocalypse, they form a group of survivors. Each participant must elaborate why their character is the best to lead the group. The goal is for them to express themselves as much as possible, and take notes on the points put forward by each of them inside the virtual environment. They were free to choose the modality to do so, i.e. they could take notes one after another, or someone could do it all. Each group could only use one tool. Then, they proofread and corrected their report outside of the immersive environment. Finally, they could test the other tools to compare them with their group. Six groups of three were recruited for this experiment. They all know VR and came from the same school environment. Hence two groups evaluated each tool. 14 men and 4 women, with 15 between 18 and 25 years old and 3 between 26 and 35 years old, participated. The application was developed under Unity3D, and the equipment used for VR was a HTC VIVE, its controllers and its stations.

3.2. Measures

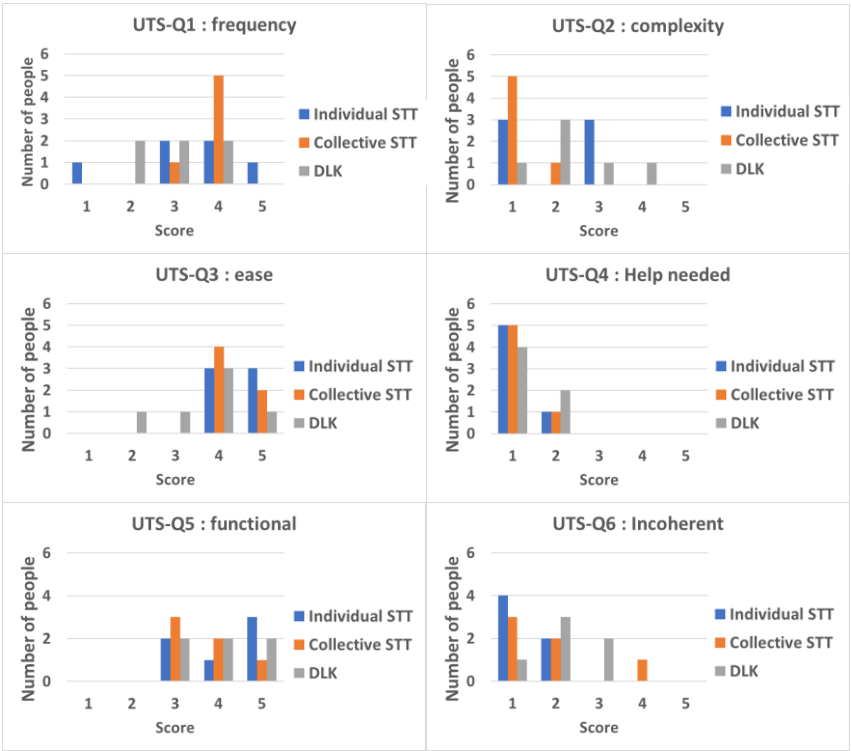
Data were collected through a questionnaire (see in appendix) as well as interviews. The first four questions were related to demographics. The other measures were:

- The perceived usability (UTS) measured with the System Usability Scale [Brooke, 1996];

- The perceived utility (UTT) according to five functionalities and uses: note taking, note exporting, report redaction, gathering and communication. Each one was rated on a 5-point scale;
- Satisfaction (based on the meCUE – Emotion [Lallemand et al., 2017]), stimulation and identity (both measured with the French translation of the AttrakDiff QHS and QHI [Lallemand et al., 2015]);
- Spoken suggestions of users during the experiment. Indeed, all those which were related to the tools were collected during the experiment.

3.3. Results

The perceived usability (see figure 3) ranged from 73 to 86% for the three tools, meaning they all have a **good** usability. In ascending order: Drum-Like Keyboard (73.3%), individual voice recognition (80%) and collective voice recognition (81.25%). A main phenomenon is responsible of this result: several subjects reported to us that the DLK requires a lot of effort in terms of attention. This can isolate some of them during the experiment and cause them not to hear what is being said.



