PLOS ONE

RESEARCH ARTICLE

Audio-based performance evaluation of squash players

Katalin Hajdú-Szücs[1]*, Nóra Fenyvesi[2], József Stéger[2], Gábor Vattay[2]

1 Dept. of Information Systems, Eötvös Loránd University, Budapest, Hungary, **2** Dept. of Physics of Complex Systems, Eötvös Loránd University, Budapest, Hungary

* szucsk@caesar.elte.hu

Citation: Hajdú-Szücs K, Fenyvesi N, Stéger J, Vattay G (2018) Audio-based performance evaluation of squash players. PLoS ONE 13(3): e0194394. https://doi.org/10.1371/journal.pone.0194394

Editor: Sven G. Meuth, Universitatsklinikum Munster, GERMANY

Received: April 3, 2017

Accepted: March 4, 2018

Published: March 26, 2018

Data Availability Statement: Data are available at: https://figshare.com/projects/Audio-based_performance_evaluation_of_squash_players/30115.

Funding: The authors thank the Hungarian National Research, Development and Innovation Office under Grant No. 125280 and the grant of Ericsson Ltd. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In competitive sports it is often very hard to quantify the performance. A player to score or overtake may depend on only millesimal of seconds or millimeters. In racquet sports like tennis, table tennis and squash many events will occur in a short time duration, whose recording and analysis can help reveal the differences in performance. In this paper we show that it is possible to architect a framework that utilizes the characteristic sound patterns to precisely classify the types of and localize the positions of these events. From these basic information the shot types and the ball speed along the trajectories can be estimated. Comparing these estimates with the optimal speed and target the precision of the shot can be defined. The detailed shot statistics and precision information significantly enriches and improves data available today. Feeding them back to the players and the coaches facilitates to describe playing performance objectively and to improve strategy skills. The framework is implemented, its hardware and software components are installed and tested in a squash court.

Introduction

At present in competitive sports there are a lot of talented sportsmen and the differences between individual performance are often very small to spot. It catalyses a race condition to be present already in the practising period, thus more and more coaches and players seek finding different means and aids to elaborate and make the preparation for the tournaments always more effective. There are a lot of new technological achievements available in the market. Small electronic devices are capable of measuring various metrics including those that are relevant for the sports, like heart rate and blood temperature and pressure registers, pedometers, speedometers and accelerometers to name a few. Using such devices is more than necessary since the results in a competition and then the final scores may depend on millesimal of millimeters. Another reason why to use measurement devices yielding objective performance metrics is because when sportsmen are overloaded in a performance, with adrenalin in their vein, it is hard if possible for them to spot and fix their failures. In certain types of sports a continuous or prompt feedback is definitely helpful, squash is one of them.

Squash is a very rapid ball and racquet game with typically 40-60 hit events per minute. Depending on the various surfaces the ball interacts during its flight defines the different shot classes. Some shot classes are very rare due to being tricky to deliver or may occur only in circumstances where the rally may seem already lost. So knowing the detailed statistics of various hits and shot patterns talks about the quality of the sportsmen and are very important information for both the coaches and the squash players. However, these data and their statistical analysis are not available at present because of the paste of squash. Given its fast speed the human processing of events enables the score registration in real-time only, but the recording of shot types and the detailed sequences of the shots are rendered definitely impossible. One possible solution might be to analyse videos of the matches using image processing as it has been shown to work for the tennis [1]. Though for the squash it turns out that this approach remains difficult even with the use of high speed and high resolution cameras, due to the small size of the ball and the view provided by the cameras. Traditionally cameras are placed behind the court, therefore the players will most often cover the sight of the ball during the match making the reconstruction of ball trajectories an inauspicious problem. To provide reliable statistics by this approach will require human processing and validation so in the end a thorough analysis of the tournament will cost many times of the duration of this sport events in man-hours.

In this study we introduce a framework to unhide these information based on the analysis of acoustic data. Playing squash produces characteristic sound patterns. The sound footprint of each rally is a projection of all the details about the strength and the position of the ball hitting various surfaces in the court. Naturally, this pattern, which maintains the natural order of the events, is contaminated by some additional noise. Recording the sound in more directions allows for inverting the problem and for giving statistical statements about where and what type of an event took place in the play. We are focusing on events generated by the ball hits, which serves as a basis for further analysis and the reconstruction of shot patterns or the ball trajectories. Note, the framework to be detailed can be applied to various other types of ball games.

Related work

Squash and soccer were the first sports to be analysed by ways of analysis systems. Formal scientific support for squash emerged at the late 1960s. The current applications of performance analysis techniques in squash are deeply investigated in the book of Stafford et al. [2].

One test that was developed by squash coach Geoffry Hunt is the "Hunt Squash Accuracy Test" (HSAT) [3], that is a reliable method used by coaches to assess shot hitting accuracy. The test is composed of 375 shots across 13 different types of squash strokes and it is evaluated based on a total score expressed as the number of successful shots.

Recent technological advances have facilitated the development of sport analytical software such as Dartfish video based motion analysis system [4, 5]. However, these systems still require a considerable amount of professional assistance.

To the best of our knowledge there is no previous research investigating the applicability of sound analysis techniques for squash performance analysis, therefore it is not possible to directly compare our system to existing solutions. In other application environments a wide literature can be found on real-time sound source localization that is the most closely related topic to our work. The emerging application of camera pointing in video conferencing environment motivated many research papers on the field of visual speaker localization [6–10]. A linear-correction least-squares estimation procedure is proposed in [6–8]. The simulation results in [7] show that the bias level of this technique is around 30 cm. In the work of Tobias Gehrig et al. [10] a method is presented to speaker tracking using audio-visual features, namely

time delay of arrival estimation on microphone array signals and face detection on multiple camera images. The sound source localization is based on a maximum likelihood approach. In the experimental results the authors measure 57.2 cm root mean square error for the audio-only solution and 49.9 cm for the audio-video approach. One conceptually simple solution for source localization is beamforming [11], where the source location is estimated by calculating the steered output power of a beamformer over a set of candidate locations. Although this concept has advantages in speech localization and enhancement, it is computationally expensive and its resolution is too low for our purposes.

The measurement equipment

This study is based on the analysis of sound waves generated during the squash play. Among many other, squash is a game where various different sources of sound are present, including the players themselves (their sighing or their shoes squeaking on the floor), the ball hitting surfaces (like the walls, the floor or the racquet) and also external sources (including the ovation of the spectators or sound generated in an adjacent court). Here we focus on audio events related to the ball.

When planning the experiments the following constraints had to be investigated and satisfied. The framework should be fast in signal processing point of view, because the target information can be most valuable when in a competitive situation it helps fine tune tactical decisions made by the coach and/or the player. The cost of the equipment should be kept low and the installation of the sensors requires a careful design to prevent them from disrupting the play. As the spatial localization of the ball is one of the fundamental goals a lower bound to the sampling rate is enforced to remain able to differentiate between displaced sound sources.

