A Reliable Non-verbal Vocal Input Metaphor for Clicking

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

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

ABSTRACT

While experiencing an immersive virtual environment a suitable trigger metaphor is often needed, e.g. for the interaction with objects out of physical reach or system control. The *BlowClick* approach [35] that is based on non-verbal vocal input has been proven to be a valuable trigger technique in previous work. However, its original detection method is vulnerable to false positives and, thus, is limited in its potential use. Therefore, we extended the existing approach by adding machine learning methods to reliably classify blowing events. We found a support vector machine with Gaussian kernel performing the best with at least the same latency and more precision than before. Furthermore, we added acoustic feedback to the NVVI trigger, which increases the user's confidence and whose absence was also stated as a limitation of the previous work. With this extended technique, we repeated the conducted Fitts' law experiment with 33 participants and could confirm that it is possible to use NVVI as a reliable trigger as part of a hands-free point-and-click interface. Furthermore, we tested reaction times to measure the trigger's performance without the influence of pointing and calculated device throughputs to ensure comparability.

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

The degree of immersion in a virtual environment (VE) is highest when the user just accesses the VE and interacts with it in the same way as in reality, without the necessity to wear and use gear, often referred to as "natural interaction". Wearing special gear or holding controllers can negatively influence the immersion by somehow making the user feel uncomfortable, i.e., being intrusive. This could be the case due to different reasons, such as being heavy, cumbersome, or just not supporting the intended interaction properly, e.g., by occupying the hands. Beyond simulating the reality, the user simultaneously wants to benefit from the possibility to extend interaction beyond what is possible in reality, for instance, the ability to select and interact with objects out of physical reach. One major requirement for the implementation of such interactions is a precise trigger. However, standard 6-DOF point-and-click devices suffer from the fact that mechanical triggers can cause small device movements when used, which then influences the pointing direction given by the same device. This effect referred to as *Heisenberg effect* [1], potentially reduces accuracy and leads to errors. However, without any gear or controllers it is difficult to perform a selection, handle a menu, or trigger an event in general. Gesture and speech recognition provide a possible replacement for this. Gesture recognition has become more interesting because of recent improvements in the field of computer vision and defining dedicated trigger gestures has been shown to basically work [7, 17, 26]. But especially when defining a trigger, approaches in both recognition fields suffer from high detection latency [13], since a gesture has to be finished or a word has to be spoken to be detected correctly.

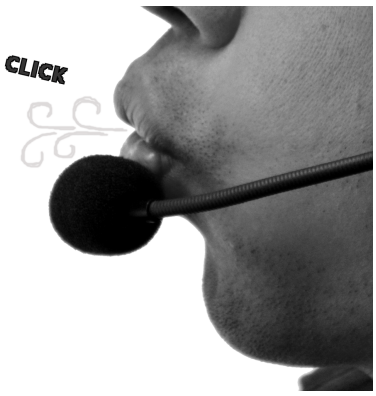

Figure 1: The idea of clicking induced by a non-verbal vocal input.

Sporka et al. [28] showed that users performed better with non-verbal vocal input (NVVI) than with speech input when controlling a Tetris game. Utilizing this, Zielasko et. al [35] proposed a prototype named *BlowClick*. In this method, blowing into a microphone is used as NVVI to trigger a click (see Figure 1). The advantages of blowing are proposed to be that a user can perform and finish it very fast and the signal is easy to distinguish from normal speech. To decide whether a click happened, the sum of amplitude in a short signal frame (about 30ms) is calculated and compared to a given threshold. The method was shown to be usable and is very easy and fast to compute. However, it suffers from detecting other audio events than blowing as a trigger, e.g., coughing, sneezing, or even speaking very loudly (see Figure 2 for an example).

In this work, we extend the idea of *BlowClick* by adding suitable machine learning methods to better distinguish blowing or other suitable vocal inputs, from other audio signals, without losing its low latency. We evaluate the improved detection mechanisms regarding their reliability and usability. Furthermore, we enrich the feedback given to the user to strengthen her confidence about actions, which was another drawback of the realization of *BlowClick* [35]. Finally, the user study design used in the related work is replicated 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.

The paper is structured as follows. In Section 2 we discuss related work in the filed of NVVI and its classification. In Section 3, two common machine learning methods are implemented to classify blow events more reliably and evaluated against the status quo. Additionally, the results are presented and discussed. In Section 4, an extended blow trigger is evaluated and the results are presented. We discuss the overall results in Section 5 and finally draw a conclusion in Section 6.

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

2 Related Work

First, we have to note that there is no *natural* interface or interaction to trigger something distant, which usually holds for pointing-based selection except, maybe, for throwing something at it. Thus, non-verbal vocal input (NVVI) is one possibility to rely on for a task like that and it is already very common in the field of accessible computing. A classic example is steering a wheelchair by *zipping* and *puffing* [10]. However, the focus of those techniques often lies more on feature-completeness than ease of use, i.e., they are often not easy to learn. This is due to the circumstance that a rich space of potential interactions usually compensates the effort to learn, for lasting motor-impaired people. Interfaces like that are the *Whistling User Interface* (U^3I) [24, 29], *The Vocal Joystick* [12], or the approach designed by Chanjaradwichai et al. [6], which in these cases open native desktop applications for disabled people. Another example is called the *Blowable User Interface* (BLUI) [23]. In this interface, the user can trigger a click by blowing into the direction of one out of nine regular grid cells within the application displayed in front of her. For this purpose, the recorded audio signals are classified and assigned to one of the cells. There are some disadvantages of this design that occur when considering it for regular use, starting with the resolution of the distinguishable click positions. In addition, this resolution is not easily scalable as the current design already requires a click confirmation due to uncertainty. Furthermore, the system is not portable, which would be desirable when using it in an IVE, as a reliable microphone position is essential for this method. Lastly, the classification algorithm has to be calibrated for each user.

Igarashi et al. [15] first mentioned the potential use of an NVVI interaction within an IVE. An application was a gestural interface together with a blowing metaphor and bottles to create a virtual music instrument [36]. Furthermore, this core concept was reduced to a general clicking metaphor with *BlowClick* [35]. In this metaphor the summed signal amplitude of a short time frame is used to make a binary decision of the system's trigger state, which is very fast to compute. At the end, the method's stake in the system's latency results nearly exclusively in waiting for the underlying signal frame to reach a meaningful size (they used approximately 30ms). As the signal's classification is that easy or general, *BlowClick* works user-independently, but is also vulnerable to false positives. A second drawback of the approach's actual implementation was the absence of additional feedback, as a typical *click* sound of a physical mouse that is often also modeled virtually. This potentially reduced the method's scores against a standard physical trigger input, which is discussed in more detail in the corresponding work [35]. That raises the idea to use the described idle time of *BlowClick* for a better classification, without increasing the total latency.

2.1 NVVI Classification

Today, there are many methods and approaches in the field of speech recognition but the classification of non-verbal sounds or features is not that densely covered. Furthermore, most work in this field tries to classify a sequence as non-verbal to early discard it for a speech recognition process.