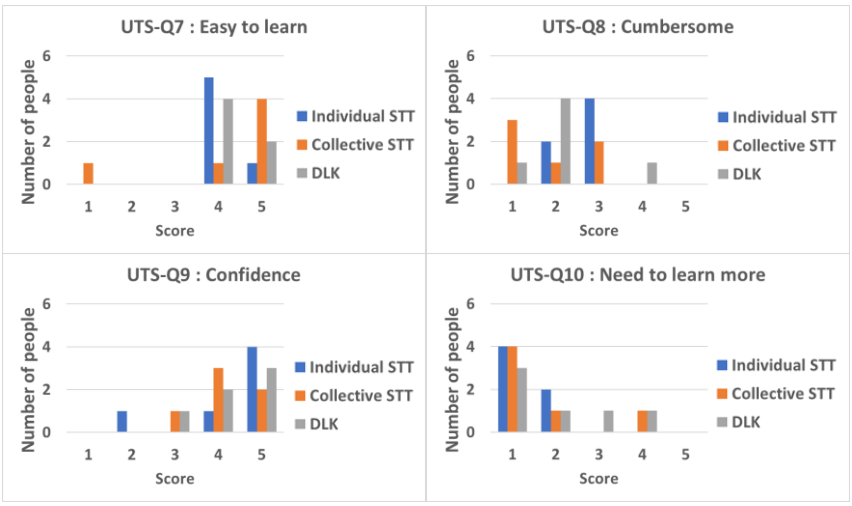


Figure 3. *Answer distribution for the perceived usability of the three tools on a 5-point Likert scale.*

The three dimensions of utility (see figure 4) that bring forth the best scores (17 out of 18 participants gave at least a score of 4 on the 5-point scale), whatever the tool used, are note export, gathering and communication. This means that our three tools fulfilled 3 out of the 5 functionalities we were interested in. However, for note export, one participant found individual STT not particularly well adapted (2/5 score). Furthermore, both voice recognition tools scored higher that the DLK: one participant gave the latter 2/5. The trend is the same for communication: one participant gave a 3/5 score for the DLK, against 4/5 for the other tools.

In the two other dimensions, the results are more mixed: if the collective STT is poorly rated, the individual STT is well rated, while the DLK moderately for note taking. For the latter, three persons gave it 2/5, while two gave it 5/5. Regarding collective STT, some report that too much irrelevant content was recorded because it records everything. Regarding DLK, it does not allow fast text entry and can quickly become tiring for users.

