A prototype for pitched gestural sonification of surfaces using two contact microphones

Alberto Novello
Royal Conservatory Den Haag
Juliana van Stolberglaan, 1
2595CA Den Haag,
The Netherlands
albynovello@gmail.com

Antony Raijekoff
University for Applied Science
Fachhochschule St. Pölten GmbH
Matthias Corvinus-Straße, 15
A-3100 St. Pölten, Austria
lbrayzhekov@fhstp.ac.at

ABSTRACT

We present the prototype of a hybrid instrument, which uses two contact microphones to sonify the gestures of a player on a generic surface, while a gesture localization algorithm controls the pitch of the sonified output depending on the position of the gestures. To achieve the gesture localization we use a novel approach combining attack parametrization and template matching across the two microphone channels. With this method we can correctly localize 80 ± 9 % of the percussive gestures. The user can assign determined pitches to specific positions and change the pitch palette in real time.

The tactile feedback characteristic of every surface opens a set of new playing strategies and possibilities specific to any chosen object. The advantages of such a system are the affordable production, flexibility of concert location, object-specific musical instruments, portability, and easy setup.

Author Keywords

Tap tracking, gesture localization, object sonification, machine learning.

1. INTRODUCTION

In the last years, an increasing amount of articles and research projects have concentrated on using everyday objects as tools to control or generate sound [2, 7, 13]. The acoustic characteristics of objects can be recorded in real-time through inexpensive sensors, analyzed by feature extraction techniques, and mapped to synthetic sound parameters. This approach creates hybrid sounds that are at the same time organic and modifiable by the computer. Using mainly contact microphones, the advantage of such approach resides in the low cost of the technology, ease of transportation and setup, and flexibility of concert location. An important aspect is also the aesthetics that such a system imposes on a performance. By using objects, with which the audience is familiar, as musical instruments, the performance extends its expressivity and connection to the public. There are already examples of electronic musicians "playing" bicycle frames, or using Lego blocks as controllers [7, 13].

The "Mogees", acronym for mosaicing gestural surface, by Bruno Zamborlin et al. [13], uses a contact microphone, connected to an accompanying software (that can be hosted on a smart phone) to drive the sound synthesis engine. The contact microphone converts the acoustic vibrations from the object into an electrical signal, which is analyzed to determine the user's gestures, i.e. a tap or a scratch each produce a different sound output. In a physical modeling analogy, whatever the sensor picks up (e.g., the gestures and the resonances of the object itself) becomes the exciter, while the resonator is emulated through the software. The sound is generated using an audio synthesis technique inspired by physical modeling, in order to create a tight coupling between the gesture and the sonic output. Mogees have been extensively used in live situations and several artists have endorsed their use in their live set.

The "Touch and Activate" by Makoto Ono et al. [7] classifies the acoustic resonances of surfaces to add interactivity to everyday objects. The authors scan the acoustic properties of the object through a transducer applied on it. A chirp signal excites the object while a contact microphone records its vibratory modes, capturing an acoustic imprint. Touching the object alters its resonance, creating a different imprint. The authors can store a pattern for every touching modality, and recognize them in a later moment by using machine learning in the form of support vector machine classification. For this system, the user needs to always have the transducer and the microphone attached to the surface. The system is thus slightly more complex than the Mogees, which requires only one transducer. Being based on classification, Touch and Activate is by its nature suited for the detection of specific patterns to trigger software cues (such as change of scene, change volume, trigger sample playback), while the Mogees is intended in the direction of sound-augmented objects.

Bisby et al. [2], propose a system that embeds both previous ideas of gestural localization and sonification. The authors estimate the location of the gesture from the different time of arrival of a gesture recorded by six contact microphones on a surface. Despite the good localization results (about 99%), this system not only requires a higher number of microphones compared to the previous systems but also a surface with uniform density and good resonance isolation.

Inspired by the previous methods, we found it challenging

to embed the previous approaches into one instrument and add the possibility for the player to control the pitch of the sound with the position of his gestures on the surface. In this article, we present a prototype of a system for pitched sonification of a surface using two common contact microphones. The software estimates the pitch for every event using a gesture localization algorithm. The timbre and decay of the pitched gestural synthesis is influenced by the acoustic characteristics of the chosen surface and material. The system requires relatively cheap hardware, it is adaptive to different surfaces, and selects the pitch from a user-defined pitch palette allowing the creation of melodies from every object.