All classifiers have in common that they do not operate on sound signals but on features of the signal, which have to be extracted first. Jarina and Olajec [18] experimented with various feature extraction methods in the context of automatic applause detection and found that *Mel Frequency Cepstral Coefficients* (MFCC) [21] performed best. Uzkent et al. [33] successfully used MFCCs and *Auto-Correlation Functions* (ACF).

As mentioned before, BLUI [23] uses NVVI classification, but the aim is to map air pressure signatures to the direction somebody is blowing in. For this purpose the authors decided to use a *k*-Nearest Neighbor (KNN) classification [9]. In a study performed by Cowling and Sitte [8] *Learning Vector Quantization* [20] outperformed a *Multilayer Perceptron* (MLP) [14] for NVVI classification. MLPs are supervised feed-forward neural networks with one input, one output, and one or more hidden layers with computation nodes or hidden units. Wang et al. [34] developed an approach to detect home environmental sounds such as coughing, laughing, etc. utilizing a hybrid approach consisting of *Support Vector Machines* (SVM) and KNN, and reported that the results outperformed Hidden Markov Model classifiers [25], which are widely used for speech recognition. SVMs [30] are supervised classifiers that separate data points along a boundary given by supporting points so that the empty space between the classes is as large as possible. In the work of Uzkent et al. [33], SVMs with a Gaussian kernel showed the best behavior in non-speech based classification, next to SVMs with a linear kernel, *Radial Basis Function* (RBF) [3] neural networks–a kind of MLPs– and KNN classifiers.

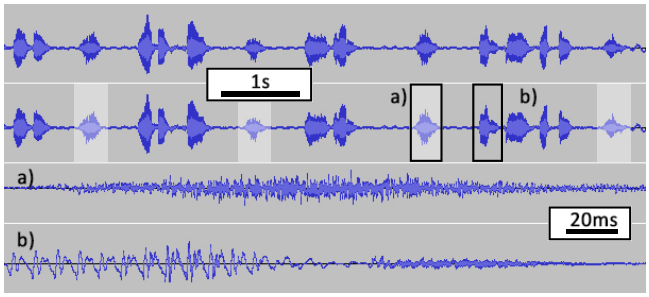

Figure 2: Cutout of an audio time domain plot (signal amplitude over time) from one of the recorded audio files. In the second row blowing events are highlighted. The third and fourth row show cutouts of the uppers, a) a blow event and b) a spoken word. It is easy to recognize blow events on this scale, as the speaker obviously took breaks before and after blowing.

In summary, MFCC in feature extraction and SVMs in classification have proven to work in diverse NVVI recognition settings, what makes them the best candidates to be considered for blow detection as well. However, there is no guarantee for success in our setting, not only because the signal features of blowing could differ from other NVVI, but even more as we are forced to consider much smaller signal frames than usually in related work. We will have a look on that in the following section.

3 Blow Classification

We further rely on blowing as trigger metaphor due to its potential signal uniqueness and inconspicuousness, i.e., the created air pressure is usually not audible. The latter can be important regarding the social acceptance of an interaction metaphor. Nevertheless, it is important to note that other non-speech sounds, or even user dependent and user preferred ones, are possible to use together with the following methods. The need for a more specific classifier to detect blowing is depicted in Figure 2. Here, the amplitudes of blowing and spoken words do not really differ as assumed in [35], but when having a deeper look on the micro scale other destructive features appear that can be general characteristics and potentially distinguishable by machine learning. Nevertheless, note that the signal structure immediately subsequent to the spoken word in Figure 2b has the same "character", as the blowing in 2a. This, furthermore, illustrates the challenge.

To detect the blowing signature, we choose SVMs, as literature seems to show a good performance in non-speech classification tasks in general. 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 an MLP in the following. In both cases, the MFCC set, completed

Table 1: SVM parameter set

training dataset size	10543
number of features	14
cross validation	3-fold
optimized value for hyper-parameter C	8
optimized value for hyper-parameter γ	$3.05176e^{-5}$
kernel type	RBF kernel
number of support vectors	1847
training accuracy	99.3%

Table 2: MLP parameter set

training dataset size	10543
batch size	100
batch size	100
number of input layers (features)	14
hidden layer activation function	ReLU
number of hidden units per hidden layers	100
loss function at output layer	sigmoid loss
momentum	0.9
weight decay	0.0005
solver mode	CPU
training accuracy	99.27%

by the sum of signal amplitude–used by *BlowClick*–were used as classification features.

3.1 Data Acquisition

To train and evaluate the classifiers, we collected 10 unsupervised recorded audio files in .mp3 file format from 10 voluntary and unpaid participants (see Figure 2 as an example). They were invited via e-mail to randomly pick one out of 12 short texts in either English (9 participants, with 8 native speakers) or German and read it out into a microphone. In advance, the sentences were randomly prepared to contain short instructions to blow into the microphone at a given point. On average, this was the case after 4.08 words, with an average text length of 54.13 words. Additionally, the participants were instructed not to create an artificially silent environment for the recording and were not instructed to blow or speak in any special way. The recordings have a length of 40s on average and were divided into training and testing corpora with a ratio of 3:2. After collecting the data it was divided into 20*ms* long, half overlapping time frames, labeled manually as belonging to a blow or not. In the previous work the underlying frames had a length of 30*ms* and we decided to shorten that time to have additional time for the classification, without increasing the latency.

3.2 Implementation & Training

To realize the SVM, the common library *LIBSVM* [5] was used. The parameters used are given in Table 1. For the MLP implementation the *Caffe* deep learning framework [19] was used. The parameters used are given in Table 2. Both classifiers were trained with a dataset of size 10.543 and reached a training accuracy of 99.3% (SVM) and 99.27% (MLP). The SVM was used together with an (Gaussian) RBF kernel as they perform well in general as long as speed is not an issue, which would then lead to a linear kernel.

3.3 Evaluation

For the evaluation, it was determined if the classification label for a time frame corresponds to the manually assigned one. When they differ it was counted as an error. This includes false positives, when not blowing was detected as blowing, and false negatives, when blowing was not labeled as blowing. The evaluation includes the method used in *BlowClick* [35] as well. For the user study in the previous work, a threshold for the a sum of signal amplitude (SSA) of 6.10% was used. We tested some additional parameters here, as the initial tests revealed that this parameter does not work well in general.

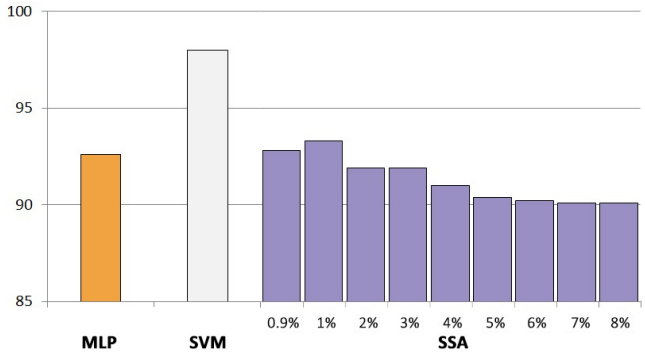

Figure 3: Accuracy of the test classification in percent, for MLP, SVM and SSA with different thresholds for the sum of amplitude.