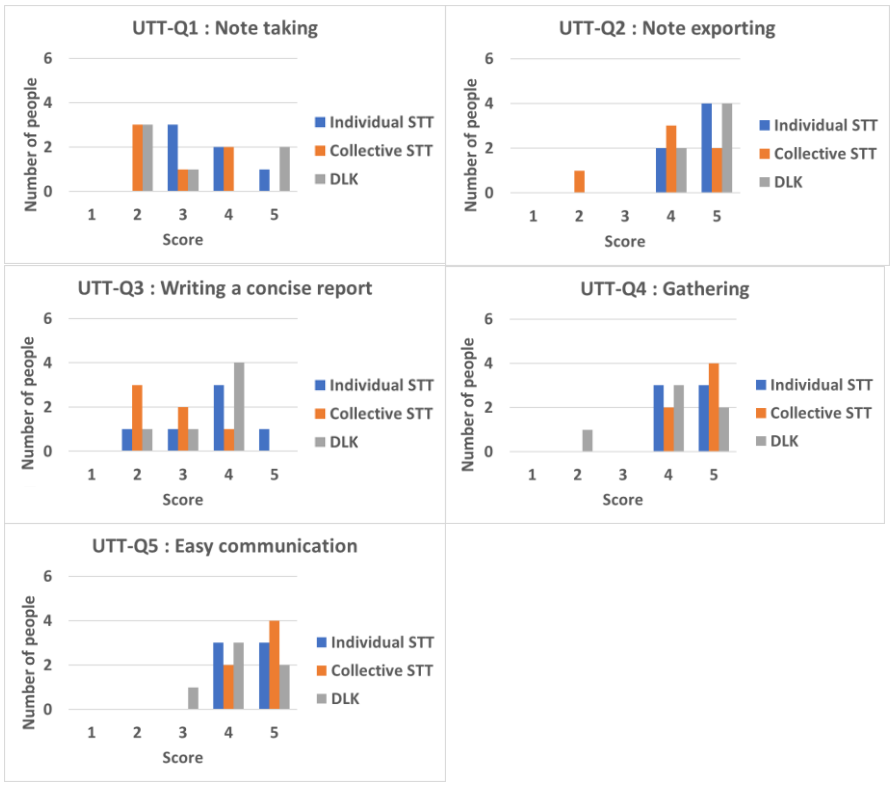
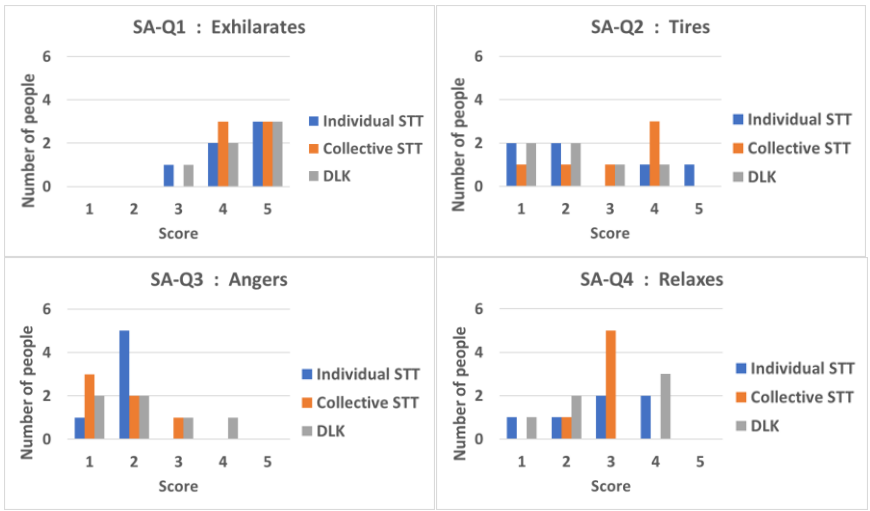


Figure 4. *Answer distribution for perceived utility of the three tools on a 5-point Likert scale (disagree/agree).*

As we can see in figure 5, most of the satisfaction criteria gave mixed results. However, the three tools generated enthusiasm, which is positive for new forms of text input, and did not frustrate users.



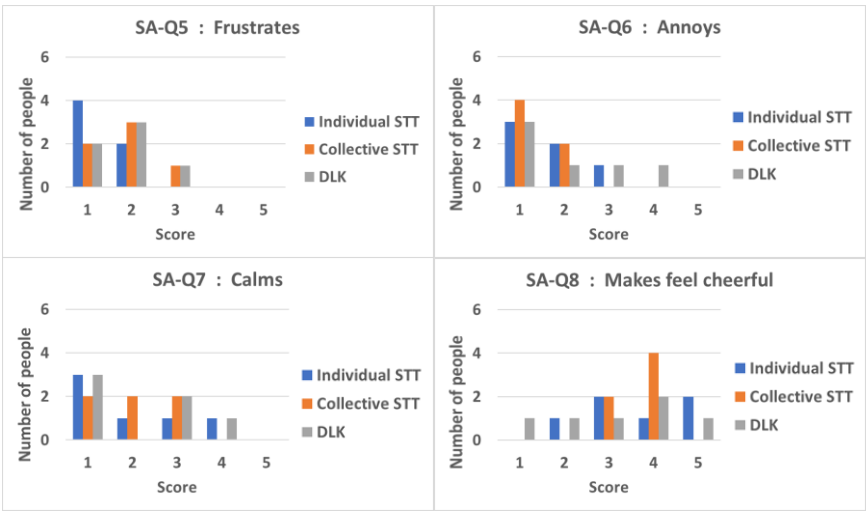


Figure 5. *Answer distribution for satisfaction criteria for the three tools on a 5-point Likert scale (disagree/agree).*

The stimulation criteria (see figure 6) highlight that voice recognition is considered as more original that the virtual keyboard. Collective STT is also judged as more "inventive" than individual STT, itself more creative than the virtual keyboard – which is even considered as "unimaginative" by some. The "innovative" character of all tools is also particularly highlighted. Finally, if the tools seem rather "captivating" and "undemanding", some find the DLK particularly "cautious", while others find it "bold".