In Fig 1 the hardware and software components are sketched. Hardware components include 6 audio sensors, three of which are omnidirectional microphones (Audio Technica ES945) sinking in the floor and the rest of them are cardioid microphones (Audio Technica PRO 45) hanging from the top. Amplification and sampling of the microphone signals are done by a single dedicated sound card (Presonus AudioBox 1818VSL) so that all channels in a

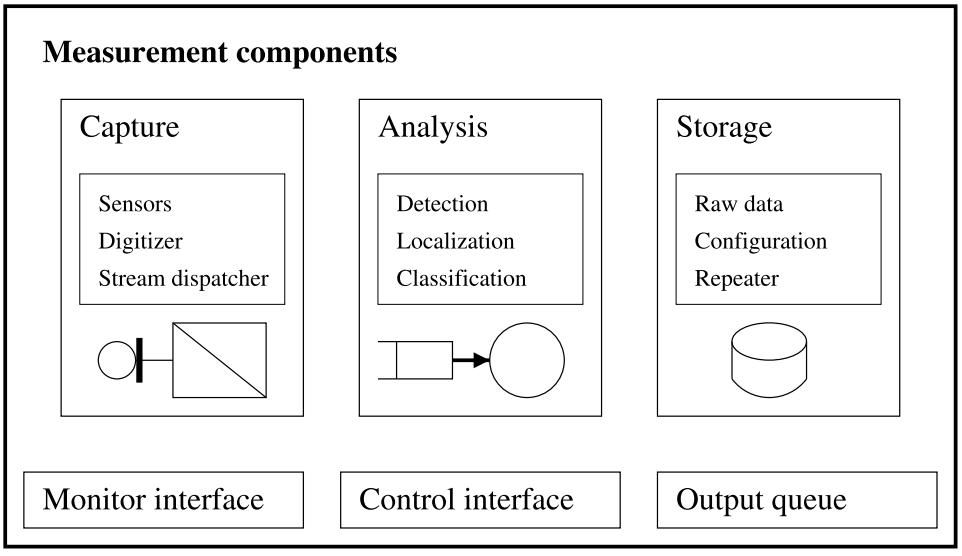

Fig 1. A schematic view of the components. To process audio events in the squash court a three component architecture was designed.

https://doi.org/10.1371/journal.pone.0194394.g001

sample frame are in synchrony. The highest sampling rate of the sound card is used (96 kHz), so by each new sample the front of a sound wave travels approximately 6 mm.

According to their functionalities software components fall in the following groups. Signal processing is done in the analysis module, which include the detection of the audio events, the classification and the filtering of the detections and after matching event detections of more channels the localization of the sound source. While these steps of signal processing can be done real-time a storage module is also implemented so that the audio of important matches can be recorded. Recording of data helps training of the parameters of the classification algorithms, and it also enables a whole re-analysis of former data with different detectors and/or different classifiers. All output generated by the Analysis module is fed to the output queue. Hardware and software components are triggered and reconfigured via a web services API exposed by the Control interface. Finally, to be able to listen to what is going on in the remote court a Monitoring interface provides a mixed, downsampled and compressed live stream across the web.

The ball impact detection

The localization and the classification of ball hits both require the precise identification of the beginning of the corresponding events in the audio streams. The detection of ball impact events is carried out for each audio channels independently and in a parallel fashion, which speeds up the overall performance of the framework significantly. Different detection algorithms of various complexities were investigated two extreme cases are sketched here. The first model assumes that the background noise follows the normal distribution. An event is detected if new input samples deviate from the Gaussian distribution to a certain predefined threshold value. Next for each channels the mean and the variance estimates of a finite subset of the samples are continually updated according to the Welford's algorithm [12].

The second method is an extension of the windowed Gaussian surprise detection by Schauerte and Stiefelhagen [13]. The algorithm tackles the problem evaluating the relative entropy [14]. It is first applied in the frequency domain and if there is a detection then a finer scale search is carried out in the time domain. The power spectrum of w-sized chunks of windowed data samples is calculated. Between detection regime the series of the power spectra is modelled by a w-dimensional Gaussian. The a priori parameters of the distribution are calculated for n elements in the past, and the posteriori parameters are approximated including the new power spectrum. The Kullback Leibler divergence between the a priori and the posteriori distributions exceeds a predefined threshold when a new detection takes place

$$S_i = \frac{1}{2}\left[\log\frac{|\Sigma_i|}{|\Sigma_i'|} + \mathrm{Tr}\,(\Sigma_i^{-1}\Sigma_i') - w + (\mu_i' - \mu_i)^T\Sigma_i^{-1}(\mu_i' - \mu_i)\right],$$

where primed parameters correspond to the posteriori distribution. The time resolution at this stage is w and to increase precision a new search is carried out in the time domain evaluating the Kullback Leibler divergence for 1-d data. In order to bootstrap a priori distribution parameters n samples from the former windows are used.

The localization of sound events

In this section we lay down a probabilistic model to determine the time and location of an audio event. For a unique event we denote these unknowns t and $\mathbf{r}_{ev}$ respectively. The location vector $\mathbf{r}_{ev}$ is a 3 dimensional array of Descartes coordinates (x, y, z), however, the calculation presented here also applies for lower dimension setups. The inputs required to find the audio

event are the locations of the $N+1$ detectors $\mathbf{r}_i^{\text{mike}}$ and the timestamps τ_i when these synchronized detectors sense the event ($0 \leq i \leq N$).

The probability that microphone i detects an event at $(\mathbf{r}, t)$ is

$$p(t_i, r_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp - \frac{(ct_i - r_i)^2}{2\sigma_i^2 c^2},$$

where c is the speed of sound, $t_i = \tau_i - t$ is the propagation delay and $r_i = ||\mathbf{r} - \mathbf{r}_i^{\text{mike}}||$ is the distance between the sound source and the microphone. The uncertainty σ_i depends on the characteristics of the microphone, which we will consider constant in the first approximation.

By introducing relative delays $\hat{\tau}_i = \tau_i - \tau_0$ the joint probability of relative delays detected is

$$p(\hat{\tau}_1, \ldots \hat{\tau}_N) = \int \mathrm{d}t_0\, p(t_0, r_0) \prod_{i=1}^{N} p(\hat{\tau}_i + t_0, r_i).$$

The formula can be rearranged

$$p(\hat{\tau}_1, \ldots \hat{\tau}_N) = \frac{1}{\sqrt{2\pi}^{N+1} \prod_{i=0}^{N} \sigma_i} \int \mathrm{d}t_0\, e^{-f(t_0)},$$

where $f(t_0) = \sum_{i=0}^{N} \frac{(c\hat{\tau}_i + ct_0 - r_i)^2}{2\sigma_i^2 c^2}$ is a quadratic function and in the expression for p the Gaussian integral follows

$$\int \mathrm{d}t_0\, e^{-f(t_0)} = \sqrt{\frac{2\pi}{f''(t_0^*)}} e^{-f(t_0^*)}.$$

The first order derivative f' vanishes in $t_0^* = \Sigma^2 \sum_{i=0}^{N} \frac{1}{\sigma_i^2} \left(\frac{r_i}{c} - \hat{\tau}_i\right)$, where $\Sigma^2 = 1/\sum_{i=0}^{N} \frac{1}{\sigma_i^2}$ is introduced for convenience.

After substitution of t_0^* we arrive at

$$f(t_0^*) = \frac{1}{2} \left\{ \sum_{i=0}^{N} \frac{1}{\sigma_i^2} \left(\frac{r_i}{c} - \hat{\tau}_i\right)^2 - \Sigma^2 \left[\sum_{i=0}^{N} \frac{1}{\sigma_i^2} \left(\frac{r_i}{c} - \hat{\tau}_i\right) \right]^2 \right\}.$$

This formula can be interpreted as a variance formula, which can be rewritten

$$f(t_0^*) = \frac{1}{2\Sigma^2} \sum_{i=0}^{N} \frac{1}{\sigma_i^2} \left[\sum_{j=0}^{N} \frac{1}{\sigma_j^2} \left(\frac{r_i - r_j}{c} - (\hat{\tau}_i - \hat{\tau}_j) \right) \right]^2.$$

A good approximation of the audio event maximizes the likelihood p, which at the same time minimizes $f(t_0^*)$, thus we seek the solution of $\nabla_{\mathbf{r}} f(t_0^*) = 0$ equations.