The results for different MF CC feature combinations are not reported here, as there was no significant effect observable between them. The SVM achieved the best results with 98.0% correct classification, followed by the SSA (93.3%, for the best matching threshold 1%) and the MLP classification with 92.6% (see Figure 3). The detailed classification results are depicted in Figure 4 as confusion matrices. The calculation for one 20ms long sound frame took less than 1*ms* for the SVM and less than 2*ms* for the MLP, both including the computation of the MFCC features. The measurements were performed on an Intel® Xeon® E5540 with 2.53GHz, 64 bit and 12GB of RAM.

3.4 Discussion

One of the first things to note is that approximately 10% of the frames are classified as blowing. Thus, a classifier that labels everything as not blowing already reaches an accuracy of 90%. This happened with the 6%-SSA classifier. As shown in Figure 4, it detected no false positives but only 1.2% of the blow frames. We have two ideas why this classifier scored that badly here but seems to provide useful results in the previous work. First, most if not all training and testing files were recorded with very cheap microphones, which additionally were not worn directly in front of the mouth, resulting in smaller signal amplitudes. As this classifier only takes the signal amplitude into account, it is obviously very vulnerable regarding varying amplitudes. This hypothesis is supported by the fact that we found much better results for SSA with lower amplitude thresholds having its peak at 1% (see Figure 4). Here 97.2% of the blowing frames are correctly classified, but the number of false negatives is elevated as well. This proportionality seems to be unavoidable in an approach that is based on the amplitude alone. Second, a blow is always longer than 20*ms* and thus a blow event is divided into several frames. Then again, only one frame of a blow has to be classified to register the event. However, in this case, other problems can occur. It is likely that the latency is increased or one blow event translates into a tremor of clicks, but it also means that the number of detected blows is proportionally higher than the number of classified blow frames. Nevertheless, for best performance and a reliable trigger for various applications, it is necessary to detect a blow event from its first frame to its last, without any interruption.

The MLP classifier performed comparable to the 1%-SSA. With the training numbers in mind, we were surprised that MLP performed so much worse then SVM. For cross-validation, we removed a small part (about 30%) of each test data file and put it in

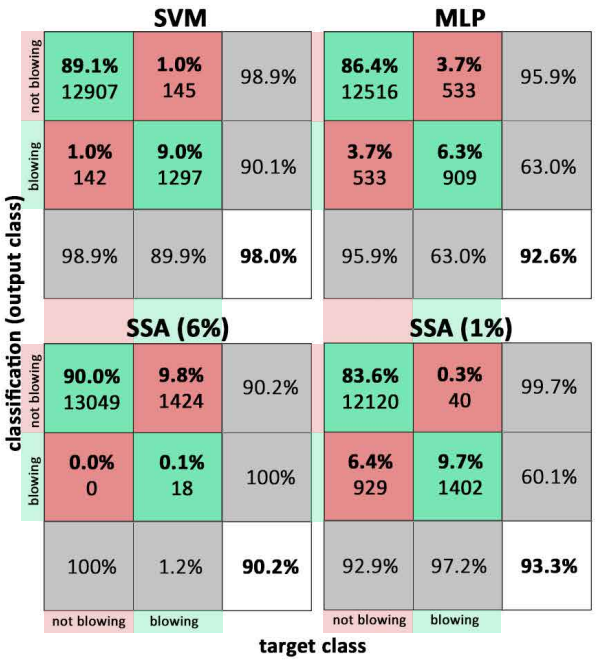

Figure 4: Confusion matrix for the test classification of SVM, MLP and SSA for 6% and 1% maximum amplitude. The matrices show, in the upper row, the number of true negatives, number of false positives and false positive rate (fall-out); in the middle row, number of false negatives, number of true positives and true positive rate (sensitivity); and in the lower row, negative predictive value, positive predictive value and finally, accuracy.

the training-pool. This did not change the results for the MLP and the accuracy of the SVM even dropped by 1%. In summary, MLPs seem to work just worse then SVMs for NVVI recognition, which is supported by related work (see section 2.1). This is consistent with the general observation that SVMs work well with small datasets and MLPs simultaneously tend to over-fit when trained with small datasets. A surprising observation is that the error rate does not change with training-data of the individual speaker. On the one hand, this suggests that the SVM can be used out of the box, and on the other hand it takes the possibility to improve the method's accuracy with additional speaker-specific training when there is time. Both MLP and SVM took less time in total than the envisaged 30*ms* for waiting on the frame buffer to get filled and the execution of the classification process, which makes them usable in an interactive application and additionally opens up the opportunity to even increase the frame sizes depending on the total system latency. In summary, SVM showed an overall good performance compared to the other classifiers and, thus, is chosen for the following work.

4 User Study

To validate the NVVI trigger using the SVM classification, we conducted a user study to measure the core performance parameters usability, speed and accuracy. Therefore, two different task designs were used. For the first, the NVVI trigger was combined with a pointing device to build a selection interface. Then, the selection performance can be compared with a 6-DOF point-and-click device utilized as ground truth, in a Fitts' Law task. Note that the standard device could not be part of an hand-free interaction interface. Second, the NVVI trigger is also compared as a stand-alone trigger in the study.

In the following subsections, first the study's implementation details of the NVVI metaphor for clicking are described in Section 4.1, followed by the used apparatus in Section 4.2. In the main part, the experiment's general procedure (see Section 4.3) with the detailed task descriptions (see Section 4.3.1 & 4.3.2) is given. Then, the subject population is described in Section 4.4. Finally, we make our hypotheses in Section 4.5 and report the results in Section 4.6.

4.1 Method Implementation

In the following, we used an SVM for the classification of a blow and rejected the other methods (see Section 3). The only thing that changes in the process of the classification described in Section 3 is that it works on uncompressed live audio frames instead of recorded audio files. Again, every frame has a length of 20*ms* and is half-overlapping with the last frame. The classification of frames over time then is treated as a binary signal, which is the output of our audio processing. This signal is wired to a selection input slot of a virtual device in a widget framework [11]. The framework is responsible for performing a click on every registered object with focus, when there is a rising flank on the input slot, i.e., when the current frame was labeled as blow and the one before not. This is a fine detail in comparison to *BlowClick*, where a click was performed with the falling flank. Nevertheless, as a result of this detail, the target object in the previous study had to have focus the whole time between raising and falling flank. Nevertheless, in the current implementation, a new click is only fired with a raising flank, i.e., the signal had to fall back to its resting state before. Together with every click a *click*-sound is played directly by the audio process.

4.2 Apparatus

The experimental part of the study took place in a 5-sided CAVE. The participants stood in the middle of it–marked with a red dot on the floor–facing the CAVEs back-wall in a distance of approximately 2.6*m*. The dimension of the walls were 3.3*m* in height and 5.3*m* in width. The participants wore tracked, active-shutter stereo glasses to provide head-tracking and stereo vision. To capture acoustics, a Sennheiser EW G2 together with a Sennheiser ME3 wireless microphone system was worn during the entire experimental part of the study, even when not used. Whenever the participants were asked to use their dominant hand as a pointing device, it was tracked by a lightweight tracking target by ART, which was attached to the hand by an elastic band (see Figure 6). Finally, an ART Flystick2 was used as the standard 6-DOF point-and-click device.

