2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Pluck and Play: Self-supervised Exploration of Chordophones for Robotic Playing

Michael Görner*[1], Norman Hendrich[1], Jianwei Zhang[1]

***Abstract*— Existing robotic musicians utilize detailed hand-crafted instrument models to generate or learn policies for playing because model-free or inaccurate policy rollouts might easily damage or wear out fragile instruments. We introduce an approach to characterize geometric models of chordophones and their audio onset responses directly through audio-tactile exploration with a physical robot arm. Initially, the system refines prior estimates of string positions, provided by kinesthetic teaching or visual estimation, through repeated attempts to pluck individual strings. A subsequent stage implements a *Safe Active Exploration* paradigm based on Gaussian Processes to explore and characterize the audio onset response of feasible plucking motions while minimizing invalid attempts. The resulting models can be used to actuate an imprecise robotic arm to play sequences of notes with varying loudness on a Chinese Guzheng.**

I. Introduction

Numerous robotic setups were demonstrated playing musical instruments, such as piano, guitar, and drums [1]. Approaches for this tend to focus on the higher-level programming/composing layer, whereas the actual actuation happens through precise servo motors or custom-built special-purpose hardware [2]. Other approaches instrument musical instruments to enhance human playing or track the system's physical state [3]. The emphasis is often put on intelligent musical behavior with the declared vision for competent robotic musicians to interact with others using musical theory as a common ground [1], [4].

Stringed instruments (or "chordophones") are particularly challenging for robots, as they afford many different and expressive playing techniques, while fully pre-programmed motions are unrealistic, as discussed below. Therefore, our work focuses on combining an online geometric reconstruction method with an active learning approach for motion generation to play unmodified original instruments with substantial dynamic range.

We demonstrate our approach with a Guzheng (古箏) shown in figure 1, a traditional Chinese chordophone from the family of zithers. The predominant variant features 21 strings in pentatonic tuning spaced approximately 1.5 cm apart. In human playing, the strings are plucked through up to four plectra taped to all except the little finger of each hand [5, 6]. Figure 2 illustrates the mounted setup on the anthropomorphic hand of our robot.

Unlike most other chordophones, which feature strings running in parallel and equidistant on a plane, the Guzheng features a more complicated geometry. The strings are mounted on a wooden body with a doubly curved surface and run over bridges at different distances from the instrument head, changing the height and angle of each string. As the bridges are mobile and moved during tuning, the strings are not positioned identically on different instruments—or different setups of the same instrument—and vary slightly. Different artists also prefer different levels of string tension, which again affects the placement of the bridges [7].

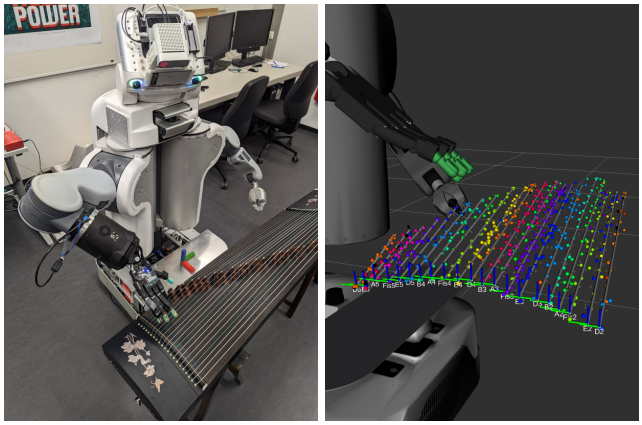

Fig. 1. (left) Modified PR2 robot with Shadow Dexterous Hand and SynTouch BioTac fingertip sensors in front of a Chinese Guzheng. (right) Projected audio note onsets (spheres) and reconstructed string representation (lines and coordinate systems) generated by self-supervised geometry exploration. Colors encode the musical note recognized from the audio channel. Note the curved and slanted geometry of the instrument.