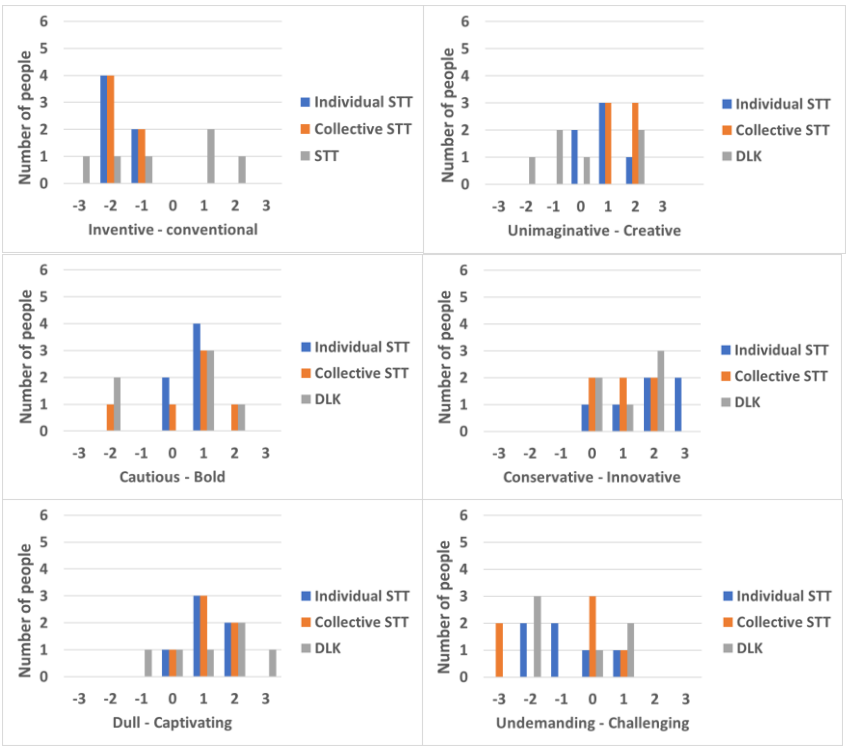


Figure 6. *Answer distribution for the stimulation of the three tools on a 7-point Likert scale.*

Finally, the identity (see figure 7) allows us to identify that with our tools, users generally feel the application is “connective”, “connects” and “integrates” them. One can however note that for the first two dimensions, isolated users considered on the contrary the application as “insulating” and “excluding” them. One can also note it does not concern collective STT at all. If the dimensions “premium – cheap” and “unprofessional – professional” are relatively centered on the mean, the application is generally considered as “stylish” and “presentable”.