4.3 Procedure

At the beginning, all participants were asked to fill out a demographic questionnaire and carefully read a printed study description. After this, every participant performed two interaction tasks in the CAVE using different device conditions in a between-subjects experimental design regarding the acoustic feedback, i.e., half of the participants were exposed to acoustic feedback for clicking (**ac**) and half were not. The first task was a repetition of the Fitts' law task conducted in the *BlowClick* study [35] and is described in detail in Section 4.3.1. As an addition to the previous study, a second task, for measuring trigger reaction times, was added after each device condition to measure differences between a classical button trigger and the NVVI trigger without any context and, thus, other components like pointing. The reaction task is described in more detail in Section 4.3.2. When the participants got acoustic feedback for clicking and selection it was added to all device conditions. The experimental part of the study took approximately 20 minutes. We did not want to exceed this time as it has been shown in tests and previous work [35] that the effects of getting exhausted by holding the flystick for pointing gets stronger and could have an unwanted effect on the overall performance of the participants. Of course, this

Figure 5: Fitts' Law task setup with task difficulty T3.

is an argument for using a light-weight alternative pointing device in tasks like the described one, although this is not the effect that should be measured here. Following the experimental part, every participant was finally asked to fill out a *System Usability Questionnaire* (SUS) [2] for each of the 3 point-and-click device combinations (see Figure 6) supplemented with a questionnaire regarding relevant subjective measures such as perceived performance, exhaustion, etc.

4.3.1 Fitts' Law Task

In this task we used a $3 \times 4 \times 21$ within-subject design, including 3 device conditions utilized to solve 4 increasingly difficult Fitts' Law tasks with 21 trails each. The task was designed according to ISO 9241-400:2007 [16, 27] and the setting is shown in Figure 5. A Fitts' Law task does not quantify a trigger alone, but a selection and thus requires a pointer. Therefore, the blow detection is combined with the direction the user's hand is pointing in, for the first device condition **BH** (trigger = blow, pointer = hand, see Figure 6). The second device combination is a 6-DOF point-and-click device, as it is interesting how the first competes against a quasi standard. Here, the surrogate for this device is a Flystick **FF** (trigger & pointer = Flystick). In this combination it is not possible to assign any possibly observed effects to the method that is used to trigger, as the pointing is different, too. Thus, a third bridging condition **BF** is introduced. Here, the Flystick is used as a pointing device only and the blow detection functionally replaces its trigger. The conditions were provided in counter-balanced order, following a latin square design.

The goal in a Fitts' Law tasks is to select, i.e., point at a target and trigger a click, alternating targets on opposite sides as fast as possible. In the given task, this had to happen 21 times for each task difficulty (**T1**-**T4**). The targets are spheres and were positioned in a circle of radius $0.75m$. The participants were asked to rank accuracy, for instance, not missing a sphere by accident, over speed. They were shown a pointing ray, starting at the pointing device's position and given no advanced selection strategy, but simply ray-casting. The spheres were rendered exactly on the projection surface to exclude any potential effects of distance estimation [4] and potential effects caused by varying distances to the selection target [32]. The current target sphere was colored green, while all others were white. When focused, the target sphere changed its color from green to white. Finally, the sphere turned blue during the *button down* phase of a click. In the acoustic feedback group, a successful selection of a sphere was accompanied by a *blub* sound instead of the usual *click* sound. The four task difficulties were designed by changing the size, i.e., the target width for selection of the spheres. The first task (**T1**) drew spheres of radius $0.1m$, followed by **T2** with $0.075m$, **T3** with $0.05m$ and finally **T4** with $0.025m$. The difficulty of the last task was designed–following pilot tests–to provoke errors from nearly all participants. Every task and subtask had to be started with the first selection, which allowed to rest between

Figure 6: 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, NVVI detection for clicking and an ART hand target for pointing.

them. In the beginning of any device combination, the participants had the opportunity to get familiar with the device and the selection task within an easy and unrecorded training task.

In the Fitts' Law task we measured the total number of clicks needed to solve a task, the time between any selection and the precise position the target sphere was selected at, i.e., the position the *button down flank* of the click was performed/registered at.

4.3.2 Reaction Time Task

In this task, the participants did not have to point, but only trigger as quickly as possible as soon as a big sphere appeared in front of them. The sphere disappeared after a click was performed. The time range in between the trials was random and varied around 1s. We measured the reaction times between sphere appearance and trigger event. Due to technical issues during the session of some of the participants it was only possible to record the first 17 trials for them. To compensate this, we consulted only the first 17 trials of any participant and session for the analysis.

4.4 Participants

33 volunteers (5 female and 28 male, ages M = 26.97 years, SD = 3.56 years) finished the study. Additionally, two participants canceled the experiment before its end and were not considered in the analysis because of incomplete data. The first felt dizzy after some time and the second one was physically not able to trigger any click by blowing, i.e. the participant could not create enough air pressure to record a sufficient acoustic signal. The participants were compensated with free candy and drinks. All reported normal or corrected-to-normal vision. Asked for their prior experience with 3D user interfaces, 7 answered that they had contact on a regular basis, 13 reported that they used a 3D user interface before and 13 answered to had no experience at all or do not know what a 3D user interface is. Additionally, 5 out of 33 participants reported to have never used any stereo display system–including 3D cinema–or head tracking.

4.5 Hypotheses

We formulate the following hypotheses regarding the results. First, we think that we confirm the previous results, e.g., regarding general usability.

- **H1** It is possible to reasonably solve the given tasks with the given NVVI metaphor for clicking.
- **H2** It is more exhausting to use blowing compared to a mechanical trigger.
- **H3** It is more exhausting to use the flystick as a pointer compared to the user's hand.

In the previous study [35] the lack of acoustic feedback seemed to had a negative influence on the task performance, especially when using the NVVI trigger. For instance, the participants sometimes accidentally did not trigger a click due to insufficient precise pointing, but assumed the error results from an unsuccessful blow detection, which led them to concentrate on that instead of trying to

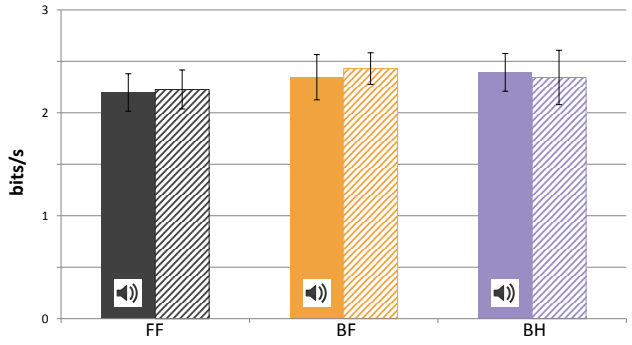

Figure 7: Device throughputs in $bits/s$. The striped bar represents the group that got no acoustic feedback. Error bars show the 95% confidence intervals.

coordinate simultaneous point and trigger. This should be compensated by additional feedback. Therefore, we expect a better overall performance that in case of the BH device catches up with the flystick when acoustic feedback is provided.

- **H4** The acoustic feedback helps increasing the users confidence in the blow conditions.
- **H5** The group with acoustic feedback performs better regarding speed and accuracy in the blowing conditions than the group without additional feedback in the same condition.
- **H6** The participants are annoyed by the additional acoustic feedback (because of false positives).
- **H7** The blow conditions perform as good as the pure flystick condition, at least when supported by acoustic feedback.