In contrast to industrial robot arms, which provide high precision end-effector control (typically $< 1\,\mathrm{mm}$), the PR2 robot we use (figure 1 left) features spring-compensated shoulder joints and belt transmissions behind the position sensors. This kinematic structure reduces the force necessary to move the joints but impedes precise control. We regularly observe significant end-effector path errors above $5\,\mathrm{mm}$ depending on motion direction and involved joints.

Overall, this work contributes the following aspects:

- We establish active exploration strategies to estimate the position and audio response of chordophone strings based on audio, tactile, and proprioceptive feedback.
- We demonstrate a local model formulation using motion primitives to pluck strings. It is readily usable with Gaussian Process inference and implicitly accounts for controller behavior, geometric environment, and physical string characteristics.
- We utilize the explored models to play intended notes with a specified loudness.

This work was supported by the DFG Transregional Research Centre CML, TRR-169.

* corresponding author goerner@informatik.uni-hamburg.de
[1] with Department of Informatics, Universität Hamburg, Germany.

Fig. 2. Close-up view of the Shadow Hand with SynTouch BioTac sensors (turquoise), prepared with different plectrum mounts for Guzheng playing. From thumb to little finger: The thumb is taped with a specially curved Chinese thumb plectrum; two fingers with a regular Chinese plectrum taped to the fingertips as in human playing, experimental 3D-printed plectrum held in place with a thimble fitted to the finger, little finger without a plectrum, as traditional in Guzheng playing.

II. Related Work

Robot Musicians: A short history of robotic musical playing, covering both robots playing standard musical instruments as well as novel instruments explicitly designed for robotic actuation, is presented in [8], Similarly, Sobh and Wang discuss the development of a full band of robot musicians that perform on real musical instruments [9].

Several works have concentrated on piano playing. Zhang et al. describe both the "robotic" and "musical" playing of (entry-level) piano pieces with the highly complex Anatomically Correct Testbed hand [10]. A much simpler robot, designed to mimic the soft skeleton and elastic properties of a human finger, was presented recently in [11]. A two-arm robot system with bamboo hammers to play a Chinese dulcimer is described in [12].

Guzheng: Despite the important role of the Guzheng in Chinese musical tradition, only comparatively few scientific studies of the instrument and its playing style have been published. A vibroacoustic study of the Guzheng soundboard was presented in [13], while [14] presents a detailed modal sound synthesis model of the Guzheng. A recent paper [15] describes an optical sensor setup to track string plucking and vibration on the Guzheng. The harp is generally considered the closest Western-style equivalent to the Guzheng. The properties of a harp-plucking robotic finger were studied in [16], with detailed motion and vibration analysis as a function of fingertip material, shape, and reference trajectory.

Expressive Play: The goal of "musical" playing has long been to invoke a large variety of emotions and impressions through the multifarious use of playing techniques. This work focuses on exploring the dynamic range of plucking strings, but expressive play also encompasses various other techniques such as damping and modulation of notes [2].

Audio Analysis: In the context of our work, algorithms are needed for automatic and robust detection of note onset and note pitch [17, 18]. As with other stringed musical instruments, this is complicated by the transient vibration patterns of the instrument until a stable vibration with distinct spectrum is reached [19, 20]. Wang and Cao [21] present an algorithm to match Guzheng audio data with a musical score based on the dynamic time warping algorithm.

Deep Learning: In addition to those earlier analytical and physics-based works, recent interest has focused increasingly on deep-learning-based approaches. Recently, Byambatsogt et al. presented a deep neural network for automatic guitar chord detection and classification [22]. The network was first trained on a human-played but highly biased dataset [23] and then trained on a larger augmented robotic dataset including all possible chords. In the context of computer graphics and virtual reality, a recent paper uses GANs to animate a Guzheng-playing avatar based on audio input [24].

Reinforcement Learning: Xu et al. [25] proposed a Bullet-based simulation setup with an Allegro hand model for learning to play short sequences of notes on a piano from music notation with fingering. The trained agents can control the model in a four-dimensional action space to press multiple keys with specific attack velocities from online input. Most recently, Zakka et al. [26] presented a MuJoCo-based piano simulation with two Shadow Dexterous Hand models. Agents need to be trained on individual music scores with fingering but demonstrate impressive advanced skills in playing well-known Western piano compositions. Neither work investigates a physical instrument, though sim-to-real transfer might be within reach with sufficient engineering. Additionally, both works forego raw audio signals through explicit MIDI `note_on` events.