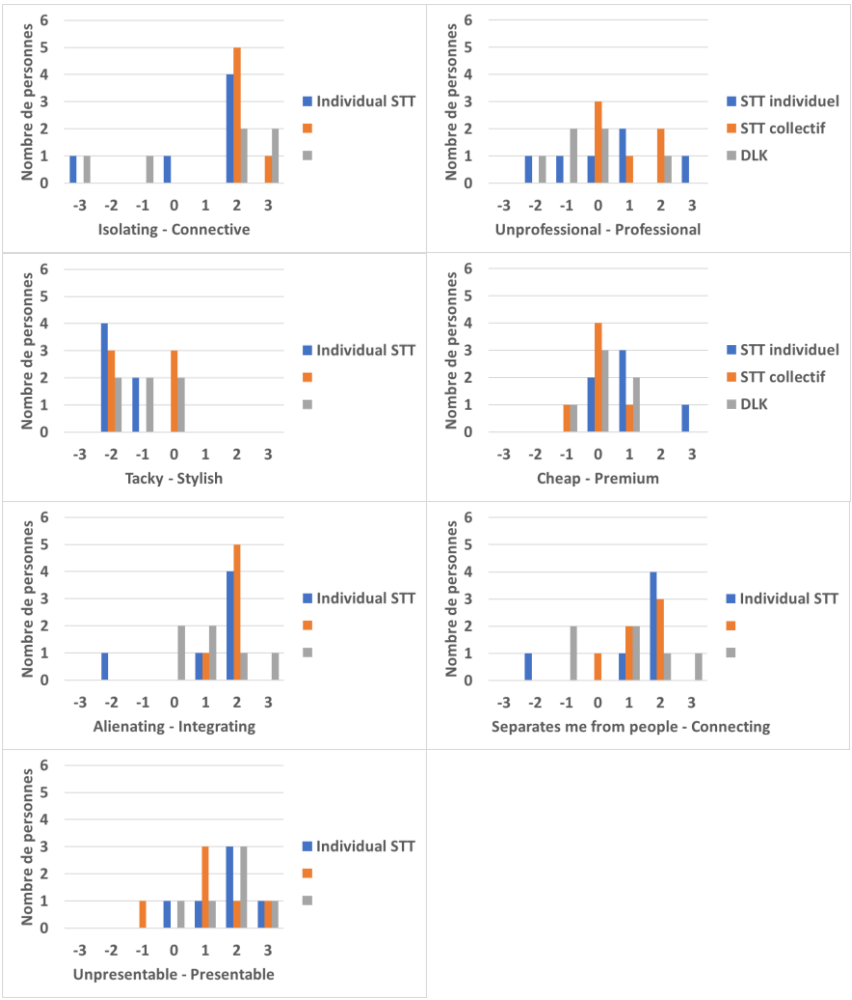


Figure 7. *Answers distribution for the identity of the three tools on a 7-point Likert scale.*

Finally, spoken comments have been very instructive. DLK induced a lot of positive remarks: the words “fun”, “funny” and “cool” together were mentioned by nine participants. However, DLK also requires more cognitive resources than the other tools. Indeed, participants reported having to cut themselves off from the discussions in order to be able to write, especially because they needed more time to write (10 participants). Regarding collective STT, the main point which arose was the need to plan if not the task in the IVE, at least the way users will organize text input (10 participants). Thereby, the report will contain more relevant content. Individual STT did not induce any particular comment. However, it caused some very interesting behavior. Indeed, it is represented as a microphone and participants have to hold the controller’s trigger to be recorded. Nevertheless, they were not informed that the real

microphone recording them was in the headset and not in their controller. Lots of them had the feeling of having a real microphone inside their hand, giving rise to the following remarks:

- It makes you feel like singing (x3);
- It's annoying to have it in your hand;
- I wanted to put the microphone on the table.

The first objective with the microphone metaphor was that people would understand how to use it without much explanations. With those remarks, and the facts that participants mostly brought their controllers up to their mouth to use the STT, we believe that goal achieved.

Finally, a problem highlighted for both STT was the absence of a correction tool in the IVE. This resulted in more tedious work for users during the proofreading and correction of the report as the STT had a noticeable error rate.

3.4. Discussion and conclusion

This experiment allowed us to identify the advantages and disadvantages of the three tools tested. Usability is considered as good for all. This is particularly due to the strong requirement of concentration from the tool. Concerning utility, note export, gathering and communication are the criteria that obtained the best scores, with a gathering stronger for the voice recognition tools. Ratings of notetaking and concise writing were more mixed, with still the notion that the DLK requires efforts. The satisfaction gave mixed results but highlights the generation of enthusiasm among participants. The stimulation showed the originality of voice recognition compared to the virtual keyboard. This is probably due to the similarities between the DLK and what we usually use to write text, namely a computer keyboard. Collective STT was also assessed as more creative than individual STT, itself more creative than the virtual keyboard. Beside the fact that the previous points explain the creative difference between the keyboard and the STTs as well, the collective nature of a STT is rather unusual. Finally, the items inventive, captivating and undemanding of stimulation are emphasized for the three tools. The identity brought out the collective aspect of the application, despite some isolated users who did not agree. This is very interesting to note: a hypothesis is that those participants were maybe appointed to write the report, and felt isolated because they had to take care of text input. Lastly, the spoken remarks allowed us to notice that DLK induced a lot of positive comments, but it also requires more cognitive resources, and more time to write. This meets a point identified in 2.1: speech is three times faster than text input.