Additionally, we expect a better performance of the blow conditions in comparison to the results of the previous study [35], due to the improvements made. However, any direct comparison is difficult because of possible differences in the overall setup, e.g., distribution of user experience.

4.6 Results

We report all results using a significance level of .05 and non-significant trends at a level of .1. The objective measures were analyzed with a two-way mixed-design ANOVA with a between-subjects factor of acoustic feedback (present, not present) and different repeated within-subjects factors reported in detail in the following. Whenever Mauchly's test indicated that the assumption of sphericity had been violated, the degrees of freedom were corrected using Greenhouse-Geisser estimates. Additionally, we will report the $F-$values and $p-$values in the case of the task difficulties T1-T4 as ordered tuples $F(x,y) = (F(\mathrm{T1}), F(\mathrm{T2}), ...)$, $p = (p(\mathrm{T1}), p(\mathrm{T2}), ...)$ to support readability.

Starting with the Fitts' law task (see Section 4.3.1), we calculated the device's throughput in $bits/s$ (see Figure 7), as suggested by [27]. We used the Shannon formulation of Fitts's law. In the following, we only discuss the *CtA* order of calculation [22], i.e. throughputs are first computed per difficulty level and then averaged. However, we also report the *AtC* order together with all other results for comparability in Table 3. Throughput combines speed and accuracy and, thus, is a good measure to compare overall device performance. The statistical analysis revealed no main effects of acoustic feedback on throughput, $F(1,31) = .033, p = .858$, nor a main effect of the used device combination, $F(2,62) = 3.048, p = .55$, and no interaction between acoustic feedback and device combination, $F(2,62) = .373, p = .690$.

As throughput is not a complete substitute for the measured time (see Figure 8) and error rate, i.e., ratio of false clicks to total number of performed clicks (see Figure 9), they were additionally inspected in the following. There were no main effects of acoustic feedback on the mean time needed to perform a selection in any task condition, $F(1,31) = (.923, 1.525, .024, .443)$, $p = (.344, .226, .878, .511)$. Furthermore, there were no main effects of the used device combination on the time for the two easier tasks T1 and T2, $F(2,62) = (.998, .892)$, $p = (.374, .415)$. However, there was a statistically significant effect for the two more difficult tasks T3 and T4, $F(2,62) = (3.162, 8.437)$, $p = (\mathbf{.049, .001})$. A subsequent post-hoc Bonferroni test revealed no statistically significant effects between the device pairs, (FF, BF) $p = .110$, (FF, BH) $p = .159$, and (BF, BH) $p = 1.0$, regarding time. The same test showed that participants performed significantly slower in T4 with the FF condition than with BH ($p = \mathbf{.009}$) and the BF ($p = \mathbf{.003}$). There was again no significant effect between the two blow conditions BF and BH, $p = .861$. Finally, we did not find an interaction effect between acoustic feedback and device combination on the time per selection for any task, $F(1,31) = (.923, 1.525, .024, .443)$, $p = (.344, .226, .878, .511)$.

Continuing with the error, there was no significant main effect of acoustic feedback on error rate found for any task condition $F(1,31) = (.919, .226, .037, .391)$, $p = (.345, .638, .849, .536)$. Regarding the device combination there was only a significant main effect for the most difficult task T4, $F(2,62) = 13.533, p < \mathbf{.001}$, while not for T1-T3, $F(2,62) = (.771, 1.336, 2.489)$, $p = (.467, .263, .091)$. A post-hoc Bonferroni test showed that there happened significantly more errors in T4 when exclusively using the flystick, than when using BF, $p = \mathbf{0.24}$, or BH, $p < \mathbf{.001}$. There was an non-significant trend between BF and BH, indicating that BH was less error-prone than BF, $p = .087$. Finally, there was no interaction effect between acoustic feedback and device combination on the numberof errors made for any task, $F(1,31) = (.919, .226, .037, .391)$, $p = (.345, .638, .849, .536)$.

The results for the reaction time task (see Section 4.3.2) are depicted in Figure 10 and also included in Table 3. The analysis revealed no significant effect on the reaction time by the acoustic feedback, $F(1,31) = .005, p = .945$, nor by the device combination, $F(2,62) = 1.717, p = .194$. Furthermore, there was no significant interaction between the two factors, $F(2,62) = .163, p = .807$.

The subjective measures were analyzed using a one-way ANOVA with a between-subjects factor of acoustic feedback (present, not present), using Welch's ANOVA instead where Levene's test indicated that the assumption of homogeneity of variances was violated. As Post-hoc test, we used Tukey's honest significant differences (HSD) or the Games-Howell test, where the assumption of homogeneity of variances was violated.

Figure 11 shows the results of a 5-point Likert scale subjective questionnaire the participants answered after the experimental phase of the study. The statistical analysis revealed no significant effects of acoustic feedback in any of the questions but the following two. Participants that received acoustical feedback less often had the feeling that they repeatedly had to blow to trigger a click (Q10), $F(1,30.163) = 7.402, p = \mathbf{.026}$. Furthermore, they felt more confident while blowing (Q11), $F(1,31) = .317, p = \mathbf{.003}$.

Additionally, we asked the participant to directly compare the two trigger methods and the two pointing methods in NASA-TLX inspired questionnaires (see Figure 12 & 13). The statistical analysis revealed no significant effects of acoustic feedback in all the subjective comparisons.

Finally the participants filled out a SUS questionnaire for the two relevant device combinations, FF and BH. The flystick condition received a score of $M = 74.3$, $SD = 14.7$ with acoustic feedback and $M = 77.3$, $SD = 19.6$, without (see Figure 14). The hand-free interaction method utilizing a blow trigger scored $M = 76.7$, $SD =$

Table 3: Mean and (SD) over the participants, of the objective measures and system usability score (SUS), per device combinations (FF, BF, BH) and divided by the factor acoustic feedback (ac).

		TP		time				error rate				reaction time	SUS
		CtA	AtC	T1	T2	T3	T4	T1	T2	T3	T4		
FF	**ac**	2.20 (.40)	2.11 (.44)	1.22 (.20)	1.37 (.29)	1.70 (.37)	3.04 (.79)	5.84 (8.54)	8.76 (10.50)	15.27 (11.98)	42.20 (12.99)	.72 (.17)	74.34 (14.69)
		2.23 (.36)	2.13 (.34)	1.26 (.16)	1.39 (.24)	1.75 (.27)	3.04 (.60)	6.64 (7.49)	8.31 (10.45)	17.10 (8.41)	39.03 (14.31)	.69 (.11)	77.31 (19.62)
BF	**ac**	2.35 (.49)	2.26 (.49)	1.28 (.19)	1.41 (.26)	1.63 (.24)	2.83 (.88)	5.89 (7.21)	14.12 (14.00)	11.97 (13.13)	35.13 (14.03)	.78 (.16)	.
		2.43 (.29)	2.43 (.35)	1.34 (.26)	1.45 (.26)	1.57 (.24)	2.59 (.56)	10.81 (12.61)	8.28 (7.58)	12.37 (10.55)	32.24 (11.88)	.79 (.23)	.
BH	**ac**	2.39 (.41)	2.31 (.44)	1.26 (.19)	1.28 (.11)	1.56 (.30)	2.69 (.87)	6.26 (7.59)	6.24 (6.48)	11.98 (9.29)	29.62 (11.51)	.72 (.28)	76.71 (18.65)
		2.34 (.50)	2.32 (.49)	1.34 (.45)	1.43 (.35)	1.61 (.45)	2.49 (.47)	6.71 (8.05)	9.17 (6.87)	11.37 (9.81)	28.67 (11.15)	.73 (.10)	71.61 (15.83)