In practice f behaves well and its minimum can be found by gradient descent method. Fig 2 shows a situation, where the ball hit the front wall and 6 microphones detect this event error free. To show the functions behaviour f is evaluated in the floor, in the front wall and in the right side wall. Finding the minimum of f takes less than ten gradient steps.

The likelihood based localization model is derived for a noiseless situation, assuming the perfect detection of samples in each channel. In real environment, however, noise is present and the error deviating the detection is exposed in the final result of the localization. In order to track this effect the method was numerically investigated as follows. 10000 points in the volume of the court is selected randomly and the sound propagation is calculated in each six microphones. Next for the ideal detections Gaussian noise is added in all channels, with

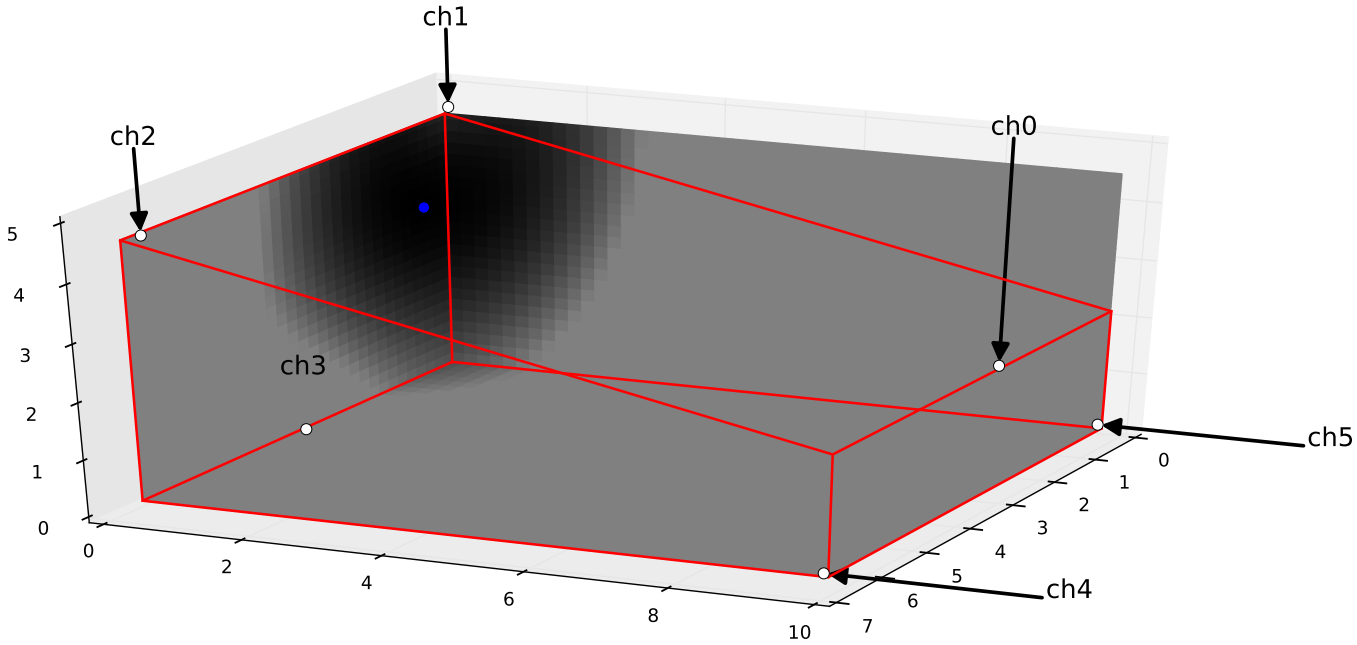

Fig 2. The visualization of the likelihood function. The ball hit the front wall. $f(t_0^*)$ can be evaluated in space given the positions of the sensors (marked by white disks) to find its minimum, which indicates where the event took place. (0.5 m from the right corner and 3 m above the floor, marked by a blue disk).

https://doi.org/10.1371/journal.pone.0194394.g002

increasing variation ($\sigma = 1, 10, 50$). In Fig 3 the noiseless case is compared to cases with increasing errors. In the figure the cumulative distribution of the error, ie. the difference between the randomly selected point and the location guess by the model is presented. Naturally, by increasing the detection error the error in the position guess is increasing, but the

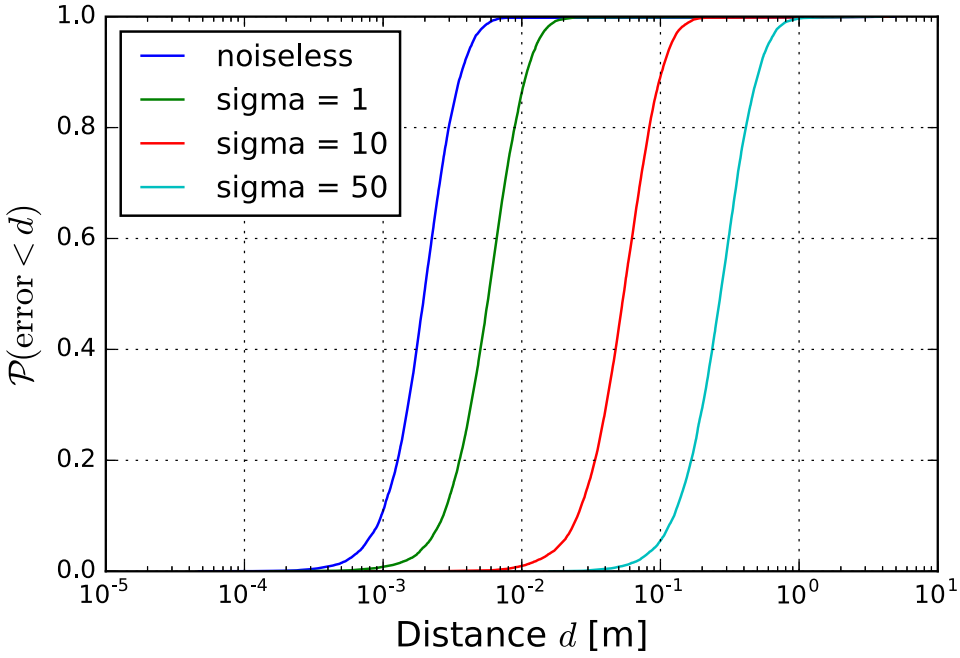

Fig 3. The cumulative distribution of the localization error. For a noiseless case most often localization will have an error comparable to the size of the ball. With a bad detector ($\sigma = 50$ samples) still the localization is exact in the order of 10 cm.

https://doi.org/10.1371/journal.pone.0194394.g003

model performs very well, for poor signal detectors the error in localization is in the order of 10 cm.

