DESIGNING ON SUBJECTIVE TOLERANCE TO APPROXIMATED PIANO REPRODUCTIONS

Federico Fontana, Stefano Zambon, and Yuri De Pra

Dipartimento di Matematica e Informatica
University of Udine, Italy
federico.fontana@uniud.it

ABSTRACT

Results from three experiments are presented, showing that the perceived acoustic and vibrotactile quality of a reproduced piano does not require models simulating every aspect of the original instrument with great accuracy. It was found that high-quality loudspeaker array passive listening at the pianist's position admits distortion of the sound field. Furthermore, pianists during playing seem to compensate for errors in the auditory scene description. Finally, they are particularly sensitive to the existence of vibrotactile musical feedback on their fingers meanwhile tolerant about the precision with which this feedback is reproduced. Based on these results we are currently working on a lightweight portable physics-based digital piano design, that should improve upon the experience a pianist with no keyboards at hand makes when interacting with a touch-screen piano software running on smartphones and laptops.

1. INTRODUCTION

Musical instrument models often expose surprisingly high levels of accuracy to the performer. Since they are employed especially to enable digital implementations which reproduce the original instrument features only partially, one wonders whether the model can be integrally transferred and, thus, appreciated in the digital counterpart. Even more radically, one may wonder whether a performer has a neat perception of the original accuracy while playing the real instrument.

Aspects of the performer's acoustic sensitivity to an instrument quality have been put into question by carefully designed experiments [1]. Fritz and colleagues have revisited a common belief about the superiority of old Italian violins [2]. The accurate perception of a piano has been criticized since long [3], but perhaps not so many times. An insightful experiment was performed by Galembo and Askenfelt, who showed that a blindfolded group of expert pianists easily recognized three previously played different pianos by randomly performing over them, conversely they lost much of their own recognition ability when just listening to the same pianos [4]. Goebl and colleagues investigated on the long-debated question about the influence of touch to piano sound production [5]. Experiments like these suggest that a musical instrument sound designer should commit him or herself to uncompromised quality only after making sure that a model under study is perceptually worth that quality.

With specific regard to the piano, the question becomes more complicate once the real instrument is substituted by a system made of digital and electro-acoustic components. Irrespectively of their quality, such components in fact further bias the perception of the sound effects that are produced by a model. A reduced keyboard mechanics working in absence of hammers and strings is likely to influence the otherwise subtle cutaneous and haptic sensations to the fingertips, but to what extent do these sensations influence a performer's self-confidence with the instrument? The replacement of a soundboard with a loudspeaker set inevitably changes the acoustics of a piano, but do performers and listeners experience a measurable decay in the sound quality and localization?

In the following we summarize the results of four experiments that we have recently conducted on the piano, with the help of other researchers. Two such experiments were intended to understand the sensitivity and possible salience of cutaneous cues during playing. The remaining two aimed at understanding the perceptual consequences of corrupting the instrument's acoustic field pointing to the performer, in terms of perceived sound accuracy and localization. The results justify to test the quality of digital implementations whose distance from the real instrument is increased, trading off the resulting minor accuracy with improved portability and reduced costs. Currently we are working on an augmented table interface prototype which implements this design approach.

2. EXPERIMENTS

In general the experiments put the focus on the multimodal relationships existing between the performer and his or her instrument. Once such relationships are strengthened by years of practice and repeatable experiences, a change in some feedback modality should bring a comparable experiential novelty unless that change is imperceptible. We have investigated this and other facts by experimenting on the dependencies between the auditory and somatosensory experience, holding a visual scenario consisting of a real piano or alternatively a digital keyboard.

2.1. Perception of interactive vibrotactile feedback [6]

Keyboard makers have long since given empirical evidence of the importance of haptic cues in defining the quality of an instrument. First of all touch, mostly depending on the keys' material along with their dynamic response due to the connection mechanism with the strings, confers a unique haptic signature to a piano. Besides the mentioned Galembo experiment [4], the relationships between the perceived quality of a piano and the haptic signature of its keyboard have been understood only to a limited extent. It is generally acknowledged that the use of a simplified keyboard mechanics along with keys made of plastic material, such as those found in consumer digital pianos, inevitably translate to a less rewarding experience for the pianist. Yet, the subjective effects of an impoverished keyboard on the perceived sound quality have not been quantified to date.

Even less is known about if and how the same quality is influenced by vibrotactile feedback arriving at the pianist's fingers

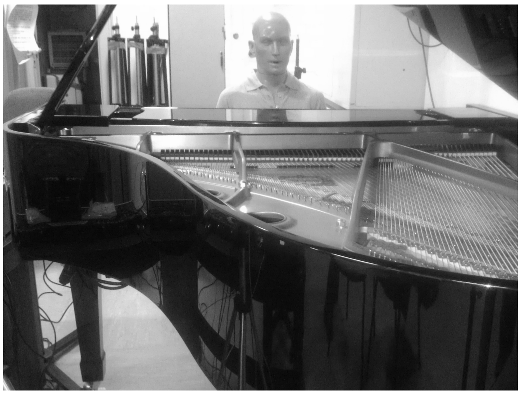

Figure 1: Setup for loudness estimation on the grand piano using a KEMAR mannequin.

once the more prominent somatosensory experience of striking the keys has ceased, leaving space to the vibrations traversing the instrument until the keys are released. In a related study [7], one of the present authors conducted a pilot experiment on a digital piano modified with the addition of vibrotactile feedback. On the other hand, while investigating the perception of vibrations on a grand piano, Askenfelt and Jansson provided quantitative evidence that even *ff* notes generate partial components whose magnitude hardly exceeds the known vibrotactile thresholds at the fingers [8]. Their measurements, hence, support the claim that neither a piano keyboard nor the keybed or the pedals should be able to convey prominent vibrotactile cues to the pianist.