III. Problem Formulation

Given a chordophone with unknown geometry located in the physical workspace of the robot, our objectives are to

1) reconstruct the instrument's string geometry by system identification using mostly self-supervised motions,
2) explore the action space of plucking motions with these strings to characterize and model the produced audio responses incrementally over time, and
3) reduce unwanted robot motions that might damage or wear out the instrument (or robot).

To better isolate the first two objectives and to achieve a stable exploration of the whole valid action space, we propose a two-stage approach.

In the first stage, we reconstruct the relevant geometry of the instrument. We consider arbitrary chordophones with reachable strings $\mathcal{S}^*$. For each string s, we model its "playable" segment together with its fundamental note N_s (audio frequency) in a set $\mathcal{S}$:

$$\mathcal{S} = \left\{ \langle \hat{N}_s, \hat{T}_s, \hat{\ell}_s \rangle \mid s \in \mathcal{S}^* \right\},$$

as a line segment of length $\hat{\ell}_s$ starting at a 6-DOF pose $\hat{T}_s \in SE(3)$ and ending at $\hat{T}_s \cdot [\hat{\ell}_s, 0, 0, 1]^T$. As even the size of the ground-truth set of strings is a priori unknown, some initial information is required to bootstrap $\mathcal{S}$.

To define the action space Φ of the robot, we assume a parameterized pluck motion primitive $\mathcal{P}$ that describes

Cartesian trajectories T:

$$\mathcal{P} : (T_s : SE(3)) \times (\ell_s : \mathbb{R}) \times (\phi : \Phi) \rightarrow (\tau_s^{\phi} : \mathrm{T}),$$

such that at least one known parameter vector $\phi_s^{\text{init}} \in \Phi$ will yield the expected audio note onset for a well-localized string s when $\tau_s^{\phi^{\text{init}}}$ is executed by the robotic finger.

Once a stable geometry $\mathcal{S}$ is reconstructed, the second exploration stage proceeds to explore the action space and audio onset responses of the instrument as two related functions

$$\mathcal{V} : \mathcal{S} \times \Phi \rightarrow [0; 1], \text{ and}$$
$$\mathcal{D} : \mathcal{S} \times \Phi \rightarrow \Psi.$$

Here, $\mathcal{V}$ estimates a probability of ϕ to be *valid*, i.e., of τ_s^{ϕ} causing an acceptable single matching note onset with perceived note N_s associated with the string $s \in \mathcal{S}$, and $\mathcal{D}$ estimates the audio onset characteristics $\psi \in \Psi$ assuming ϕ were valid. In this work, we define Ψ as the measured loudness of the fundamental note of an audio onset in A-weighted decibels [dBA]. The general approach supports any measurable metric on note onsets, though, including overtone profiles and temporal envelope, both related to the timbre of the instrument and playing technique.

This decomposition into *functional* and *discriminative* functions is established in *Safe Active Exploration* approaches using Gaussian processes (e.g., [27, 28]) and facilitates reasoning over a set of safe (in our case *valid*) action parameters $\Phi_s^{\alpha} = \{\phi \in \Phi | \, \mathcal{V}(s, \phi) > \alpha\}$, where $1 - \alpha$ describes the remaining risk of an invalid action.

Lastly, we aim to utilize the combined system model $\langle \mathcal{S}, \mathcal{V}, \mathcal{D} \rangle$, which comprises all three data-driven components, to infer motion parameters that produce audio onsets for note N^t with characteristics close to ψ^t. To do so, we eventually optimize

$$\underset{\substack{s \in \mathcal{S} \\ \phi \in \Phi}}{\arg\min} (\mathcal{D}(s, \phi) - \psi^t)^2$$
$$s.t.\ N_s = N^t, \text{and } \phi \in \Phi_s^{\alpha}$$

IV. Methods

A fundamental requirement of our approach is the ability to detect and characterize plucks of tuned strings in the audio and tactile modality. To this end, we implement dedicated unimodal detectors. Aligned multimodal detections are then integrated to estimate the geometric model $\mathcal{S}$ through string fitting. After introducing a motion primitive $\mathcal{P}$ (and associated action space Φ) for plucking motions, we detail the exploration stages to autonomously estimate and refine the model components $\langle \mathcal{S}, \mathcal{V}, \mathcal{D} \rangle$ over time.