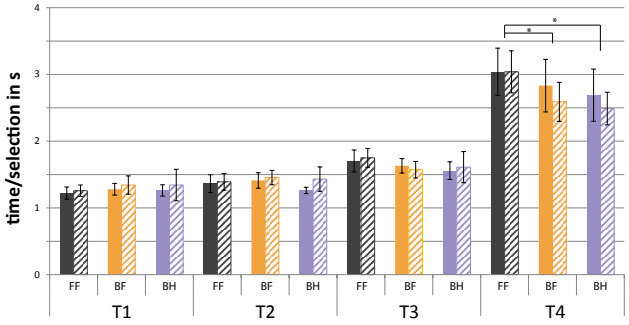

Figure 8: Average time needed to perform one selection, per task T1-T4 and, per device. The striped bar represents the group that got no acoustic feedback. Error bars show the 95% confidence intervals. A * denotes statistical significance, $p < .05$.

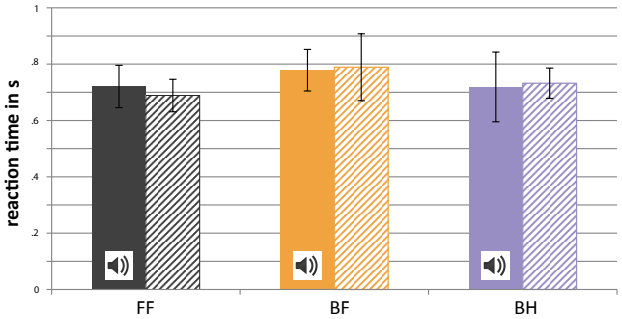

Figure 10: Average reaction time in seconds, per device. The striped bar represents the group that got no acoustic feedback.

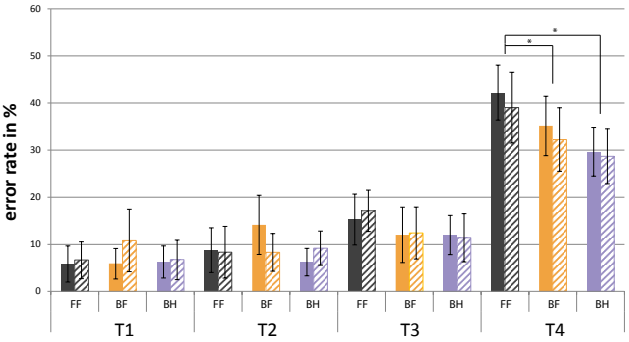

Figure 9: Average error rate with regard to number of clicks performed in total in percent, per task T1-T4 and, per device. The striped bar represents the group that got no acoustic feedback. Error bars show the 95% confidence intervals. A * denotes statistical significance, $p < .05$.

18.6 with acoustic feedback and $M = 71.6$, $SD = 15.8$, without. The statistical analysis revealed no significant effects of acoustic feedback on the results.

4.7 Secondary Results

One notable difference to the *BlowClick* study setup as described in [35] is the average user experience. While the pool of participants here was more homogeneous (see Section 4.4), the previous one consisted mainly of experts in the field of 3D interaction. Thus, we were additionally interested in potential effects, that may occur in special groups of users, first time users ($n = 13$) and expert users ($n = 7$). We performed a one-way repeated measures ANOVA on both of the groups with an within-subjects factor of device combination. The analysis revealed no interesting effects within the group of experts, but found significant effects within the group of first time users in the time per selection of T3 and T4, $F(2,24) = (3.939, 4.649), p = (.033, .020)$, and in the number of errors made of T3, $F(2,24) = 9.636, p = \mathbf{.001}$,. and T4, $F(2,24) = 14.150, p < \mathbf{.001}$. A post-hoc Bonferroni test showed that first-time users made significant more errors in the both more difficult tasks with the FF device, than with the BF, ($p = \mathbf{.010}, p = \mathbf{.043}$), and with the BH condition, ($p = \mathbf{.007}, p = \mathbf{.001}$). In the time needed per selection there was no significant effect, but a non-significant trend suggesting that the BH device condition was faster than the BF in T3, $p = 0.63$. However, no significant effect between the overall device throughputs could be shown, $F(2,24) = 3.310, p = .054$.

5 Discussion

First of all, we are surprised that the presence of additional acoustic feedback had no significant effect on most of the subjective or objective measures for any of the device conditions (**H5**). This was and is especially unexpected as participants reported and were observed in the previous study [35] to have had often problems to recognize when a trigger induced by blowing was successful. This showed up in two effects. First, they went for the next target after not having selected the current one correctly and, thus, had to come back. Second, the participants started to blow harder, because of a mis-selection, but the reason for that was not the trigger, but the pointing. Both should actually happen less often when provided with suitable feedback, and at least, the subjective measures show a significant difference in the acoustic feedback factor in the two questions dealing with this (see Q10 and Q11). The latter confirms **H4** which states that users feel more confident with the additional feedback, however it was not measured to have an effect on the overall task performance (**H5**), even though the participants were not more annoyed from the overall feedback as expected (**H6**). We can also exclude the possibility that the tasks were not difficult enough to benefit from the acoustic feedback, especially the feedback given for a successful selection, which differed from the standard trigger feedback, as the error rate for T4, for example, lay between 30% and 40% for all devices. Of course, it seems to be possible that most of the errors were induced by the pointing

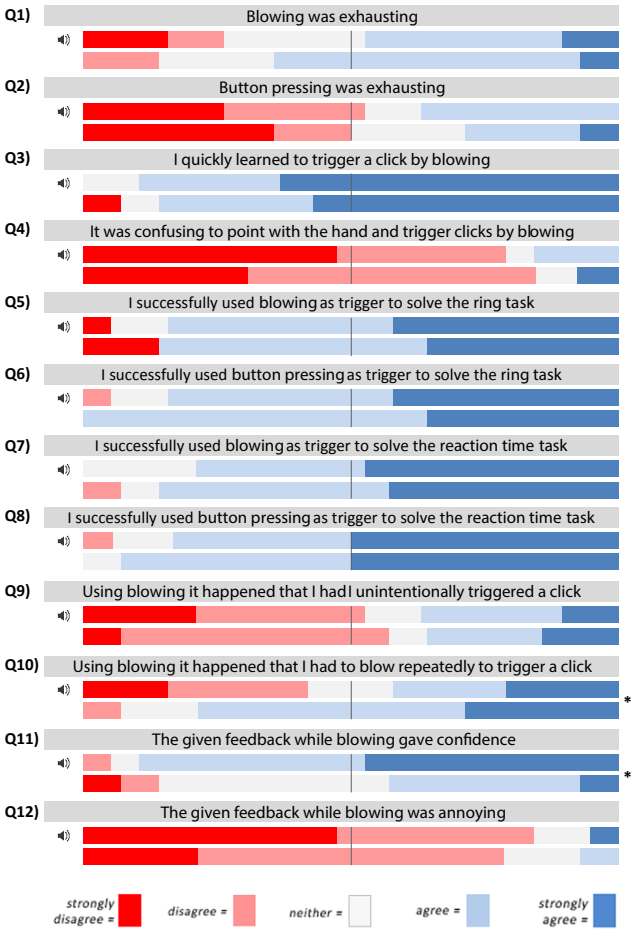

Figure 11: Answers to the subjective questionnaire, divided by the factor acoustic feedback. A * denotes statistical significance, $p < .05$.