Classification

It is the task of the classification module to distinguish between the different sound events according to their origin. Sound events are classified based on the type of the surface that suffered from the impact of the ball. This surface can be the wall, the racquet, the floor or the glass. When the sound does not fit any of these classes, like the squeaking shoes, then it is classified as a false event. The classification enhances the overall performance of the system by two means. First, skipping to localize the false events speeds up the processing. And second, in doubtful situations when the calculated location of the event falls near to multiple possible surfaces, by knowing the type of the surface that suffered from the impact can reinforce the localization. For example a sound event localized a few centimetres above the floor could be generated by a racquet hit close to the floor or by the floor itself.

Classification utilizes feed-forward neural networks that had been trained with backpropagation [15–18]. The training sets are composed of vectors belonging to 5461 audio events, which have been manually labelled. Based on these audio events two types of input were constructed for teaching.

In the first case temporal data is used directly. A vector element of the training set T_1 is the sequence of the samples around the detections for each channels.

$$T_1 = \{(a_{d-w}, \ldots, a_d, \ldots, a_{d+w})\},$$

where the channel index is dropped and d is a unique detection and w sets the length of the vector. Given the sampling rate 96 kHz and setting $w = 300$ the neural network is taught by 6.25 millisecond long data.

The second feature set T_2 is built up of the power spectra.

$$T_2 = \{|\mathcal{F}(a_d, \ldots, a_{d+w})|\},$$

where $\mathcal{F}$ denotes the discrete Fourier transform.

A single neural network model where all event classes are handled together performed poorly in our case. Therefore, separate discriminative neural network models were built for all four classes (racquet, wall, floor and glass impact) and for both of the training sets. It has also been investigated if any of the input channels introduce discrepancy. In order to discover this effect models were built and trained for each unique channels and another one handling the six channels together. Note, that not all possible combinations of the models were trained due to the fact that some channels poorly detected certain events, for example microphones near the front wall detected glass events very rarely.

In the training sets the class of interest was always under-represented. To balance the classifier the SMOTE [19] algorithm was used, which is a synthetic minority over-sampling technique. A new element is synthesized as follows. The difference between a feature vector from the positive class and one of its k nearest neighbours is computed. The difference is blown by a random number between 0 and 1, to be added to the original feature vector. This technique forces the minority class to become more general, and as a result, the class of interest becomes equally represented like the majority set in the training data.

Different network configurations were realized to find that for the direct temporal input a 20 hidden layer network (with 10 neurons in each layer) performed the best, while for the spectra input a 10 hidden layer (each layer with 10 neurons) is the best choice.

Analysis

In this section the performance of each modules of the framework and the datasets are presented.

Datasets

In order to analyse the components of the framework implementing the proposed methods two audio and video record sets were used. Datasets are available at https://figshare.com. *Audio 1* was recorded on the 18th of May 2016 when a squash player was asked to target specific areas of the wall. This measurement was necessary to increase the cardinality of the different hits significantly in the training datasets T_1 and T_2, and it was also manually processed to be able to validate the operations of the detector and the localization components. *Audio 2* resembles data in a real situation as it contains a seven minutes squash match recorded on the 8th of March 2016. Table 1 summarizes the details of these audio recordings.

Training the neural network models require properly labelled datasets. After applying the ball impact detection algorithm to the audio records the timestamps of the detected events were manually categorized as front wall event, racquet event, floor event or glass event. In the categorization procedure video files helped in doubtful situations. Every sudden sound effect that was detected by the algorithm but does not belong to these relevant classes was labeled as false event. In *Audio 1* prescribed audio events were generated and recorded and it does not contain any false events. In contrast, *Audio 2* was captured in a real situation and it presents several false events by nature.

Detection results

The performance of the detector is analysed by comparing the timestamp reported by the detector d_{detector} and the human readings d_{human}. For *Audio 1* in Fig 4 the cumulative probability distribution of the time difference is shown for each channel and in Table 2 the average error and its variance are shown grouped by the event types present in the dataset. One can observe that the detectors in channels *ch4* and *ch5* perform poorly for front wall and racquet events. When estimating the position discarding one of or both of these channels will enhance the precision of the localization. However, for floor events, these two channels performed the

Table 1. The content of the audio files.

	Class	Ch0	Ch1	Ch2	Ch3	Ch4	Ch5	Total
Audio 1	Front wall	165	165	165	165	165	165	990
	Racquet	166	166	166	166	166	166	996
	Floor	30	30	30	30	30	30	180
	Glass	25	25	25	25	25	25	150
	Total	386	386	386	386	386	386	2316
Audio 2	Front wall	100	109	108	110	107	111	645
	Racquet	112	112	113	110	109	99	655
	Floor	85	70	75	19	115	11	375
	Glass	46	20	24	15	62	11	178
	False event	227	274	254	264	456	147	1622
	Total	570	585	574	518	849	379	3475

The count of events in *Audio 1* and *Audio 2* broken down for each class and each channel. In total 5791 events have been labeled.

https://doi.org/10.1371/journal.pone.0194394.t001

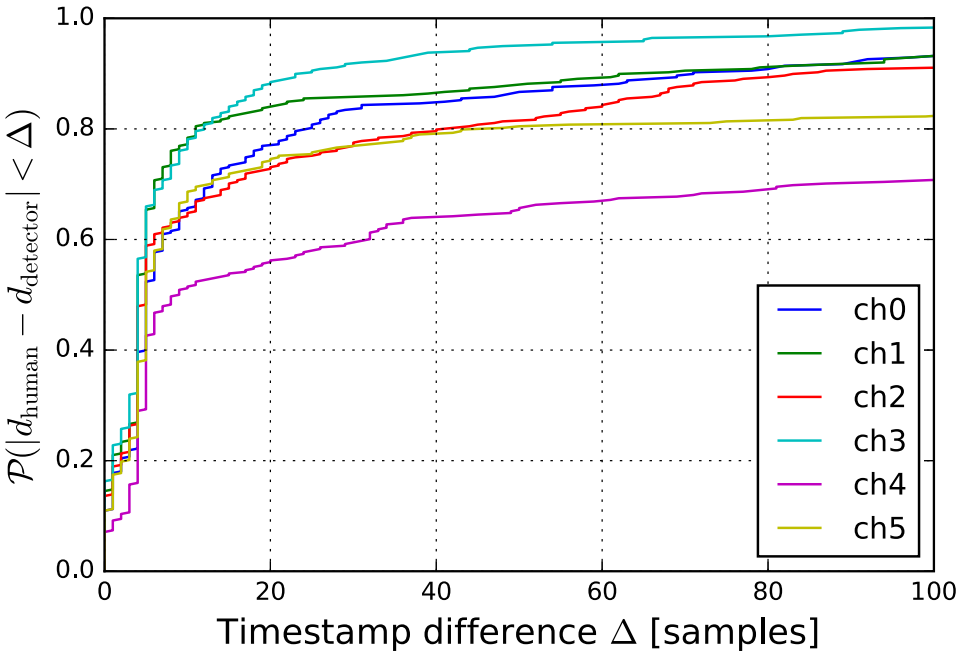

Fig 4. The error of the detector. The detection error is defined as the difference between the timestamps generated by the module and read by a human.

https://doi.org/10.1371/journal.pone.0194394.g004

best along with channel *ch2*. For glass events the smallest deviations were measured on channels *ch2*, *ch3* and *ch5*.

In Table 3 the error statistics for dataset *Audio 2* is shown. Intensive events, like front wall impacts, can be detected precisely, whereas the detection of milder sounds like a floor or glass impact is less accurate.