We measured the pianists' sensitivity to piano key vibrations at the fingers while playing an upright or a grand Yamaha Disklavier piano. We took advantage of the switchable quiet mode to either provide vibrations or not in both pianos during playing. Subjects had to be prevented from hearing the Disklaviers; therefore, MIDI OUT data were used to control a Pianoteq software piano synthesizer that was configured to simulate a grand or an upright piano. The synthesized sound was provided by means of isolated headphones. The loudness of the acoustic pianos at the performer's ear was estimated by recording with a KEMAR mannequin all the A keys played at various velocities (Figure 1 shows the grand piano setup).

The test was a yes-no experiment. The task was to play a loud, long note (*mf* to *fff* dynamics, lasting 4 metronome beats at 60 BPM) and then to report whether vibrations were present or not. Only the A keys across the whole keyboard were considered, in this way reducing the experiment's duration while maximizing the investigated pitch range. A randomized sequence of 128 trials was provided, made up of 16 occurrences of each A key. Half of the trials were in the "vibration OFF" condition, corresponding to the Disklavier set to quiet mode. The total duration of the experiment was about 20 minutes per participant.

2.1.1. Summary of results and discussion

Proportions of correct responses, given by

$$p(c) = \frac{(\text{"yes"} \cap \text{vibes present}) + (\text{"no"} \cap \text{vibes absent})}{\text{total trials}},$$

were calculated for each participant individually for each A key. Average results for the upright and grand configurations are presented respectively in Figures 2 and 3, showing a similar trend. For the lowest three pitches (A0 to A2), the subjects could easily discriminate between the trials with and without vibrations.

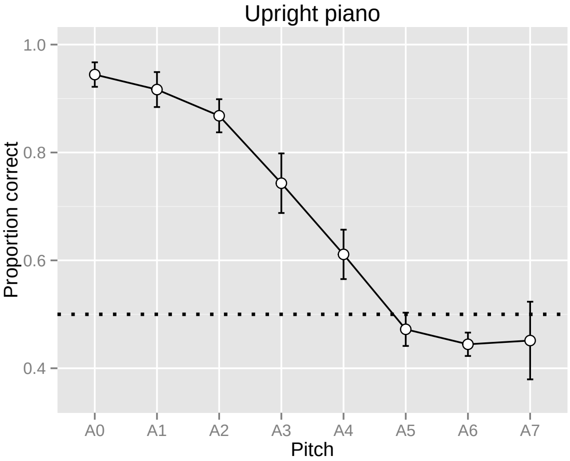

Figure 2: Mean proportions correct for the upright piano configuration. Chance performance given by dashed line. Error bars present within-subjects confidence intervals.

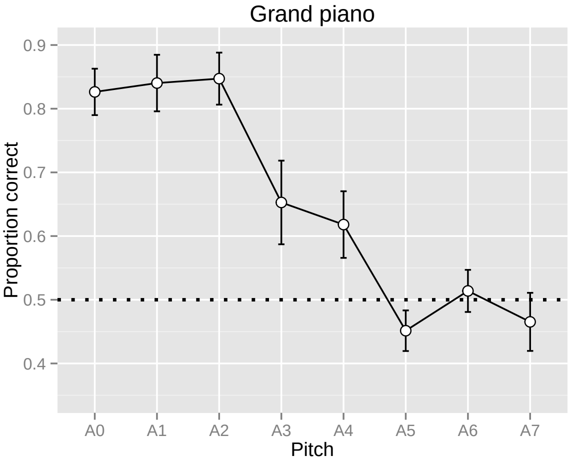

Figure 3: Mean proportions correct for the grand piano configuration. Chance performance given by dashed line. Error bars present within-subjects confidence intervals.

In the middle register the proportion of correct responses was still over 60%, while it finally dropped to chance level at A5 ($f_0 = 880$ Hz).

Our findings complement the Askenfelt and Jansson results [8] especially in the low range up to 110 Hz, where detection was clearly easier than in the range of highest sensitivity, where only two thirds of the subjects were successful at detecting key vibrations. This may be explained by the nature of the vibratory signal which was not sinusoidal, unlike in the threshold measurements by Verrillo. More in general the pianist is engaged in an enactive experience where every key depression produces a distinct audio-haptic contact event, immediately followed by the transmission of vibrotactile cues from the keyboard, caused by the vibrating strings and resonating body of the instrument. Such cues are subjected to disparate temporal, spatial and spectral summation or interference effects, depending on the sequence of played notes and chords, as well as on the position of the hands on the keyboard. For all such effects the literature provides only sparse data.

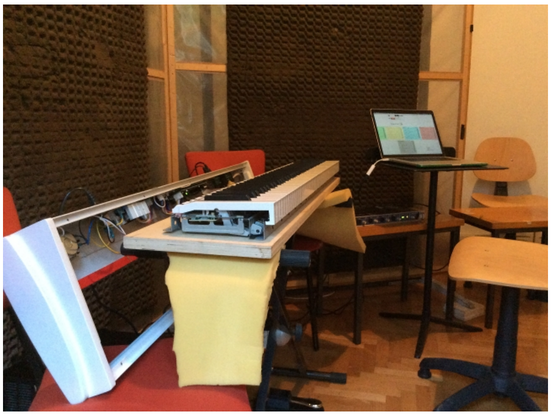

Figure 4: Experimental setup.