Thus, in light of users' comments and user experience measures, we can observe that each tool has its own qualities and drawbacks. As a participant underlined, the voice recognition and the virtual keyboard can complement one another. Indeed, the main limitation of STT is its error rate. Yet, it cannot correct itself: it needs another tool, and the DLK is a good candidate. Its amusing side may engage users to use it to correct the text rather than a physical keyboard once outside the IVE, which would also require more time. Therefore, given these feedbacks, it would be particularly interesting to combine individual voice recognition and a virtual keyboard like a Drum-Like Keyboard.

With the recent implementation of hand-tracking in the Oculus Quest, one might wonder if there is a point to use hand-tracking as a text-input tool. It would certainly be interesting to compare it to the DLK, we do believe that the lack of haptic feedback could be a drawback, as well as the difference of precision with classic controllers.

In conclusion, we propose a text input system accessible and usable by any virtual reality application developer, using modern technologies of virtual reality and voice recognition. We think that those results will allow to create text input systems in IVE more suitable to users' needs.

12. Bibliography

Amodei D., Anubhai R., Battenberg E., Case C., Casper J., Catanzaro B., Chen J., Chrzanowski M., Coates A., Diamos G., Elsen E., Engel J., Fan L., Fougner C., Han T., Hannun A., Jun B., LeGresley P., Lin L., Narang S., Ng A., Ozair S., Prenger R., Raiman J., Satheesh S., Seetapun D., Sengupta S., Wang Y., Wang Z., Wang C., Xiao B., Yogatama D., Zhan J., Zhu Z., Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, *CoRR*, Vol. abs/1512.02595, 2015

Bowman D., Rhoton C. J., and Pinho M.S., Text input techniques for immersive virtual environments: An empirical comparison, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46, N°26, 2002, p2154-2158.

Brooke, J. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, vol. 189, v°194, 1996, p4-7.

Feeman S.-M., Wright L.-B. & Salmon J.-L., Exploration and evaluation of CAD modeling in virtual reality, *Computer aided Design & Applications*, Vol. 53, N°4, 2018, p892-904.

Jimenez J. G., A Prototype for Text Input in Virtual Reality with a Swype-like Process Using a Hand-tracking Device, Ph.D. Dissertation, UC San Diego, 2017.

Karat C.-M., Halverson C., Horn D., Karat J., Patterns of entry and correction in large vocabulary continuous speech recognition system, CHI '99, ACM, p. 568–575, Pittsburgh, Pennsylvania, USA, 1999

Kongsvik S., Text Input Techniques in Virtual Reality Environments - An empirical comparison, mémoire de maîtrise, University of Oslo, 2018.

Lallemand, C. & Koenig, V, How Could an Intranet be Like a Friend to Me? – Why Standardized UX Scales Don't Always Fit, *Proceedings of ECCE 2017*, Umeå, Sweden, 2017

Lallemand, C., Koenig, V., Gronier, G., & Martin, R., Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs, *Revue Européenne de Psychologie Appliquée*, 2015

Ruan S., Wobbrock J. O., Liou K., Ng A. Y., Landay J. A., Speech is 3x faster than typing for english and mandarin text entry on mobile devices, *CoRR*, Vol. abs/1608.07323, 2016

Yang X., Lin L., Cheng P.-Y., Yang X., Ren Y., & Huang Y.-M., Examining creativity through a virtual reality support system, *Educational Technology Research and Development*, vol. 66, n°5, 2018, p1231-1254.