BlowClick 2.0: A Trigger Based on Non-verbal Vocal Input

Daniel Zielasko* Neha Neha† Benjamin Weyers* Torsten W. Kuhlen*

Visual Computing Institute, RWTH Aachen University, Germany
JARA – High-Performance Computing

ABSTRACT

The use of non-verbal vocal input (NVVI) as a hand-free trigger approach has proven to be valuable in previous work [7]. Nevertheless, *BlowClick*'s original detection method is vulnerable to false positives and, thus, is limited in its potential use, e.g., together with acoustic feedback for the trigger. Therefore, we extend the existing approach by adding common machine learning methods. We found that a support vector machine (SVM) with Gaussian kernel performs best for detecting blowing with at least the same latency and more precision as before. Furthermore, we added acoustic feedback to the NVVI trigger, which increases the user's confidence. To evaluate the advanced trigger technique, we conducted a user study ($n = 33$). The results confirm that it is a reliable trigger; alone and as part of a hands-free point-and-click interface.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—[Voice I/O] I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—[Virtual reality] I.2.6 [Artificial Intelligence]: Learning—Parameter Learning

1 INTRODUCTION

One factor that can negatively influence the immersion of an immersive virtual environment (IVE) is the need to wear or use additional gear as it may make the user feel uncomfortable. This could be due to different reasons, as being heavy, cumbersome or just not supportive for the intended interaction. However, without any gear or controllers it is difficult to perform a selection, handle a menu or in general trigger events. Gesture and speech recognition offer s possible replacement for the gear. Gesture recognition gets more relevant, because of recent improvements in the field of computer vision and because, the definition of dedicated trigger gestures have been proven to basically work [1, 3]. However, especially when defining a trigger, approaches in both recognition fields suffer from high detection latency, since a gesture has to be finished or a word spoken to be detected correctly.

Sporka et al. [4] were able to show that users performed better with non-verbal vocal input (NVVI) than with speech input when controlling a Tetris game. Utilizing this, Zielasko et al [7] proposed a prototype named *BlowClick*. In this method blowing into a microphone is used as NVVI to trigger a click. The advantages of blowing are proposed to be that a user can perform and finish it very fast and the signal is easily distinguishable from common speech. To decide on a click, the sum of amplitudes in a short signal frame (about 30ms) is calculated and compared to a given threshold. The method has shown to be usable, to be very easy and fast to compute. However, the technique suffers from detecting other audio events than blowing as a trigger, e.g., coughing, sneezing or even speaking very loud.

In this work, we extend the idea of *BlowClick* by adding suitable machine learning methods to better distinguish blowing, or other suitable vocal inputs, by other audio signals, without losing its low calculation latency. We evaluate the advanced detection mechanisms regarding their reliability and usability. Furthermore, we enrich the feedback given to the user to strengthen their confidence about actions, which was another drawback of the realization of *BlowClick* [7]. Finally, the user study design used in the related work is reused and extended under the changed feature set and gives even more evidence for NVVI being a suitable trigger; alone and as part of a hands-free point-and-click interface.

*e-mail: {zielasko, weyers, kuhlen}@vr.rwth-aachen.de
†e-mail: neha.neha@rwth-aachen.de

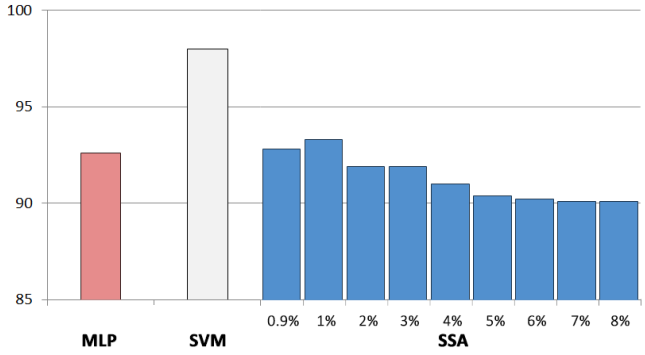

Figure 1: Accuracy of the test classification in percent, for MLP, SVM and SSA with different thresholds for the sum of amplitude.