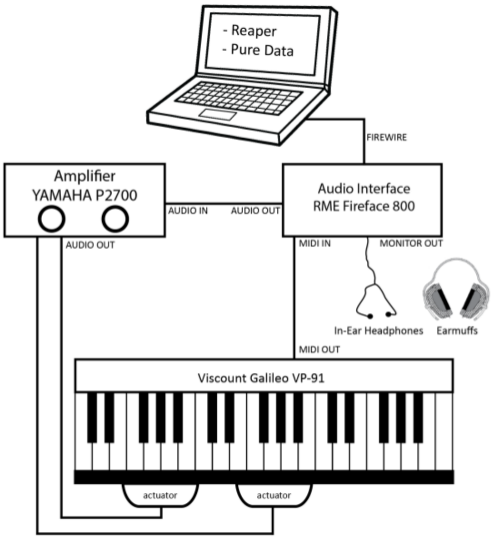

Figure 5: Schematic of the setup.

The effects of active pressing force on vibrotactile perception are not thoroughly known, but there is evidence that vibrotactile magnitude sensation increases under a passive static force. Promising results have been very recently obtained in this sense by Papetti and colleagues [9].

2.2. Interactive reproduction and subjective evaluation of real vs. synthetic vibrotactile cues [11]

On the light of the previous experiment, recently we have investigated on the quality of the vibrotactile feedback. In other words, we hypothesized that pianists appreciate the reproduction of real as opposed to simplified synthetic key vibrations. The experiment required to disassemble a digital piano keyboard, and instrument it so as to convey vibratory signals to the user (see Figure 4); then, to record key vibrations on an acoustic piano and to synthesize simplified counterparts, which were organized in two respective sample banks.

Audio-tactile stimuli were produced at runtime: the digital keyboard played by the participants sent MIDI messages to the computer, where a piano synthesizer plug-in generated the related sounds and, in parallel, a sampler plug-in played back the Disklavier grand piano vibration samples then processed by an amplitude & spectral equalizer plug-in (see Figure 5).

Subjects wore earphones and ear-muffs on top of them, in the same fashion as the mannequin did during a previously made loudness matching procedure. In this way they were not exposed to the sound coming by air conduction from the transducers, as a by-product of their vibration.

Three vibration conditions were assessed relative to a non-vibrating standard stimulus A:

B: recorded real vibrations;

C: recorded real vibrations with 9 dB boost;

D: synthetic vibrations.

Synthetic vibrations consisted of noise filtered around the fundamental note, possessing similar amplitude envelope across time. Sound feedback was generated by a Pianoteq piano synthesizer playing the same configuration as in the previous experiment. The task was to play freely on the digital keyboard and assess the playing experience on five attribute rating scales: Dynamic control, Richness, Engagement, Naturalness, and General preference. The dynamics and range of playing were not restricted in any way.

Subjects could switch freely among setups α and β: Setup α was always the non-vibrating standard, while setup β was one of the three vibration conditions (B, C, D). The rating of β was given in comparison to α. The presentation order of the conditions was randomized. Also, participants were not aware of what could actually change in the different setups, and in particular they did not know that sound feedback would not be altered. The free playing time was 10 minutes per couple of conditions (A, B), and participants were allowed to rate the five attributes at any time during the session by means of a point & click graphical user interface (GUI). In the end, each subject gave one rating in each attribute scale for each vibration condition.

2.2.1. *Summary of results and discussion*

Ratings were given on a continuous Comparison Category Rating scale (CCR), ranging from -3 to +3, which is widely used in subjective quality determination in communications technology. Inter-individual consistency was assessed for each attribute scale by computing the Lin concordance correlations ρ_c for each pair of subjects. The average ρ_c were 0.018 for general preference, 0.006 for dynamic control, -0.04 for richness, -0.02 for engagement, and -0.04 for naturalness. In all scales, a few subjects either agreed or disagreed almost completely and, due to this large variability, ρ_c was not significantly different from 0 for any of the scales ($t(54) < 0.77, p > 0.05$). The low concordance scores indicate a high degree of disagreement between subjects.

Responses were positively correlated between all attribute scales. The weakest correlation was observed between richness and dynamic control, (Spearman correlation $\rho_s = 0.18$), and the highest between general preference and engagement ($\rho_s = 0.75$). The partial correlations between general preference and the other attribute scales were as follows: $\rho_s = 0.39$ for dynamic control, $\rho_s = 0.72$ for richness, and $\rho_s = 0.57$ for naturalness.

Results are plotted in Figure 6, and the mean ratings for each scale and vibration condition are given in Table 1. On average, each of the vibrating modes was preferred to the non-vibrating standard, the only exception being condition D for Naturalness. For conditions B and C Naturalness received faintly positive scores. The strongest preferences were for Dynamic range and Engagement. General preference and Richness had very similar mean scores though somewhat lower than Engagement and Dynamic control. Generally, C was the most preferred

Vibration	Dyn.	Rich.	Eng.	Nat.	Pref.
B	0.92	0.30	0.50	0.26	0.24
C	1.28	0.67	1.21	0.17	0.81
D	0.87	0.42	1.00	-0.23	0.29

Table 1: Mean ratings over all subjects for each attribute and vibration condition.