2. GESTURE LOCALIZATION

Localization of gesture on surfaces is a topic that has been approached in literature from different disciplines, such as signal processing, machine learning, and video tracking [1, 4, 5, 8, 9, 10, 14]. Several processes make the estimation of the transfer function of sound in solids more difficult than in the case of air. The speed of sound is variable and depends on the properties of the material through which the wave is traveling. The speed of transverse waves depends on the shear deformation under shear stress and the density of the medium. Longitudinal waves in solids depend on the same two factors with the addition of a dependence on compressibility. The different reflections and varying homogeneity of solid objects make the prediction of the transfer function of a gesture computationally difficult to determine. This problem becomes increasingly difficult considering the different gesture typologies, intensities of attack, and nonlinearities of each material [6, 11, 12].

Several papers overcome these physical and computational complexities by looking at cross correlation between different microphone channels in microphone arrays [1, 11]. By estimating the "similarity" of output, an algorithm can theoretically determine the time of attack and the delay of two signals, without need to understand the physical characteristics of the surface or its exciter. The use of cross correlation in this case requires high sample rate to determine the location of impact with precision.

In Table 1, we report the speed of sound in different common materials and the theoretically required sample precision to discriminate tap positions on a 1-meter surface for a chromatic scale. Even with high sample rates of 88.2 and 192.4 KHz, approaches using cross correlation in our tests lead to unsatisfying results. We ascribe these errors to the sound reflections and cancellations of the surface, which distort the waveform through the medium. The peak of every gesture is deformed, resulting in the cross correlation misalignment of the channels. Different articles report a better prediction obtained using the GCC-PHAT algorithm for cross correlation, which weighs the final waveform to reduce the effect of echo cancellations [11]. However, in our tests this method also lead to poor delay determination.

Table 1. Speed of sound in different materials and number of samples required in discriminating one key (on a chromatic scale spread over a surface 1 meter)

	Sound Speed (m/s)	# samples for 1 key (88Khz)	# samples for 1 key (192Khz)
Air	343	23	46
Glass	3962	2	4
Wood	3960	2	4
Iron	5130	1	3
Plastic	3940	2	4

A second complication is the non-uniformity of most surfaces. Materials contain non-linear homogeneity causing filtering and diffraction of sound waves in the medium, resulting in frequency-dependent energy losses and phase distortions. Several methods have been proposed to solve the aforementioned complications using classification algorithms [1, 4, 5, 8, 9, 10]. Treating each surface as a black box, the system uses artificial intelligence methods to store different patterns for specific tapping locations (training phase). In a latter stage, the algorithms compare the incoming signal to match the patterns and determine the gesture location (playing phase). From our implementation of the aforementioned methods, we observed, however, high dependence of the system's performance on the type of exciter (e.g., a metal coin or a finger generate different reactions of the material, thus different patterns) and its intensity of impact. A system using machine-learning approaches thus requires recalibration for different percussion typologies.

3. PARAMETRIZED ATTACK DETECTION

In this paper, we use a novel approach with a combination of methods to estimate the position of percussive events on the surface. We first estimate the attack-time difference of excitation between channels with a two-stage low-latency gate. In every channel the gate opens when it detects a sudden change of amplitude that indicates the presence of a hit on the surface. A peak detector first smooths the signal to eliminate fluctuations due to noise and cancellations. It then estimates the position of the first peak in the selected audio sub-frame by looking at the maximum of the smoothed curve. A second threshold determines the signal-smoothing factor, to check whether there are secondary maxima in the following samples. This information is used in a pattern-matching algorithm with a predefined decaying envelope template to adjust the gate threshold and ignore the consequent peaks of the signal belonging to the same gesture. In this way we disregard oscillations from echo cancellations in the material, and we can better locate the first peak relative to the hit. We perform these operations independently on both channels. We finally compute the attack difference in time between the two channels to

estimate the best pitch candidate.

4. EVALUATION

We evaluated the performance of our algorithm on estimating the tap locations of a trained percussionist on a surface. The percussionist was requested to maintain constant velocity across his hits. The surface used was 1.20 m long and was divided in 12 equidistant keys, to represent the chromatic scale over one octave. In this way every key was 10 cm long, which is also the sensibility of our evaluation procedure. We asked the performer to tap 10 permutated sequences of the 12 notes of the chromatic scale on the surface (120 hits in total). This was done to equally use every location with different order. The performance was estimated as number of correctly identified locations out of 120 hits.

We repeated these measurements using two sampling frequencies: 88KHz and 96KHz), two percussion typologies: finger tap (smooth attack) and drumstick (sharp attack) and three velocities (piano, medium and forte). If we convert the ratio of correctly identified hits in percentage, we achieve on average a successful estimation of about 90 ± 8 % hits in the case of the stick, and 80 ± 9 % in the case of a finger tap. Considering the velocities, the medium and forte conditions give statistically better results than the piano condition, especially in the finger case. These results were replicated and hold true for surfaces of different resonating materials (glass, plastic and wood) and different dimensions (between 0.75 and 2m).