The false discovery and the false negative rate of the detector were examined on *Audio 2*. False positives are counted if detector signals for a false event, and false negatives are the missing detections. The results are summarised in Table 4.

Classification results

Approaching the problem at first and to use as much information as possible to teach the neural networks a large training set was constructed of the union of the detections of all the six channels. However, this technique gave poorer results than treating all the channels separately.

Table 2. The class and channelwise error of the detector.

	Front wall	Racquet	Glass	Floor
ch0	9.6 ± 46.0	-5.8 ± 63.7	22.1 ± 33.6	38.3 ± 81.3
ch1	3.1 ± 1.9	-9.3 ± 130.6	88.4 ± 42.8	12.0 ± 15.8
ch2	3.5 ± 5.4	21.3 ± 129.3	7.7 ± 6.8	7.8 ± 10.9
ch3	3.0 ± 1.9	7.3 ± 39.9	9.7 ± 23.2	33.1 ± 63.4
ch4	221.4 ± 476.5	116.4 ± 401.3	31.0 ± 55.2	5.7 ± 16.5
ch5	210.8 ± 512.3	23.5 ± 136.2	7.7 ± 26.6	2.4 ± 2.1

The error of the detector algorithm is measured in samples for the various classes and all channels. The sampling rate is 96 kHz (1 sample ≈ 0.01 ms).

https://doi.org/10.1371/journal.pone.0194394.t002

Table 3. Classwise error of the detector.

Class	Audio 1	Audio 2
Front wall	4.8 ± 23.3	6.9 ± 19
Racquet	3.4 ± 99.8	107 ± 85
Floor	38.0 ± 141.1	125 ± 149
Glass	*n.a.*	183 ± 173

The statistics of the dataset *Audio 1* is calculated for 660 events for each class excluding Floor events, counting 24 pieces. For *Audio 2* 200 events were available for each class. The sampling rate is 96 kHz (1 sample ≈ 0.01 ms).

https://doi.org/10.1371/journal.pone.0194394.t003

Table 4. Performance of the detector.

False alarm	Ch0	Ch1	Ch2	Ch3	Ch4	Ch5
FDR	39%	47%	44%	51%	54%	39%
FNR	16%	24%	22%	38%	5%	43%

False Discovery Rate (FDR: $\frac{n_{fp}}{n_{tp}+n_{fp}}$) and False Negative Rate (FNR:$\frac{n_{fn}}{n_{fn}+n_{tp}}$) of the detector based on 3475 events.

https://doi.org/10.1371/journal.pone.0194394.t004

The different settings of the microphones and the distinct acoustic properties of the squash court at the microphone positions are found to be the reasons of that phenomenon.

Eight-fold cross-validation [20] was used on the datasets to evaluate the performance of the classifiers. Three measures are investigated closer: the accuracy, the precision and the recall. Accuracy (in Fig 5) is the ratio of correct classifications and the total number of cases examined ($\frac{n_{tp}+n_{tn}}{n}$). Precision (in Fig 6) is the fraction constrained to the relevant cases ($\frac{n_{tp}}{n_{tp}+n_{fp}}$). Recall (in Fig 7) is the fraction of relevant instances that are retrieved ($\frac{n_{tp}}{n_{tp}+n_{fn}}$).

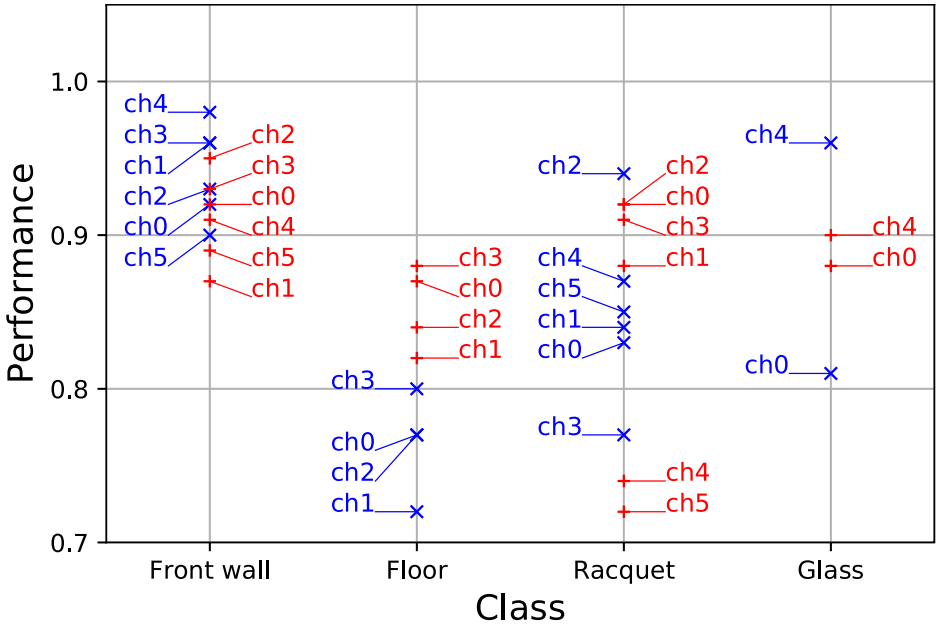

Fig 5. The classifiers' accuracy. The classwise accuracy of each channel is presented in T_1 (blue) and T_2 (red) input sets. Front wall classification gives high accuracy on all channels in both sets. It is interesting to observe that floor classification is more accurate in input T_2. Racquet classification performs best on channel 2 in both sets.

https://doi.org/10.1371/journal.pone.0194394.g005

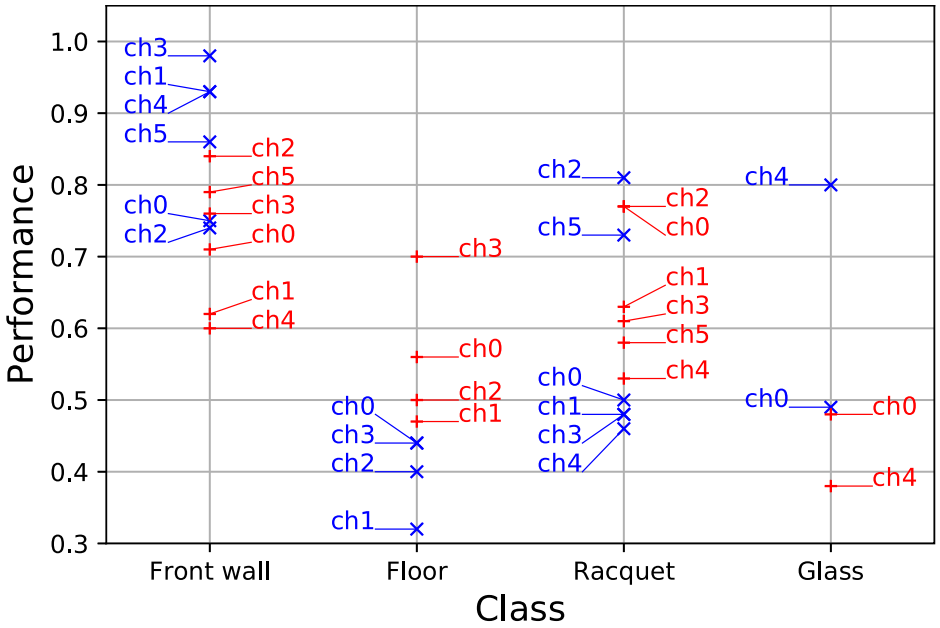

Fig 6. The classifiers' precision. The classwise precision of each channel is presented in T_1 (blue) and T_2 (red) input sets. Front wall classification gives high precision in input T_1. The precision of floor classification is low. Racquet classification still performs best on channel 2. The precision of glass classification is only acceptable on channel 4.