A. Audio Analysis

The audio noise floor near our robot reaches 60-68 dBA SPL, caused mainly by cooling fans. As we capture audio onsets from around 40 dBA SPL, we rely on a contact microphone attached to the resonating body of the instrument. However, note that this entails an uncalibrated audio signal that cannot be directly interpreted as sound pressure level. Instead, we consider dB above signal noise floor.

As we expect our audio signal to come from the musical domain and originate from a tuned instrument, we limit our analysis to frequencies associated with the western music scale. A well-established way for fast spectrum analysis in this field is described by the *constant Q-transform* (CQT) [29, 30], which can be computed based on the classical Fast-Fourier transform. The Guzheng's usual playing range encompasses notes from D2 (73.42 Hz) to D6 (≈1.17 kHz). To include essential harmonics, we include two more octaves in our analysis and consider a CQT of 84 semitones above D2. We sample the microphone with 44.1 kHz and use a standard hop length of 512 samples for the analysis, yielding an effective temporal resolution of 12 ms for further analysis. To detect note onsets, we extract maxima in the spectral flux envelope of the estimated CQT [31] in overlapping processing windows of 500 ms. The CQT exhibits maxima at the fundamental frequencies associated with each string. However, with strong overtones and instrument resonance, simple maximum detection yields many false positives to classify onsets by note. Instead, we use the *CREPE* network [32] to estimate the fundamental note of each onset.

B. Tactile Pluck Detection

The PR2 robot we use for our experiments is equipped with a Shadow Robotic Hand and *BioTac* tactile fingertips [33] (figures 1 and 2). As we attach the plectrum directly to the sensor, contact of the plectrum with a string of the instrument clearly reflects in the measured tactile feedback. The fingertip sensor provides a number of signals that can be used to evaluate contact position and state. For our purposes, reading the absolute pressure values generated by the BioTac at 100 Hz is sufficient to detect when the plectrum plucks, i.e., deflects and releases a string. We compute the numeric derivative of the signal and apply a tunable threshold to trigger a pluck event upon string release.

C. Pluck Validation

In order to evaluate the performance and location of an attempted pluck and reduce ambiguities in either detector, we align the two modalities and evaluate events together. Audio onsets are associated with detected tactile plucks if they occur in a short time frame (50 ms in our experiments) after the tactile pluck. All plucking attempts by the system are assigned a binary validity score $\nu \in \{1, -1\}$. Any audio onset that does not match a tactile pluck is assumed to be unintentional or externally caused and yields a negative score. Additionally, plucking attempts that fail to trigger any onset at all, trigger multiple onsets, or multiple tactile events, yield negative scores. Lastly, a plucking attempt that targets a specific note onset receives a positive score $\nu = 1$ if and only if a single tactile-grounded audio onset is observed during the pluck.

D. Kinematic Projection

Assuming a known position of the plectrum tip w.r.t. the finger it is attached to, each audio onset event, which is

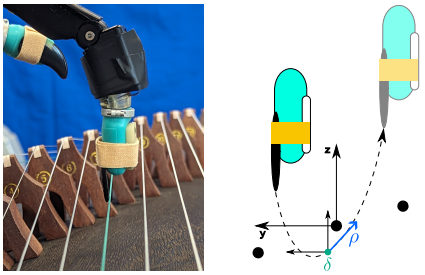

Fig. 3. (left) The fingers of the Shadow Hand with BioTac sensors in playing position, with the plectrum on the index finger touching the eighth string, tuned to *A4*. (right) Sketch of the parameterized pluck primitive in the local coordinate frame of a string located at the origin. The trajectory starts on one side of the string at a fixed position, dives between two strings, and accelerates upwards to reach the waypoint δ with velocity ρ. The last waypoint specifies a position above and behind the string to achieve the actual pluck.

grounded in a tactile detection, can be associated with a Cartesian position of this tip using the known forward kinematics of the robot. While the exact plectrum positions are usually unknown, the described exploration stages are robust enough to cope with discrepancies of several millimeters.