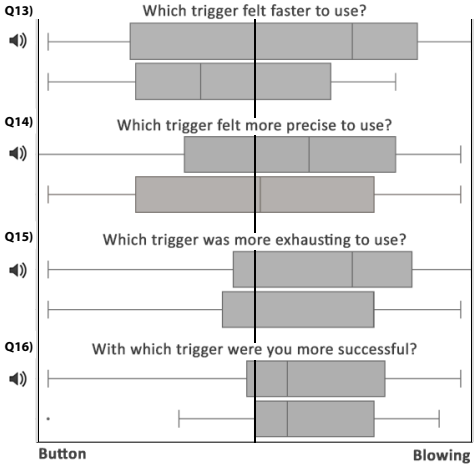

Figure 12: Subjective comparison between trigger, divided by the factor acoustic feedback.

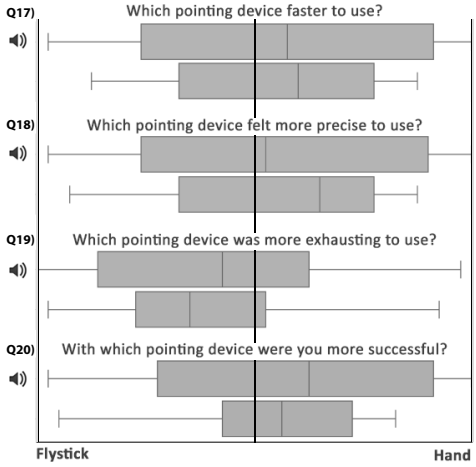

Figure 13: Subjective comparison between pointer, divided by the factor acoustic feedback.

metaphors or pointing process and not by the trigger. Nevertheless, there is no other evidence for that and we may need further measurements to divide triggering from pointing effects. In summary, people felt more confident with the additional feedback, but the effect is not strong enough to be measurable in the objective measures or SUS (see Figure 14). This led us to the recommendation to use additional acoustic feedback with an NVVI trigger to keep the user aware of the current system state (as the classification is not 100% precise) even when it does not directly increase the user's performance.

Independently from the feedback, the overall subjective and objective results confirm that blowing into a microphone is a suitable metaphor for clicking (**H1**). Furthermore, it does not perform worse than a standard interaction device, the flystick (**H7**), and both reached good system usability scores (see Figure 14). Moreover, the blowing conditions BF and BH performed significantly faster and more precise than the pure flystick condition within the most difficult task. However, we think that the reason for that is not the trigger alone, but the hand-eye coordination, or more precisely the trigger-pointing coordination. It is usually even challenging to keep a sphere focused for a longer time without trembling in this task. Its actual moderate task difficulty ($ID = 6$) or the corrected one ID_e ranging from 5.4 to 6.3, does not indicate this, but its level does not seem to translate one-to-one to a midair interaction context, which is a point we will discuss below. This is notable in the disproportionate increase of time and error for all conditions in T4 (see Figure 8 & 9). Having noted that, one can imagine that it is even harder to trigger in the exact right moment, when the sphere is hit. In our opinion, this is easier, first, with a trigger that is detached from the pointing device, as the use of the mechanical trigger causes small movements of the device, an effect called *Heisenberg effect* [1]. Second, blowing may allow for more sensitive hitting the right moment for triggering. Interestingly, this is the effect which seems to occur earlier, i.e., already within T3 and for T4 again, if first-time users of 3D user interfaces (see Section 4.7) are considered. This might point out that this problem is even bigger for unexperienced users but can be compensated with some training. The subjective trigger comparisons (see Figure 12) also tend towards the NVVI trigger, except the one asking for the exhaustion, as expected (**H2**). Also as expected, the participants felt more exhausted using the flystick as pointing device, rather than their hand (**H3**). Furthermore, we can also confirm the results of the previous study [35], showing that users seem to prefer their hand as a pointer over a flystick (see Figure 13).

We made another observation that we want to share. The task difficulty (ID) is one parameter of the throughput calculation, as

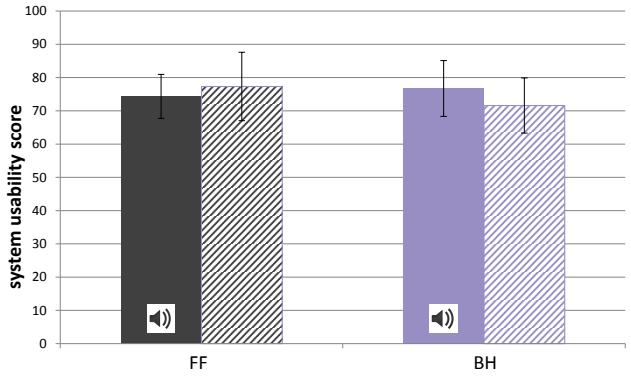

Figure 14: System usability score (SUS), per device combination. The striped bar represents the group that got no acoustic feedback.

it results from the average of the throughputs measured for each difficulty. To calculate a robust score, it is recommended to consider tasks with *ID*s approximately filling the interval between 2 and 8 [16]. As mentioned before, it seems to be hard for 3D selection tasks, as the *ID* dies not translate in the same way to subjective difficulty levels as in a 2D task. The tasks were designed to aim for those interval borders and while this worked for the easiest task T1, with an $ID = 4$ and a corrected ID_e between 2.7 and 3.1, it was not possible for the most difficult task, as mentioned above. However, our throughput values are comparable with the ones measured for a device similar to ours, the *Pen Ray*, by Teather et al. [31], even when in another display setting.

Compared to the results from the previous study [35], the main difference between both is the advanced recognition method for the blow trigger. Nevertheless, there might be others that are not obvious and two were mentioned before including the experience level of the participants and a slightly different definition of a click, regarding raising and falling signal flanks (see Section 4.1). However, the first notable aspect is that most of the measurements for time and error are in the same intervals, which arguably shows that the results are at least roughly comparable. Moreover, the task T4 is standing out and excluded from this observation. Participants performed worse regarding time and error within this task for all device conditions and only in this task. One explanation for this can be that the less experienced participants had much more problems with this task, which supports the observations made above. Furthermore, the FF condition performed significant faster and with fewer errors than the BH condition especially in the two easier tasks, T1 and T2. This differences disappeared completely. Additionally, the subjective comparison of the more successful trigger (Q20) switched from a light tendency for the button to a light one for the NVVI trigger. Together with the overall results, this shows that depending on the application an NVVI trigger should be considered as a possible solution.