https://doi.org/10.1371/journal.pone.0194394.g006

Table 5 summarises the results of the best classifiers for each class. It can be seen that the classification of the front wall and the racquet events is reliable. However, the precision and the recall of floor and glass events are poor. The reason for it is that these classes are under-represented in the data sets. Whenever x, an unseen sample comes, the best classifiers of each

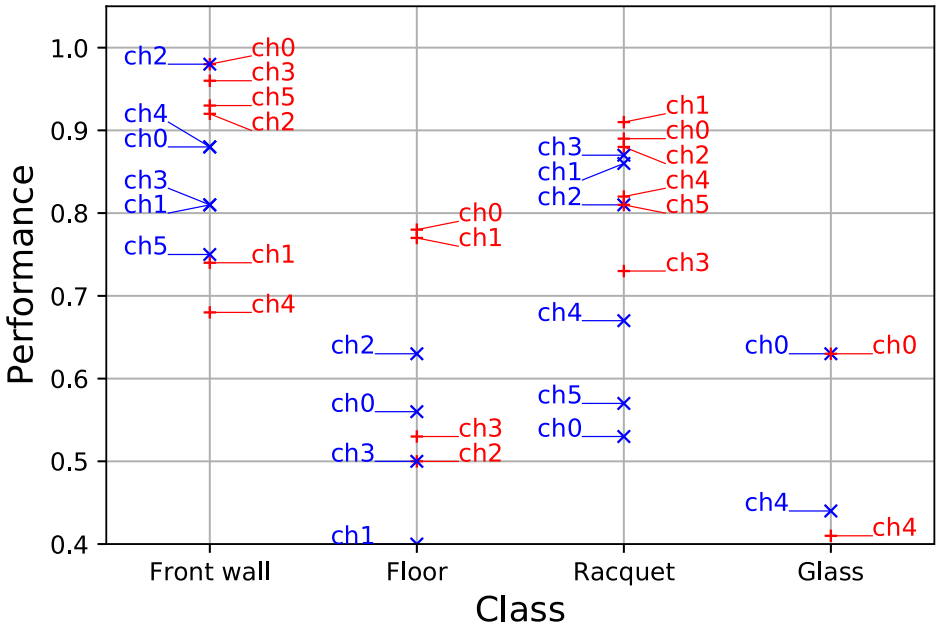

Fig 7. The classifiers' recall. The classwise recall of each channel is presented in T_1 (blue) and T_2 (red) input sets. The performance of front wall classification is reliable. The recall of racquet classification is high on channels 1 and 2 in both sets. However, the performance of floor and glass classifications is low.

https://doi.org/10.1371/journal.pone.0194394.g007

Table 5. The classwise performance of the best classifiers.

Class	Channel	Input	Acc	Prec	Rec
Front wall	ch4	T_1	0.98	0.93	0.88
Racquet	ch2	T_1	0.94	0.81	0.81
Floor	ch4	T_2	0.88	0.53	0.7
Glass	ch0	T_2	0.88	0.63	0.5

https://doi.org/10.1371/journal.pone.0194394.t005

class are applied on the new element. The prediction of the class label $\hat{y}$ to which x belongs to is computed by the following formula:

$$\hat{y} = \begin{cases} \arg\max_{k \in C} \left\{ \frac{f_k(x) - \text{cut}_k}{1 - \text{cut}_k} \frac{\text{prec}_k}{\sum_{i \in C} \text{prec}_i} \right\}, & \exists k : f_k(x) > \text{cut}_k \\ \text{false event}, & \text{otherwise} \end{cases}$$

where C is the set of class labels without the class of false events and $f_k(x)$, cut_k and prec_k are the confidence, the cutoff value and the precision of the best classifier in class k respectively.

Fig 8 depicts the combined output generated by the detector and the classifier modules. A 1.77 seconds long segment of channel 1 audio samples are grabbed from *Audio 2*. Detections and resolved classes are also shown. From the snapshot one can observe the different intensities of the events. Generally the change in the ball's moment happens when a racquet or a front wall impacts and the sample amplitudes are higher, whereas floor and glass events tend to generate lower intensity and are harder to detect.

Localization results

Based on the geometry of the court, the placement of the microphones and using the localization technique detailed in this study for each set of detection timestamps the 3-d position of the source of the event can be estimated. In case not all source channels provide a detection of the event localization is still possible. Four or more corresponding timestamps will yield a 3-d estimate, whereas with three timestamps the localization of events constrained on a surface (e.g. planes like wall or floor) remains possible.

In Fig 9 the located events present in dataset *Audio 1* are shown. In this measurement scenario the player was asked to hit different target areas on the front wall. It was a rapid exercise, as the ball was shot back at once. Only a few times the ball hit the floor, most of the sound is composed of alternating racquet and front wall events. In Fig 10 the front wall events are shown. The target areas can be seen clearly, and also it is visible the spots scatter a little more

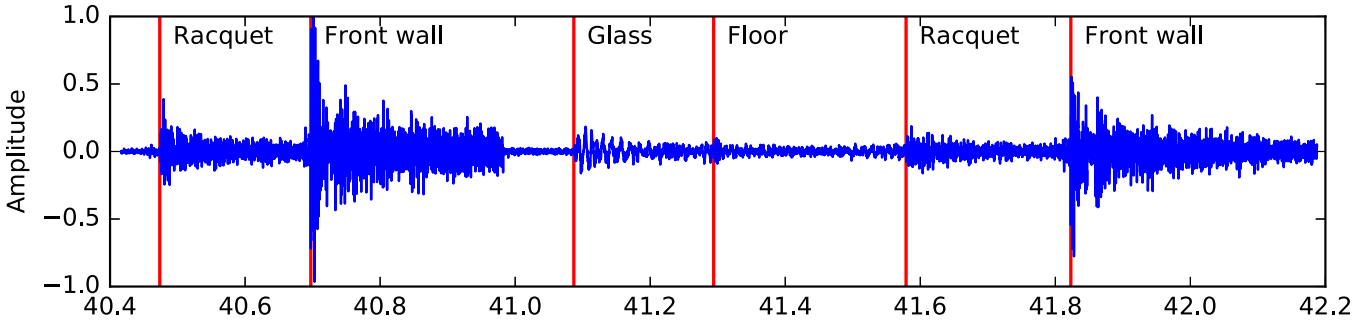

Fig 8. Labelled audio signal. 1.77 second long samples from channel ch1 in *Audio 2*. Detected timestamps and the event classes are marked.