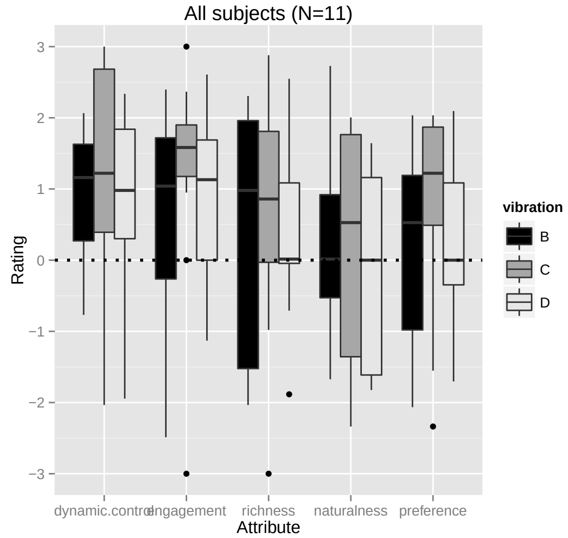

Figure 6: Results of the quality experiment. Boxplot presenting median and quartiles for each attribute scale and vibration condition.

of the vibration conditions: It scored highest on four of the five scales, although B was considered the most natural. Interesting enough, B scored lowest in all other scales.

Heterogeneity was observed in the data, as might be expected due to the high degree of variability in the inter-individual agreement scores ρ_c. A k-means clustering algorithm was used to divide the subjects *a posteriori* into two classes according to their opinion on General preference. Eight subjects were classified into a “positive” group and the remaining three into a “negative” group. The results of the respective groups are presented in Figure 7. A difference of opinion is evident: The median ratings for the “winning” setup C are nearly +2 in the positive group and -1.5 in the negative group for General preference. In the positive group, the median was > 0 in all cases except one (Naturalness, D), whereas in the negative group, the median was positive in only one case (Dynamic control, B).

We concluded that key vibrations increase the perceived quality of a digital piano. Although the recorded vibrations were perceived as the most natural, amplified natural vibrations were overall preferred and received highest scores on all other scales as well. The other interesting outcome is that the vibrating setup was considered inferior to the non-vibrating standard only in Naturalness for synthetic vibrations. This suggests that pianists are indeed sensitive to the match between the auditory and vibrotactile feedback.

The high degree of disagreement between subjects suggests that intra- and inter-individual consistency is an important issue in instrument evaluation experiments. Due to only one attribute rating per subject and condition, intra-individual consistency could not be assessed in the present study and will be left for a future revision. However, the heterogeneity in the data was similar across all attributes and conditions, making it hard to believe it was caused by inconsistency alone. Roughly two thirds of the subjects clearly preferred the vibrating setup, perhaps less rewarded by the synthetic vibrations, while the remaining one third had quite the opposite opinion. It is interesting that both the jazz pianists, having probably more experience of digital pianos than the classical pianists, were in the “negative” minority: would a vibrating digital keyboard be perceived as less pleasant than a neutral one, reflecting a preference of those pianists to the digital piano’s traditional tactile response?

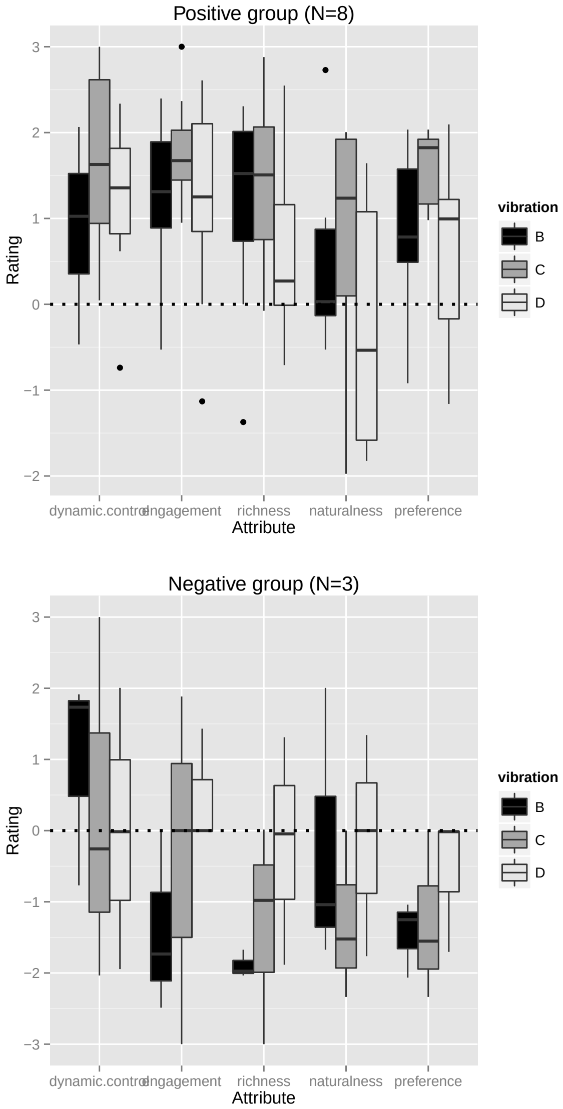

Figure 7: Quality results for the positive and negative groups.

2.3. Sensitivity to loudspeaker permutations [12]

Performers declare to be especially sensitive to changes in the sound coming from their instrument. On the other hand, the role and importance played by the auditory cues when a piano is perceived to sound different is not obvious. Recent literature marks the difference existing between playing as opposed to listening to a piano: such two activities would in fact lead the pianist to develop different impressions about the quality of the instrument [13, 5].