Lastly, we want to point out that the NVVI trigger as presented here requires the user to wear a microphone, which is a limitation to what we motivated to achieve. Although a wireless microphone is usually a lightweight device, the user has to equip it and might feel uncomfortable wearing it. However, this introduces no additional effort in application interfaces that already use speech input, for example. Furthermore, especially with the advanced classification presented here, it should no longer be necessary to blow directly into a microphone, especially when considering different NVVI types. This opens up the possibility to pick up the sound by an external sensor, as long as the background noise is not too dominant or the sensor is a dynamic directional microphone.

6 Conclusion

In this work, we presented a reliable NVVI metaphor for clicking. We evaluated different classification methods and found an SVM with Gaussian Kernel to perform best. Furthermore, this opened up the possibility to add acoustic feedback to the NVVI trigger, without annoying the user to much, because of false positives. We conducted a user study to, among other goals, compare the NVVI trigger included in a hand-free selection interface with a standard 6-DOF point-and-click device and could show that it is able to perform similarly. Additionally, our results led us to recommend the use of triggers that are not mechanical and detached from the pointing device, at least for difficult 3D midair selection tasks. Moreover, we want to advise the use of multi-modal feedback in combination with NVVI triggers to increase confidence. Finally, the subjective measures indicate that the results are highly relevant, as users seem to prefer a hand-free point-and-click interface over a device, at least in the presented configuration. In summary, the NVVI metaphor for clicking showed enough potential to be considered in various IVE application contexts and beyond that, just to exemplary name accessible computing or mobile devices.

References

[1] D. Bowman, C. Wingrave, J. Campbell, and V. Ly. Using Pinch Gloves®for Both Natural and Abstract Interaction Techniques in Virtual Environments. *In Proc. of HCI International*, pages 629–633, 2001.

[2] J. Brooke. SUS-A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.

[3] D. S. Broomhead and D. Lowe. Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. *Complex Systems*, 2:321–355, 1988.

[4] G. Bruder, F. A. Sanz, A.-H. Olivier, and A. Lécuyer. Distance Estimation in Large Immersive Projection Systems, Revisited. *In Proc. of IEEE Virtual Reality*, pages 27–32, 2015.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[6] S. Chanjaradwichai, P. Punyabukkana, and A. Suchato. Design and Evaluation of a Non-Verbal Voice-Controlled Cursor for Point-And-Click Tasks. *In Proc. of the 4th International Convention on Rehabilitation Engineering & Assistive Technology*, pages 48:1–48:4, 2010.

[7] 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.

[8] M. Cowling and R. Sitte. Analysis of Speech Recognition Techniques for Use in a Non-Speech Sound Recognition System. *International Symposium on Digital Processing and Communication Systems*, pages 16–20, 2002.

[9] P. Cunningham and S. J. Delany. K-Nearest Neighbour Classifiers. *Multiple Classifier Systems*, pages 1–17, 2007.

[10] L. Fehr, W. E. Langbein, and S. B. Skaar. Adequacy of Power Wheelchair Control Interfaces for Persons with Severe Disabilities: A Clinical Survey. *Journal of Rehabilitation Research and Development*, 37(3):353–360, 2000.

[11] S. Gebhardt, T. Petersen-Krauß, S. Pick, D. Rausch, C. Nowke, T. Knott, P. Schmitz, D. Zielasko, B. Hentschel, and T. W. Kuhlen. Vista Widgets: A Framework for Designing 3D User Interfaces from Reusable Interaction Building Blocks. *In Proc. of ACM Conference on Virtual Reality Software and Technology*, pages 251–260, 2016.

[12] S. Harada, J. Landay, and J. Malkin. The Vocal Joystick: Evaluation of Voice-Based Cursor Control Techniques. *In Proc. of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 197–204, 2006.

[13] S. Harada, J. O. Wobbrock, and J. A. Landay. Voice Games: Investigation Into the Use of Non-speech Voice Input for Making Computer Games More Accessible. *In Proc. of the 15th IFIP International Conference on Human-Computer Interaction*, pages 11–29, 2011.

[14] J. J. Hopfield. Artificial Neural Networks. *IEEE Circuits and Devices Magazine*, 4(5):3–10, 1988.

[15] T. Igarashi and J. F. Hughes. Voice as Sound: Using Non-Verbal Voice Input for Interactive Control. *In Proc. of the 14th annual ACM symposium on User interface software and technology*, 3(2):155–156, 2001.

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

[17] 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.

[18] R. Jarina and J. Olajec. Discriminative Feature Selection for Applause Sounds Detection. *IEEE International Workshop on Image Analysis for Multimedia Interactive Services*, pages 13–13, 2007.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *In Proc. of ACM International Conference on Multimedia*, pages 675–678, 2014.

[20] T. Kohonen. Learning Vector Quantization. *Self-Organizing Maps*, pages 175–189, 1995.

[21] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. *In Proc. of International Symposium of Music Information Retrieval*, 2000.

[22] H. B. Olafsdottir, Y. Guiard, O. Rioul, and S. T. Perrault. A New Test of Throughput Invariance in Fitts' Law: Role of the Intercept and of Jensen's Inequality. *In Proc. of BCS Interaction Specialist Group Conference on People and Computers*, pages 119–126, 2012.

[23] S. N. Patel and G. D. Abowd. BLUI: Low-Cost Localized Blowable User Interfaces. *In Proc. of the 20th ACM Symposium on User Interface Software and Technology*, pages 217–220, 2007.

[24] O. Poláček, A. J. Sporka, and P. Slavík. A Comparative Study of Pitch-Based Gestures in Nonverbal Vocal Interaction. *IEEE Transactions on Systems, Man and Cybernetics*, 42(6):1567–1571, 2012.

[25] L. Rabiner and B. Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[26] J. Segen and S. Kumar. Shadow Gestures: 3D Hand Pose Estimation Using a Single Camera. *In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:479–485, 1999.

[27] R. W. Soukoreff and I. S. MacKenzie. Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI. *International Journal of Human Computer Studies*, 61(6):751–789, 2004.

[28] 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.

[29] A. J. Sporka, S. H. Kurniawan, and P. Slavik. Whistling User Interface (U^3I). *8th ERCIM Workshop on User Interfaces for All*, 3196:472–478, 2004.

[30] J. A. Suykens and J. Vandewalle. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[31] R. J. Teather and W. Stuerzlinger. Pointing at 3D Targets in a Stereo Head-Tracked Virtual Environment. In *In Proc. of IEEE 3D Symposium on User Interfaces*, pages 87–94. IEEE, 2011.

[32] R. J. Teather and W. Stuerzlinger. Pointing at 3D Target Projections with One-Eyed and Stereo Cursors. *In Proc. of SIGCHI*, pages 159–168, 2013.

[33] 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.

[34] 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.

[35] 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.

[36] D. Zielasko, D. Rausch, Y. C. Law, T. C. Knott, S. Pick, S. Porsche, J. Herber, J. Hummel, and T. W. Kuhlen. Cirque des Bouteilles: The Art of Blowing on Bottles. *In Proc. of IEEE 10th Symposium on 3D User Interfaces*, pages 209–210, 2015.