https://doi.org/10.1371/journal.pone.0194394.g008

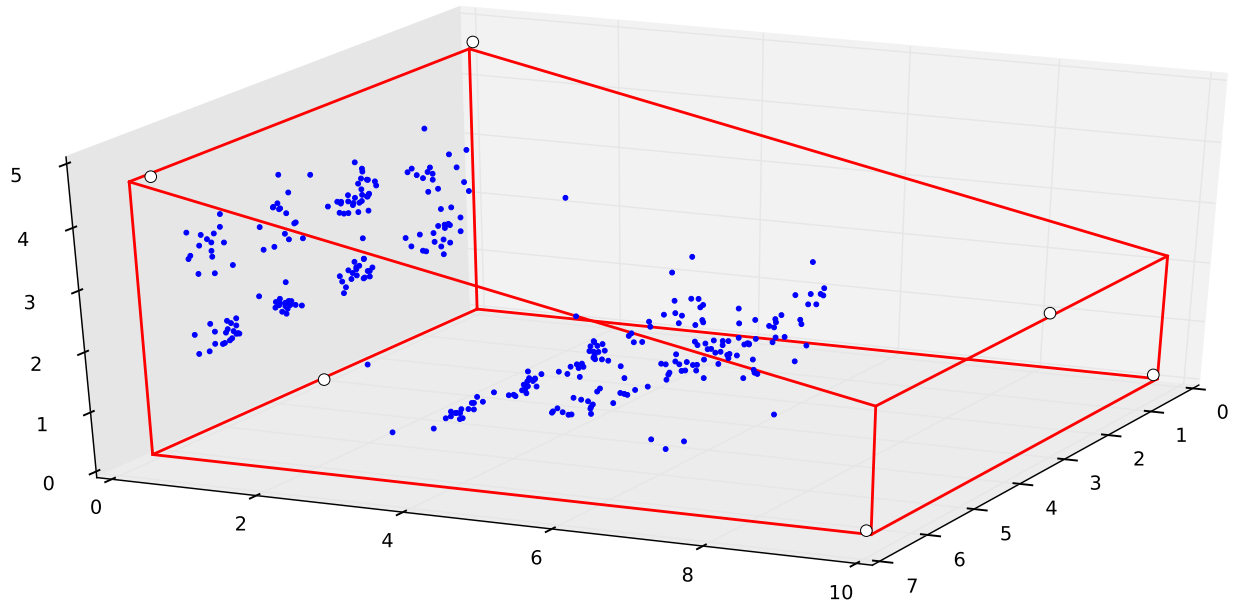

Fig 9. The position of impacts. Visualize the localized events embedded in 3-d.

https://doi.org/10.1371/journal.pone.0194394.g009

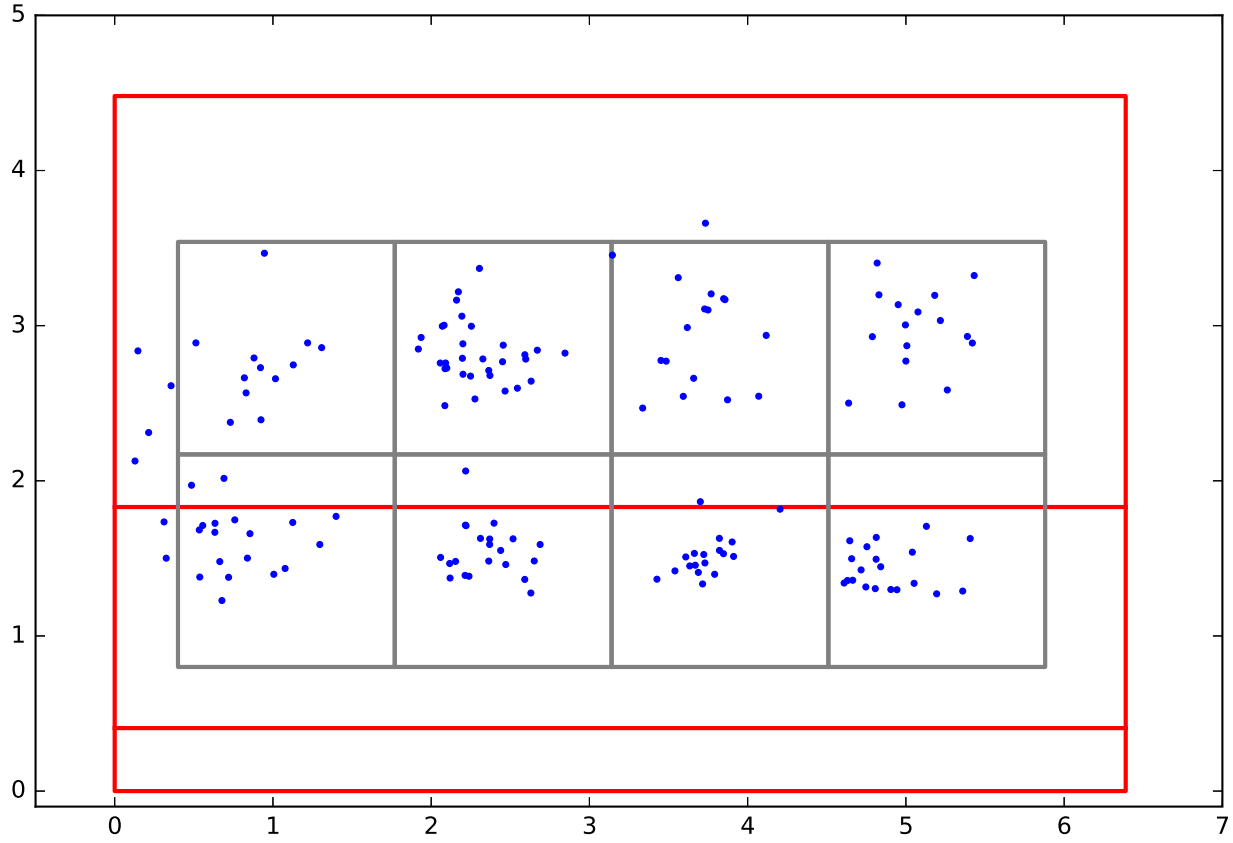

Fig 10. Front wall impacts. Gray squares embrace the eight target areas.

https://doi.org/10.1371/journal.pone.0194394.g010

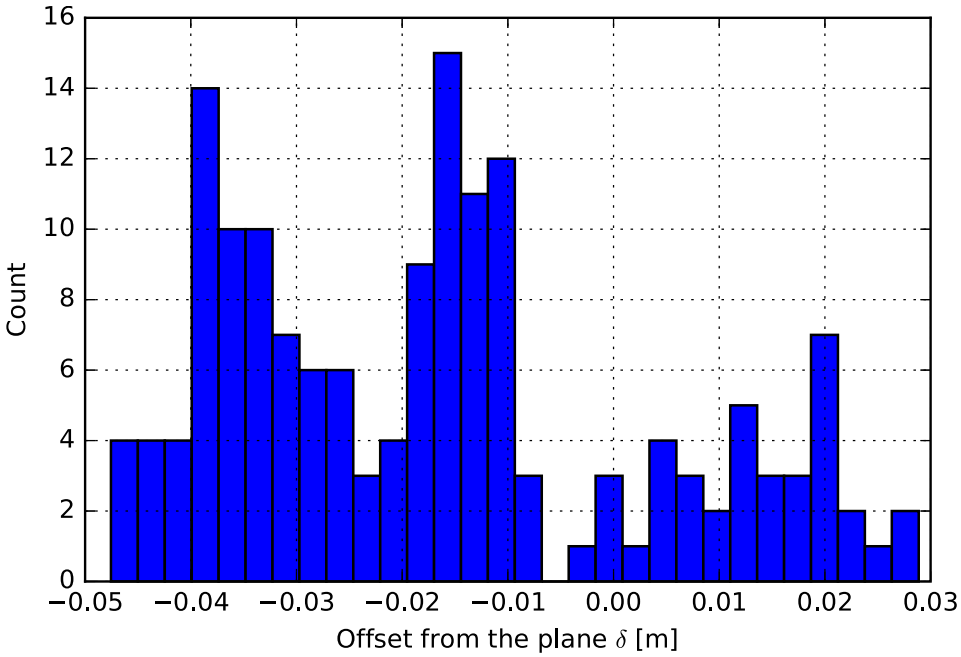

Fig 11. The front wall offsets. The distribution of the offset δ from the front wall ($\sigma(\delta) \approx 0.02$ m).

https://doi.org/10.1371/journal.pone.0194394.g011

on the left. The reason could be the player being right handed or the fact the target area was hit later during the experiment and the player showed tiredness.