We noticed in these experiments that the system is rather sensitive to feedback. For instance, in a concert situation with strong amplification the audio might enter in resonance with the material. Another limitation concerns frame size used for the attack detection: larger frames make the software estimates more reliable. However this element affects the size of the pitch envelope. With large frames and fast playing the pitch envelope might receive updates too slowly. The window frame for attack processing still needs to be adapted manually in the case of rapid events.

5. GENERAL FUNCTIONALITY

For this article we used two AKG C411 III contact microphones through a MOTU Ultralight mk3 audio interface at a sample rate of 88KHz. The software is written completely in Max/MSP environment. We use the *gen~* object with *codebox* for the DSP calculations to achieve estimations at sample precision. Our calculations require one audio frame, usually consisting of 256 or 512 samples, which introduces acceptable latency for most common playing styles. The system was tested also using several types of contact microphones and provided consistent results.

5.1 Signal flow

The signal flow of our system is described in Figure 1. The incoming signals of the two channels are first passed through an adaptive FFT filter to remove the background noise, and calibrated in level through a compressor/gate system. The signal is then processed through the gesture classifier using the MUBU library by IRCAM [14], and recognized as one of the trained gestures, i.e. tap, scratch, knock, etc. If the gesture is classified as a fast attack, the tap detection algorithm extracts the time delay between channels, as mentioned in the above section, and estimates the position of the impact on the surface and sends a value to the pitch estimator. The position is converted into a pitch value from the set of pitches assigned by the user to every tap position on the surface and sent to the synthesis engine.

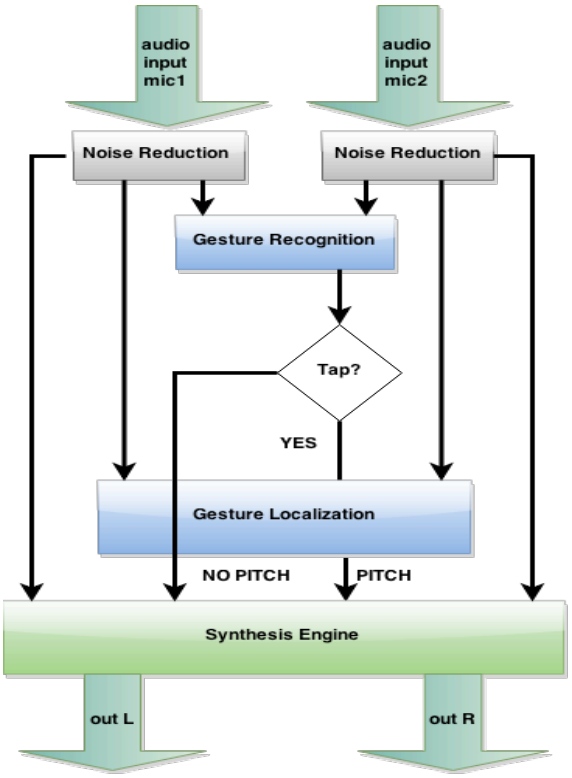

Figure 1. Signal flow for the described system.

5.2 Setup

After positioning the contact microphones on the surface, the software performs a calibration in three stages, each requiring about one minute. In the first calibration stage, the user records the incoming signal in absence of gestures. The system trains the FFT noise removal filter and adjusts the amplitude threshold just above the remaining noise level.

The second calibration stage involves the gesture-recognition engine. The user records a sample of the sound produced by few different gestures on the surface or different exciters. The recordings are analyzed through feature extractors to determine the relevant sound characteristics for later pattern matching. In this stage the user can select which synthesis engine to use according to the gesture typology and materials employed to play the surface.

In the last calibration stage, the system measures the

maximum time delay across microphone channels of the percussive events. For this task the user is required to tap on different positions with different intensities, very close to the first and second contact microphone and in the approximate center of both. This phase is used to furthermore extract a rough characteristic of how much energy is dispersed to every extreme.

After the three calibrations, the user can choose the number of desired pitches. The larger the number of pitches, the more precise the pitch detector should be, and consequently the higher the possibility of false detections in the system. The user can also customize which pitches to associate to a particular position. For example the system can use a pentatonic scale in the usual ascending pitch order, or modify an existing scale, e.g. reversing it, or swapping pitches. All presets can be changed in real time to quickly adapt to varying playing strategies.