This research considered a collection of accurately recorded multi-channel piano notes, that were presented to a group of pi-

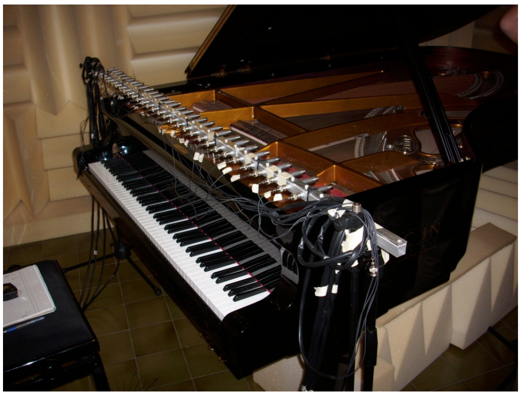

Figure 8: Recording session: microphone setup.

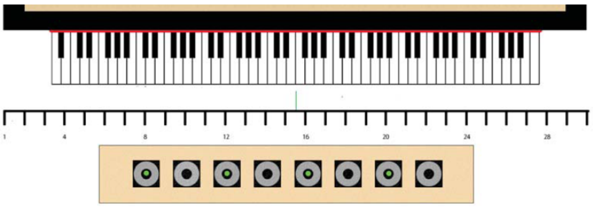

Figure 9: Microphone/loudspeaker alignment.

anists via a calibrated array of eight small loudspeakers. Distortions were introduced during the listening test by exchanging the output channels, and subjective impressions about the realism of the sound and the auditory scene were gathered along with the apparent listening position. Our analysis suggests that only the largest permutations, in a sense that will be defined later, cause significant corruption of both qualities furthermore without clear implications on the auditory scene description.

Six *mezzo forte* piano notes (C4, E4, C2, A4 major, D4, C5) were selected from a huge collection, result of a recording session made in July 2012 at the Viscount International SpA semi-anechoic room based in Mondaino (RN) - Italy, using a Seiler model 1849 piano that was tuned and prepared for the occasion, and then played by a professional pianist and sound designer consulting for the company. Such notes were collated together one after the other, hence forming a slow scale lasting about thirty seconds.

Fig. 8 illustrates the recording setup consisting of a linear array of 30 Bruel&Kjær model 4188 omnidirectional microphones, calibrated and made available by Angelo Farina's acoustics research group at the University of Parma, along with an M-Audio multi-channel sound interface. The array was positioned in such a way to capture the soundfield in front of the cover, which was left open. The reproduction was realized avoiding any signal processing, by just reporting eight equally-spaced recorded channels onto a single-pressure chamber linear array made with 2.5" Ciare loudspeaker units.

Fig. 9 shows the alignment between the microphone and the loudspeaker array, with the piano keyboard taken as reference: the eight loudspeakers, hence, reproduced the recorded channels no. 8, 10, 12, 14, 16, 18, 20 and 22, respectively. From here on we will associate such recorded channels respectively to the loudspeakers 1, 2, 3, 4, 5, 6, 7, 8, numbered left to right.

Ten reproduction patterns were prepared using the eight channels: two of them were formed respectively by quadruplicating two, and duplicating four recorded channels over the loudspeakers; the third one was left untouched; the remaining seven were obtained by permutations of the inputs. All patterns are listed in Table 2.3 below.

Pattern no.	Configuration	Label
1	11118888	Magnified stereophony
2	11336688	Magnified quadraphony
3	12345678	Original
4	21436587	Swapped adjacent ones
5	34127856	Swapped adjacent pairs
6	56781234	Swapped quadruples
7	73258146	Random no. 1
8	78345612	Swapped edge pairs
9	87654321	Reverse panning
10	51843276	Random no. 2

The experiment was set up in a silent, dry room (approximately 3 × 3 × 2.75 meters) having walls partially covered with damping foam. In addition to the active array, four loudspeakers were located each at one corner of the room, furthermore two additional eight-channel arrays were put in front of the listener: the presence of such idle systems added uncertainty in the listeners about the sources that were going to be used during the experiment.

Subjects had to sit on a chair at the center of the room, approximately one meter far from the loudspeaker array. While sitting, every subject was given a tabletop computer on which (s)he could respectively rate the *realism of the sound* R_S and the *realism of the auditory scene* R_A on a scale ranging 1 (poor) to 7 (excellent), as well as choose his or her own *relative position* R_P in the virtual scene among nine possible listening points, labeled A to I in alphabetical order. Before the test, the subject was given verbal instructions about the scale (s)he was going to listen to, as well as about the use of the graphical interface.

The test consisted of listening to a balanced random distribution of the patterns, each repeated five times for a total of fifty trials. During each trial, every subject listened to the musical scale and then rated R_S, R_A and R_P by selecting the corresponding value in the graphical interface; finally, (s)he submitted her or his selections by pushing a software button. After each submission a new trial was started: this procedure allowed in particular for rating a scale and go to the next one by submitting before the end of the current sound, or conversely to pause at the end of a trial by delaying the respective submission. In this way subjects could optimize the flow of the test, which took approximately 40 minutes to be completed.

2.3.1. Summary of results and discussion

Fig. 10 (above) plots, for each pattern in the respective box, the median of the corresponding rate R_S, the 25th and 75th percentiles with their extreme datapoints, the average values and the outliers. The same boxes are displayed for each pattern rated R_A, below in the figure. Both plots have been obtained using the `boxplot` function of Matlab.

An informal inspection suggests the existence of a significant decay in both sound and scene realism *only* when the patterns 7,8 and 10 are displayed. In all other cases the decay appears to be not significant.