Measuring the error of the localization method is not straight forward because the ball hitting the main wall does not leave a mark, where the impact happened and there was no means to take pictures of these events. Taking advantage of the geometry of the front wall an error metric can be defined for front wall events. The error δ is defined by the offset of the approximated location from the plane of the front wall. In Fig 11 the error histogram is shown. The mean of δ should vanish and the smaller its variance the better the framework located the events. From this exercise one can read the standard deviation is $\sigma(\delta) < 3$ cm, which is smaller then the size of the squash ball.

Another way to define the error is based on relying on human readings of the events. In the dataset *Audio 1* all of the sound events were marked by human as well as by the detector algorithm. Localizing the events using both inputs the direct position difference can be investigated. The mean difference between the positions is 11.8 cm and their standard deviation is 39.9 cm.

Discussion

Our results support that in sports, where the relevant sound patterns are distinguishable, careful signal processing allows the localisation of shots. The described system is optimized for handling events and as a consequences the real-time analysis of data is possible, which is important to give an instant feedback. The framework can be extended to provide higher level statistics of events such as the evolution of shots types. From the wide range of possible applications we highlight three use cases. Firstly, during a match the players can get to know their precision in short time and if is necessary they can change their strategy. Secondly, during practice coaches can track the development of the players hit accuracy. Or thirdly, certain

exercises can be defined, which can be automatically and objectively evaluated, without the need for the coach be present during the exercise.

In this study our framework was adopted to squash. In theory it can be extended to any other sports where the stereotypical events are associated with a specific sound pattern. In those applications, where typical patterns are present but the surroundings introduce significant amount of noise, a solution could be to use additional microphones with possibly special characteristcs to record the noise allowing to subtract its contribution from all other input signals. For example in tennis games played in the open field.

Ethics statement

In this study human participants were instructed to carry out specific squash excercises. Participants were informed beforehands about the fact that during the excercise the sound is to be recorded. During the excercises the sound emerging mainly from the ball impacts was recorded. The recording itself do not contain any sensitive information. Along with the raw recordings no additional information about the participant is saved or published. Participants do not object that these recordings are made public.

Acknowledgments

The hardware components enabling this study are installed at Gold Center's squash court. We thank them for this opportunity and squash coach Shakeel Khan for the fruitful discussions.

Authors thank the Hungarian National Research, Development and Innovation Office under Grant No. 125280 and the grant of Ericsson Ltd.

Author Contributions

Conceptualization: Nóra Fenyvesi, József Stéger.

Data curation: Katalin Hajdú-Szücs, József Stéger.

Formal analysis: Katalin Hajdú-Szücs, Nóra Fenyvesi, József Stéger, Gábor Vattay.

Investigation: Nóra Fenyvesi, József Stéger.

Methodology: Katalin Hajdú-Szücs, Nóra Fenyvesi, József Stéger.

Project administration: Gábor Vattay.

Resources: Gábor Vattay.

Supervision: Gábor Vattay.

Validation: József Stéger.

Visualization: Katalin Hajdú-Szücs, József Stéger.

Writing – original draft: Katalin Hajdú-Szücs, Nóra Fenyvesi, József Stéger.

Writing – review & editing: Katalin Hajdú-Szücs, Nóra Fenyvesi, József Stéger, Gábor Vattay.

References

1. Broadbent DP, Ford PR, O'Hara DA, Williams AM, Causer J. The effect of a sequential structure of practice for the training of perceptual-cognitive skills in tennis. PLOS ONE. 2017; 12(3):1–14. https://doi.org/10.1371/journal.pone.0174311
2. OBE NM. Current applications of performance analysis techniques in squash. Science of Sport: Squash. 2016;.

3. Williams BK, Hunt GB, Graham-Smith P, Bourdon PC. Measuring squash hitting accuracy using the 'Hunt squash accuracy test'. In: ISBS-Conference Proceedings Archive; 2014.
4. Barris S, Button C. A review of vision-based motion analysis in sport. Sports Medicine. 2008; 38 (12):1025–1043. https://doi.org/10.2165/00007256-200838120-00006 PMID: 19026019
5. Travassos B, Davids K, Araújo D, Esteves PT. Performance analysis in team sports: Advances from an Ecological Dynamics approach. International Journal of Performance Analysis in Sport. 2013; 13 (1):83–95. https://doi.org/10.1080/24748668.2013.11868633
6. Huang Y, Benesty J, Elko GW, Mersereati RM. Real-time passive source localization: a practical linear-correction least-squares approach. IEEE Transactions on Speech and Audio Processing. 2001; 9 (8):943–956. https://doi.org/10.1109/89.966097
7. Huang Y, Benesty J, Elko GW. Passive acoustic source localization for video camera steering. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100). vol. 2; 2000. p. II909–II912 vol.2.
8. Huang Y, Benesty J, Elko GW. An efficient linear-correction least-squares approach to source localization. In: Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575); 2001. p. 67–70.
9. Friedland G, Yeo C, Hung H. Visual Speaker Localization Aided by Acoustic Models. In: Proceedings of the 17th ACM International Conference on Multimedia. MM'09. New York, NY, USA: ACM; 2009. p. 195–202.
10. Gehrig T, Nickel K, Ekenel HK, Klee U, McDonough J. Kalman filters for audio-video source localization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.; 2005. p. 118–121.
11. Chen JC, Yao K, Hudson RE. Source localization and beamforming. IEEE Signal Processing Magazine. 2002; 19(2):30–39. https://doi.org/10.1109/79.985676
12. Welford BP. Note on a Method for Calculating Corrected Sums of Squares and Products. Technometrics. 1962; 4(3):419–420. https://doi.org/10.1080/00401706.1962.10490022
13. Boris S, Stiefelhagen R. "Wow!" Bayesian surprise for salient acoustic event detection. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013;.
14. Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics. 1951; 22(1):79–86. https://doi.org/10.1214/aoms/1177729694
15. Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine. 2012; 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597
16. Bugatti A, Flammini A, Migliorati P. Audio classification in speech and music: a comparison between a statistical and a neural approach. EURASIP Journal on Advances in Signal Processing. 2002; 2002 (4):1–7. https://doi.org/10.1155/S1110865702000720
17. Shao X, Xu C, Kankanhalli MS. Applying neural network on the content-based audio classification. In: Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. vol. 3. IEEE; 2003. p. 1821–1825.
18. Wang Y, Lee CM, Kim DG, Xu Y. Sound-quality prediction for nonstationary vehicle interior noise based on wavelet pre-processing neural network model. Journal of Sound and Vibration. 2007; 299(4):933–947. https://doi.org/10.1016/j.jsv.2006.07.034
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002; 16:321–357.
20. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010; 4:40–79. https://doi.org/10.1214/09-SS054