The resulting point projections of haptically-validated audio onsets can be seen in figure 1 (right). Relevant sources for the observable noise in this projection include (1) uncertainties in the plectrum position, (2) the deflection of the string during the plucking motion that differs per pluck, (3) an unmeasured deformation of the plectrum itself, (4) temporal skew between the different modalities, and (5) systematic errors in joint measurements and robot calibration.

E. String Fitting

To estimate the string coordinate frames $\hat{T}_s$ and length $\hat{\ell}_s$ of $\mathcal{S}$, we employ RANSAC [34] to fit lines through the groups of projected events classified with note $\hat{N}_s$. An outlier-resistant line fitting is required because *CREPE* sometimes misclassifies note onsets due to harmonics and instrument resonances, and background noise (talking, sneezing) can produce false-positive events. We use an inlier threshold of 1 cm for all single note onset events. We associate a coordinate frame T_s with each string that originates at the RANSAC model's right- and bottom-most inlier and points along the string in its x- and upwards in its z-axis.

Lastly, we can optionally assume a planar bridge for the mounting of all strings on the instrument and fit a 2D line on the regression plane of all inliers of string models that fits the origins of the computed coordinate systems. To align all frame origins with this line, we recompute them as intersections of the string model and the 3D plane associated with the fitted bridge line that is orthogonal to the regression plane. The resulting frames and fitted string segments can also be seen in figure 1.

F. Pluck Primitive

There are various plucking techniques for chordophones depending on the intended dynamic expression and musical context [6]. As a formal definition of relevant motions essentially includes unconstrained motion policies, the respective action space Φ to parameterize plucks via a motion primitive $\mathcal{P}$ is effectively intractable, and the automatic evaluation of their execution safety in the partially unknown physical world is very limited. Instead, we limit $\mathcal{P}$ to a strongly reduced parametrization that can be specified as a position on the string and a trajectory of the plectrum tip in the two-dimensional Cartesian coordinate system orthogonal to the string. The primitive connects fixed start- and end positions for either plucking direction d (inward or outward) with a parameterized intermediate position/velocity waypoint $\langle \delta, \rho \rangle$ close to the string, as illustrated in figure 3. Note that any such pluck model includes an implicit spring component, as the string is deflected. In our case, the finger joints exhibit sufficient spring compliance to store all additional potential energy until the plectrum releases the string. We chose the fixed start position closer to the string and the goal position slightly further away to allow for a broader range of transit motions to smoothly reach the start point and have more flexibility in continuing the trajectory after the physical pluck occurs. To generate smooth spline trajectory profiles with intuitive dynamics limits, we utilize the Ruckig trajectory generator [35] in Cartesian space. In summary, this primitive defines action parameters $\phi \in \Phi$ as

$$\langle d, \eta, \delta, \rho \rangle \in \{\textbf{in}, \textbf{out}\} \times [0; 1] \times \mathbb{R}^2 \times \mathbb{R}^2,$$

where $\eta \cdot \hat{\ell}_s$ defines the string position to pluck.

To actuate the described Cartesian trajectory τ_s^ϕ, we map it to the robot's joint space. Various methods can be used here, such as point-wise mapping, pseudo-inverse Jacobian steering, and constraint trajectory optimization. We utilize an iterative tracking loop based on the efficient *bio_ik* kinematics solver [36]. It tightly integrates with the MoveIt system [37] and can thus avoid collisions with the modeled environment also for longer motions moving between strings.