2.4. Active sound localization

Similarly to the explanation given for motivating the stability of the perceived realism in presence of loudspeaker permutations in the previous array, independently of the existence of a

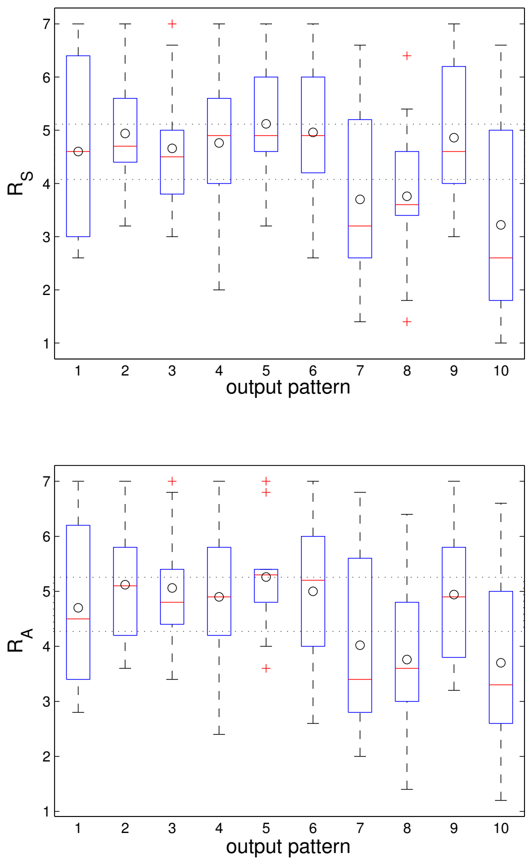

Figure 10: Boxplots showing median, 25th and 75th percentiles with their extreme datapoints, average values ('∘') and outliers ('+') for each output pattern rated R_S (above) and R_A (below), respectively. On either boxplot, the rectangle in dashed line gathers average values within the respective HSD range, edged by the largest average.

precedence effect we conjecture that the localization of a note during playing is locked to the corresponding key position by the somatosensory cues of hand proprioception. Holding this locking effect, then the same position can be robustly recalled during listening: this previously learnt process may suppress the auditory localization of the same note via lateralization cues, in particular resolving any potential incongruence between the proprioceptive and auditory information.

Here we illustrate some preliminary results of an experiment using the same stimuli as before, this time employing a 14 channel array of 2.5" Ciare loudspeaker units, in which we asked subjects to localize the direction of arrival of piano notes. Subjects either passively listened to the stimuli, or conversely they activated them by playing the corresponding key on the Galileo keyboard—see Sec. 2.2. The array channels were manipulated so to create also reversed, random, and monophonic sounds.

Figure 11 shows that subjects during active playing gave higher scores to manipulated sound fields. The significance of these scores is left to a future analysis, along with any further discussion on the support of the active playing task to the sound localization.

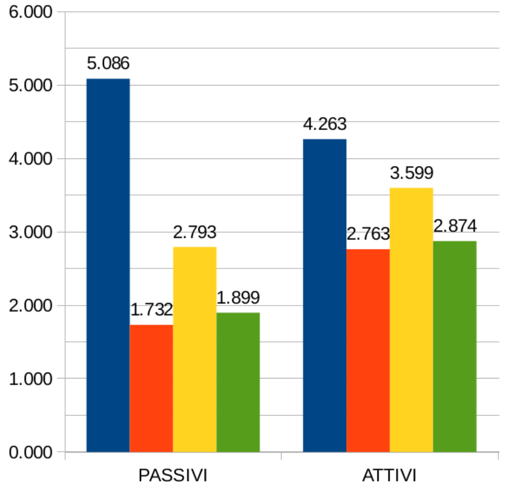

Figure 11: Histograms showing perceived degree localization for listening (left) vs. playing performers (right) respectively for normal (blue), reversed (red), random (yellow), and monophonic (green) array listening.

3. AUGMENTED TABLE INTERFACE

Recently there has been a lot of attention, both in the musical interface research community and in the industry, in developing keyboard-like interfaces which are highly portable and that can be the right companion for computing devices such as smartphones and tablet PCs. Such devices usually provide touch-screen interfaces, but they suffer from limitations in size, high latency and are in general not sensitive to force variations. Therefore, we focused on developing an interface that aims to augment any surface, e.g. a common table, into an immaterial digital keyboard, requiring no more space than the portable computing device to operate.

In this work we aimed at developing a prototype which can capture the musical gestures using mostly a common camera, following standard algorithms for finger and hand tracking [14]. In this way, the augmented interface could be built without any additional hardware making it the best solution in terms of portability. While there are already some solutions that work in a similar way, such as Augmented Piano by Amit Ishai and Moshe Liran Gannon [1], they all suffer from high failure rate in the detection algorithms and high latency.

Our primary goal was to provide a system which could detect precisely notes and velocities, working with a total latency lower than 30 milliseconds. Unfortunately, this requirements are very strict for a generic mobile computing platforms such as smartphones or tablets, mainly because of the extra latency imposed by either the video and audio processing pipelines of the underlying operative systems. Therefore, we centered our prototype on a Linux-based single-board PC, which is similar in terms of computing hardware to the aforementioned devices but gives us the flexibility that we need in terms of software development. Moreover, we were able in this way to experiment with enhancements that rely on external hardware, such as the use of piezoelectric microphones to improve detection latency or tactile transducers for haptic rendering on the surface. However, it is important to notice that the features provided by them are optional and the system is still functional just with video

[1] https://sites.google.com/site/pianoreality/

input from a camera.