In a future development of the system, we intend to use the tracked position of the gesture not only to determine pitch but also to send cue signals, such as sample playback, or change of synthesis engine, as in the case of Touch and Activate [7]. The pitch values, and an estimation of the sound envelope for velocity, can also be sent as MIDI or OSC messages for further processing. In this sense, our system can be used to augment existing surfaces adding functional buttons in a form of interactive objects or piano keys.

6. CONCLUSIONS

We presented a prototype of a hybrid instrument using two contact microphones to sonify the gestures of a player on a generic surface. A gesture localization algorithm using the information from the two contact microphones allows the player to choose the pitch of each percussive event depending on the tapping position on the surface. The user can customize the pitch palette in real time. The advantages of such reduced and hardware choice are the affordable production price, customization of object-specific musical instruments, portability, as well as easy setup. The system's design encourages experimentation using different types of exciting objects/surfaces. While the haptic feedback, specific to every material and surface, forces the player to develop new different playing strategies to new combinations.

Gesture localization using only two contact microphones is still a challenging problem. With the current calibrated system we reach correct detections of about 80 ± 9% of the percussive events depending on the exciter used. Compared to machine learning methods, our algorithm performs 10% worse on average. However, allowing a reasonable degree of recognition, even with a new untrained exciter or different surface, we found it to have the best compromise between robustness in detection and flexibility for improvisation. Because most false detections occur with neighboring positions, we decided to assign consonant pitches in adjacent positions, or use scales that have no hard internal interval dissonance, e.g. pentatonic scales.

Furthermore the use of cross synthesis, that uses the acoustic characteristics of the object, requires the players to choose their objects carefully for their sound, as every resonance of the object will affect the final sound result. This aspect brings the attention back from the computer to the sound of the object itself, similar to the ideas of musique concrete. In this sense, rather than the object being "just" a controller or a sound generator, it is both.

7. REFERENCES

[1] Bouzid, O. M., Tian, G. Y., Neasham, J., Sharif, B., Envelope and Wavelet Transform for Sound Localisation at Low Samplig Rate in Wireless Sensor Networks, *Journal of Sensors,* Volume 2012, Article ID 680383, (2012).

[2] Bisby, H., Cooper, A., James, S., Robertson, A., Ng, K., A Tactile Visual Instrument using Sound Source Localisation, *Electronic Visualisation and the Arts* (2014).

[3] Crevoisier, A., Future-instruments.net: Towards the Creation of Hybrid Electropnic-Acoustic Musical Instruments, (2008).

[4] Holm, S., Holm, R., Hovind, O. B., System and Method for Position Determination of Objects, *US patent* 7,535,796, 19 (2009).

[5] Lenz, C., Localization of Sound Sources, Studies on Mechatronics, PhD Thesis, Autonomous Systems Lab, *Swiss Federal Institute of Technology Press* (Spring 2009).

[6] Nazarov, V., Radostin, A., Nonlinear Acoustic Waves in Micro-inhomogeneous Solids, *Wiley* (2015).

[7] Ono, M., Shizuki, B., Tanaka., J., Touch & Activate: Adding Interactivity to Existing Objects using Active Acoustic Sensing, *UIST '13, ACM* (2014).

[8] Paradiso, J., King Leo, C., Checka, N., Hsiao, K., A simple System for Determining the Position and Characteristics of Knocks on a Large Sheets of Glass, *MIT Media Laboratory Cambridge* (2000).

[9] Paradiso, J., King Leo., C., Tracking and Characterizing Atop Large Interactive Displays, *MIT Media Laboratory Cambridge* (2005).

[10] Poletkin, K., Yap, X., and Khong, A Touch Sensitive Interface exploiting the use of Vibration Theories and Infinite Response Filter Modeling Base Localization Algorithm, *IEEE International Conference of Multimedia and Expo* (ICME 2010).

[11] Rui, Y., Florencio, D., Time Delay in The Presence of Coorrelated Noise and Reverberation, *Microsoft Research* (2004).

[12] Warren, P., Mason, J., Physical Acoustics and the Properties of Solids, *Journal of Acoust. Soc. Am.* 28, 1197 (1956).

[13] Zamborlin, B., Mogees, Gesture Recognition with Contact Microphones, **http://mogees.co.uk** (retrieved 11 January 2014).

[14] Schnell, N., Röbel, A., Schwarz, D., Peeters, G., Borghesi, R., MuBu & Friends - Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP. *International Computer Music Conference (2009)*.

[15] Montag, M., Sullivan., S., Dickey, S., Leider, C., A Low-Cost, Low -Latency Multi-Touch Table with Haptic Feedback for Musical Applications. *Proceedings of New Interfaces for Musical Expression* (2011).