2 METHOD

To detect the blowing signature, we choose SVMs, as literature seems to show a good performance in non-speech classification tasks in general [5, 6]. Nevertheless, as blowing was not explicitly investigated before in NVVI classification and neural nets are a common classifier in speech recognition, we cross check the classification with a *Multilayer Perceptron* (MLP), in the following. Both classifiers were trained on a *Mel Frequency Cepstral Coefficients* (MFCC) feature set completed by the sum of signal amplitude–used by *BlowClick*–, generated for unsupervised recorded audio files. The files contained a mix of speech and blowing. In the following, the trained classifiers were tested together with the original method used in *BlowClick* [7]. For the user study in the previous work, a threshold for the sum of signal amplitude (SSA) of 6.10% was used. We tested some additional parameters, as a useful threshold seemed to depend on factors like the used microphone. The results are depicted in Figure 1. Note that approximately 10% of the test frames were labeled as being a blow frame. This already leads to an accuracy of 90% in the results for a classifier, when it just classifies every frame being not a blow. We leave out a more detailed evaluation here and just determine that the SVM with a Gaussian Kernel reached the best results, having classified 89.9% of the blowing frames, and 98.9% of the non-blowing frames correctly. Thus, we will use this method in the following.

3 EVALUATION

To validate the advanced NVVI trigger using the SVM classification, we conducted a user study to measure the core performance

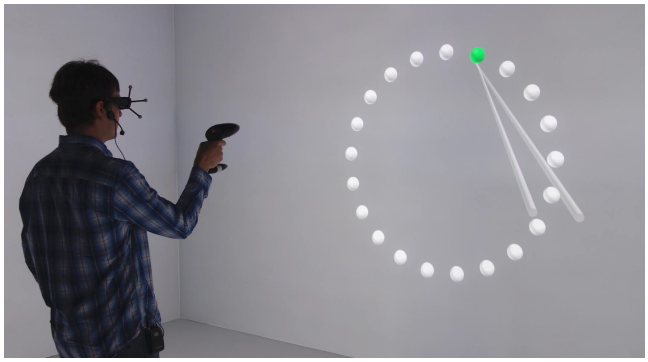

Figure 2: Fitt's Law task setup in an IVE displayed by a 5-sided CAVE.

Figure 3: The 3 tested device conditions from left to right, an ART Flystick2 for clicking and pointing (FF), NVVI detection for clicking and the Fystick2 for pointing (BF), NVVI detection for clicking and an ART hand target for pointing (BH).

parameters, speed and accuracy next to subjective measures. Additionally, one half of the participants got additional acoustic feedback for successful triggering. The study consisted of two task types which will we explained in the following.

The first was a Fitt's law selection task according to ISO 9241-400:2007 [2] (see Figure 2). Therefore, the NVVI trigger together with a pointing device builds a selection interface (see Figure 3, **BH**; clicking = **b**lowing, pointing = **h**and). Then, its selection performance was compared with a 6-DOF point-and-click device utilized as control condition **FF** (clicking = **f**lystick, pointing = **f**lystick). As pointing and triggering differ in both device combinations, we created a third one as a mixture of both (see Figure 3, **BF**), to allow differentiating causal connections in the results. To reasonably compare the different performance metrics, the throughput according to the ISO standard [2] was calculated for any device combination and is shown in Figure 4. The results reveal no statistically significant effects between all the device combinations, nor for the presence of acoustic feedback.

The second task aimed for independently comparing the triggers. Therefore, we used a reaction time task, where the user had to trigger as fast as possible when a sphere appeared in front of her. The results are depicted in Figure 5. Again, the results show no statistically significant effects between all the device combinations, nor for the presents of acoustic feedback.