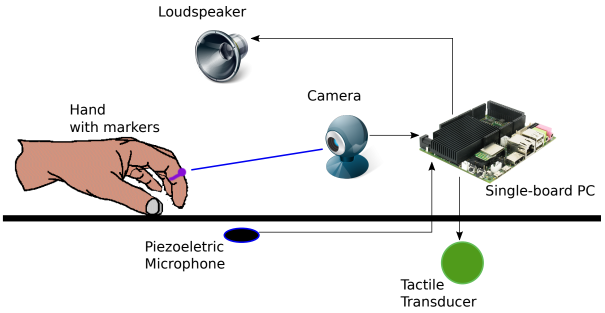

Figure 12: Test picture

3.1. Prototype Architecture

We chose to employ the Udoo single-board PC [2] as the computing core of our system. While maintaining a compact size, the board features a quad-core ARM Cortex-A9 processor clocked at 1GHz, plus an additional ARM Cortex-M3 processor compatible with the Arduino Due environment. The board also features an integrated codec for audio connections and a WiFi module which comes useful when integrating a device like a PC or smartphone for e.g. control with a graphical user interface. We tried several solutions for the camera and finally settled on the Playstation Eye Camera, a compact and cheap consumer device which is still capable of capturing video with a low latency and a high frame-per-second (FPS) rate (60FPS at 640x480 resolution).

The connection between the components of the system can be seen in Fig. 12. Besides the main board and the camera, there is an optional input in the form of a piezoelectric microphone placed on the table. The output hardware components are a generic loudspeaker or headphones, plus an optional tactile transducer driven by a separate DAC on the board controlled by the Cortex-M3 processor.

The system runs on top of a Linux system based on a highly customized Debian Wheezy distribution. The most difficult part has been the integration of the RealTime Linux Kernel patch [15], since the sources provided by the board manufacture are not aligned with the mainline Linux Kernel. With this modification we were able to reduce the audio latencies from 40ms to 4ms, thus resulting in a minimum impact on the overall latency of the system. The audio synthesis system is driven by the JACK audio server [16] and is composed by a sample-based synthesis engine [3], able to reproduce e.g. a high-quality piano soundfont with a polyphony of 256 notes. We implemented a dedicated service in the form of a Jack audio program to obtain the haptic output signal from the audio, by filtering and downsampling the output of the synthesizer. The result is then encoded and sent over a serial communication bus connected to the Cortex-M3 processor, where a separate piece of software, written using the Arduino language, is in charge of collecting the data and driving one of the additional DACs accessible on the board.

3.2. Gesture Recognition Algorithm

At its core, the system tracks the movement of the fingers using the video stream obtained from the camera. Using only this information, it is very hard to obtain latencies lower than 30-40ms, due to the low frame rate and buffering involved in video processing. Therefore, we included the option to use the data acquired by a contact microphone placed on the surface to improve the overall latency. The audio signal obtained in this way is analyzed using an onset detector calibrated to find the peaks at medium-high dynamics, thus enabling the triggering of Note ON events in advance compared to when the video-only algorithm would have detected them.

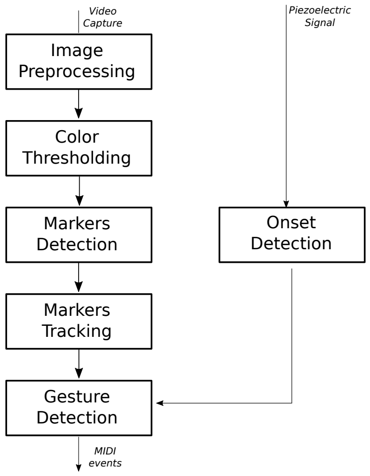

Figure 13: Data flow

We wrote a C++ program using the popular OpenCV library [17] for most of the vision-related tasks. The algorithm implemented requires coloured markers placed on rings that can be worn on the fingers, excluding the thumb, and a camera placed in a fixed position in front of the hand. Processing follows standard motion-tracking techniques [14] and is summarized by the dataflow diagram presented in Fig. 13.

During an initial one-time calibration phase, the colour of the tracking markers is analyzed finding a two-dimensional threshold on Hue and Saturation by looking at the histogram of a predefined region, using an interval having half-length of a standard variation σ for Hue and three times σ for Saturation, since Hue is a more reliable feature to track. A second calibration is performed to estimate the pose of a keyboard drawn on a sheet of paper, in order to have a map from camera coordinates to keyboard keys.

At run-time, each image is first preprocessed by applying a conversion from RGB to HSV color space. Then, a two-dimensional thresholding operation, using the limits derived from the calibration, is applied to obtain a binary image where the selected pixels correspond to the tracked colors. A sequence of dilate and erode filters is applied to the result so that false positives are minimized and noise inside the tracked area is reduced. A number of predefined (e.g. four) markers are then detected in the binary image using a standard algorithm by Suzuki et al. [18] which gives as a result a set of closed contours. For each contour, a single relevant and stable point is extracted by using features based on the momentum of the detected area such as the barycenter.

The detected video coordinates are then fed to a tracking algorithm based on a Kalman filter, where we used a dynamic system of dimension four, with two states assigned to position and two to velocity. The usage of this algorithm reduces significantly detection artifacts that can result in sudden jumps of the

[2] `http://www.udoo.org/`

[3] `http://www.fluidsynth.org/`

tracked position, which can severely impact gesture and velocity detection in the following phase. At the same time, however, the filtering process might add some latency, so the parameters of the Kalman filter have to be calibrated in order to get a good compromise. Finally, the tracked positions are analyzed at each frame for detecting musical relevant gestures, which are then converted to MIDI events such as Note ON/Note OFF or continuous controls. Note triggering works using two virtual contacts placed vertically above the table and measuring the time between the crossing of the two contacts. After an event is triggered, the horizontal coordinates are mapped from camera to keyboard space using the transformation obtained from the calibration step, so the exact key pressed can be detected.

If the piezoelectric audio is taken into account, the trigger of the events is formed by the logical or of the two detection systems, with some extra logic that takes care of not detecting a single event more than one time. Typically, the audio-based detection works with medium to high dynamics, while only the video is used for lower dynamics, which luckily do not suffer much from higher latency. Overall, detection latency using the audio trigger is around 10ms, of which 4ms are fixed and due to the audio buffer size and the rest vary depending on dynamics of the hit and type of the surface. Video-only latency is harder to estimate but we judged it to be around 35-40ms from empirical comparison with other systems.

4. ACKNOWLEDGMENTS

Alberto Amendola, Federico Avanzini, Yuri De Pra, Hanna Järveläinen, Giorgio Klauer, and Stefano Papetti have collaborated to the mentioned experiments. The authors acknowledge support from the Proof of Concept Network national project entitled ”Virtual piano system on a tablet pc” (”Sistema di pianoforte virtuale su tablet pc”) called by the AREA Science Park in Trieste, participated by the University of Udine and Julia SRL, and co-financed by Viscount SpA through the University of Verona, all from Italy.

5. REFERENCES

[1] D Murray Campbell, “Evaluating musical instruments,” *Physics Today*, vol. 67, no. 4, pp. 35–40, 2014.

[2] Claudia Fritz, Joseph Curtin, Jacques Poitevineau, Palmer Morrel-Samuels, and Fan-Chia Tao, “Player preferences among new and old violins,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 3, pp. 760–763, 2012.

[3] Mary Cochran, “Insensitiveness to tone quality,” *The Australasian Journal of Psychology and Philosophy*, vol. 9, no. 2, pp. 131–133, 1931.

[4] A. Galembo and A. Askenfelt, “Quality assessment of musical instruments - Effects of multimodality,” in *5th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM5)*, Hannover, Germany, Sep. 8-13 2003.

[5] Werner Goebl, Roberto Bresin, and Alexander Galembo, “Once again: The perception of piano touch and tone. Can touch audibly change piano sound independently of intensity?,” in *Proceedings of the International Symposium on Musical Acoustics (ISMA2004)*, Nara, Japan, Mar. 31 – Apr. 4 2004, pp. 332–335.

[6] F. Fontana, F. Avanzini, H. Järveläinen, S. Papetti, F. Zanini, and V. Zanini, “Perception of interactive vibrotactile cues on the acoustic grand and upright piano,” in *Proc. Joint ICMC/SMC Conf.*, 2014.

[7] F. Fontana, S. Papetti, V. dal Bello, M. Civolani, and B. Bank, “An exploration on the influence of vibrotactile cues during digital piano playing,” in *Proc. 8th Sound and Music Computing Conference (SMC2011)*, Padua, Italy, July 6-9 2011, pp. 273–278, Padova University Press, Available at `http://www.padovauniversitypress.it/`.

[8] A. Askenfelt and E. V. Jansson, “On vibration and finger touch in stringed instrument playing,” *Music Perception*, vol. 9, no. 3, pp. 311–350, 1992.

[9] S. Papetti, H. Järveläinen, and G.-M. Schmid, “Vibrotactile sensitivity in active finger pressing,” in *World Haptics Conf.*, 2015.

[10] Andrew M. Galica, Hyun Gu Kang, Attila A. Priplata, Susan E. DAndrea, Olga V. Starobinets, Farzaneh A. Sorond, L. Adrienne Cupples, and Lewis A. Lipsitz, “Subsensory vibrations to the feet reduce gait variability in elderly fallers,” *Gait & Posture*, vol. 30, no. 3, pp. 383 – 387, 2009.

[11] F. Fontana, F. Avanzini, H. Järveläinen, S. Papetti, G. Klauer, and L. Malavolta, “Interactive reproduction and subjective evaluation of real vs. synthetic vibrotactile cues on a digital piano keyboard,” in *Proc. SMC Conf.*, 2015, To appear.

[12] F. Fontana, Y. De Pra, and A. Amendola, “Sensitivity to loudspeaker permutations during an eight-channel array reproduction of piano notes,” in *Proc. SMAC/SMC 2013*, Stockholm, Sweden, Jul. 30 - Aug. 3 2013.

[13] Anders Askenfelt, Alexander Galembo, and Lola L. Cuddy, “On the acoustics and psychology of piano touch and tone,” *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2873, 1998.

[14] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly, “Vision-based hand pose estimation: A review,” *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 52–73, 2007.

[15] Steven Rostedt and Darren V Hart, “Internals of the rt patch,” in *Proceedings of the 2007 Linux symposium*. Citeseer, 2007, vol. 2, pp. 161–172.

[16] Paul Davis and T Hohn, “Jack audio connection kit,” in *Proceedings of the 1st Linux Audio Developer Conference*. Institute for Music and Acoustics, ZKM, Karlsruhe, Germany, 2003.

[17] Gary Bradski et al., “The opencv library,” *Doctor Dobbs Journal*, vol. 25, no. 11, pp. 120–126, 2000.

[18] Satoshi Suzuki et al., “Topological structural analysis of digitized binary images by border following,” *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.