4 Conclusion

In this work we refined a reliable NVVI metaphor for clicking. Therefore, we evaluated different classification methods and found an SVM with Gaussian Kernel to perform the best. Furthermore, this opened the possibility to add acoustic feedback to the NVVI trigger, without annoying the user too much. We conducted a user study to, inter alia, compare the NVVI trigger included in a hands-free selection interface (BH) with a standard 6-DOF point-and-click device (FF). Since our data failed to reject the null hypothesis, we think that both methods arguably perform similar. Supported by the subjective measures, which are not presented here, we want to advise the use of multi modal feedback in combination with NVVI

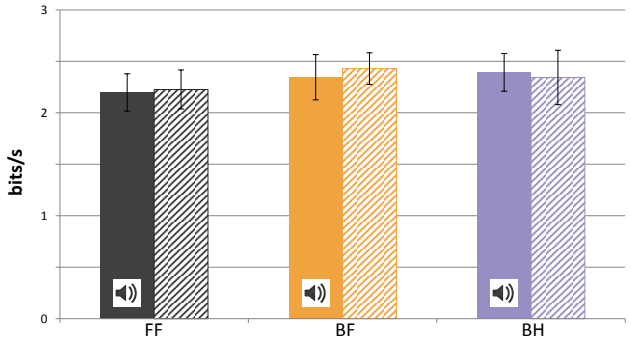

Figure 4: Device throughputs in $bits/s$. The striped bar represents the group that got no acoustic feedback. Error bars show the 95% confidence intervals.

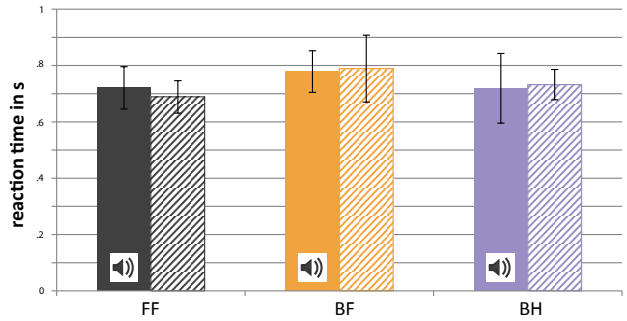

Figure 5: Average reaction time in s, per device. The striped bar represents the group that got no acoustic feedback.

triggers to increase confidence. Furthermore, the subjective measures indicate that the overall results are highly relevant, as users seem to prefer a hands-free point-and-click interface over a device, at least in the presented configuration.

References

[1] A. Choumane, G. Casiez, and L. Grisoni. Buttonless Clicking: Intuitive Select and Pick-Release Through Gesture Analysis. *In Proc. of IEEE Virtual Reality*, pages 67–70, 2010.

[2] ISO. *Ergonomics of Human-system Interaction: Principles and requirements for physical input devices (ISO 9241-400:2007, IDT)*. International Organisation for Standardisation, 2007.

[3] Y. Jang, S.-T. Noh, H. J. Chang, T.-K. Kim, and W. Woo. 3D Finger CAPE: Clicking Action and Position Estimation under Self-Occlusions in Egocentric Viewpoint. *IEEE Transactions on Visualization and Computer Graphics*, 21(4):501–510, 2015.

[4] A. J. Sporka, S. H. Kurniawan, M. Mahmud, and P. Slavík. Non-speech Input and Speech Recognition for Real-time Control of Computer Games. *In Proc. of ACM SIGACCESS Conference on Computers and Accessibility*, pages 213–220, 2006.

[5] B. Uzkent, B. D. Barkana, and H. Cevikalp. Non-Speech Environmental Sound Classification Using SVMs with a New Set of Features. *International Journal of Innovative Computing, Information and Control*, 8(5):3511–3524, 2012.

[6] J.-C. Wang, J.-F. Wang, K. W. He, and C.-S. Hsu. Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor. *In Proc. of IEEE International Joint Conference on Neural Network*, pages 1731–1735, 2006.

[7] D. Zielasko, S. Freitag, D. Rausch, Y. C. Law, B. Weyers, and T. W. Kuhlen. BlowClick: A Non-Verbal Vocal Input Metaphor for Clicking. *In Proc. of ACM Symposium on Spatial User Interaction*, pages 20–23, 2015.