G. Geometry Exploration

Combining the reconstruction system for $\mathcal{S}$, and the pluck parametrization $\mathcal{P}$, the system gathers spatial evidence autonomously and refines $\mathcal{S}$ over time. Still, $\mathcal{P}$ relies on estimates $\hat{T}_s, \hat{\ell}_s$ for the string geometry as Cartesian reference space. Such initial estimates could come from a previous reconstruction, visual estimations without associated fundamental note annotations, or kinesthetic demonstrations close to the reachable ends of the strings. We adopt the latter approach and sample further plucks in between. Thus, these initial demonstrations determine the length of reconstructed string models $\hat{\ell}_s$. During autonomous operation, we repeatedly sample a direction, a string $s \in \mathcal{S}$, a string position $\eta \cdot \hat{\ell}_s$, execute the associated pluck with predetermined δ, ρ and evaluate the pluck's validity. On invalid plucks, we reattempt heuristically-modified parameters for a fixed number of steps, where we slightly lower δ and move it along the pluck direction on missing onsets, or stepwise rotate ρ upwards and move δ opposite the direction on invalid onsets. Valid plucks are integrated to update $\mathcal{S}$.

Experimentally, we found Halton sampling for η to improve model stability as uniform sampling is known to yield spatial clusters in few samples, which can cause string fitting to diverge. While the best sampling points to ground the string geometry are near the ends, we found it preferable to attempt plucks along the whole length of the string. Plucks near the ends of a string tend to excite higher harmonics, which degrade note onset detection, and initial onset misclassifications slow down convergence significantly. To account for long transit motions, we sample from $\mathcal{S}$ with a shaped distribution to encourage strings closer to the current finger position and penalize strings with strong evidence. Note that this is entirely optional, and target strings might also be determined by sheet music at the cost of a less balanced overall exploration.

H. Active Valid Pluck Exploration (AVPE)

After successful geometry exploration, we aim to actively select informative and likely-valid parameters from Φ to improve our approximations of $\mathcal{V}$ and $\mathcal{D}$. As a basis for both functions, we freeze $\mathcal{S}$ and fix reference geometries of the pluck primitive. Additionally, we reduce the number of free parameters Φ by freezing δ_z and ρ (in the respective direction), effectively describing actions through a *target string* s, a *binary direction* d, a *normalized string position* η, and δ_y, which we label the "*deflection offset*". Plucks in one direction can provide information about their opposing plucks, but in our experiments we observed no clear data correspondence due to varying controller responses. Thus, we assume independence in s and d, and model $\mathcal{V}, \mathcal{D}$ as independent functions $\mathcal{V}^s_{\textbf{in}}/\mathcal{D}^s_{\textbf{in}}, \mathcal{V}^s_{\textbf{out}}/\mathcal{D}^s_{\textbf{out}}$ per string s.

With each attempted pluck with parameters ϕ, we collect data points $\langle \phi, \nu, \psi \rangle \in \Phi \times \{1, -1\} \times \Psi$ on its validity and audio response. We aggregate $\langle \phi, \nu \rangle$ in $X^s_{\mathcal{V}}$, and $\langle \phi, \psi \rangle$ in X^s_{Ψ} whenever $\nu = 1$, to provide evidence for $\mathcal{V}$ and $\mathcal{D}$. Following existing safe active learning paradigms [27], we model both functions via Gaussian Processes with stationary squared exponential kernel and zero mean after normalizing X^s_{Ψ}. We directly apply GP regression with X^s_{Ψ}:

$$\begin{aligned}\mathcal{D}(s,\phi) &= \mathbb{E}[g|X^s_{\Psi}(\phi)];\ g \sim GP(rbf(\lambda_d)) \\ &= \mathbb{E}[\mathcal{N}(x;\mu_\psi,\sigma_\psi)] = \mu_\psi\end{aligned}$$

To model $\mathcal{V}$, we apply probit regression [38] to $X^s_{\mathcal{V}}$, assuming noisy measurements:

$$\begin{aligned}\mathcal{V}(s,\phi) &= p(\mathbb{E}[g|X^s_{\mathcal{V}}(\phi)] > 0);\ g \sim GP(rbf(\lambda_v)) \\ &= \int_{-\inf}^{0} \mathcal{N}(x;\mu_\nu,\sigma_\nu)dx \\ &= \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(-\frac{\mu_\nu}{\sigma_\nu\sqrt{2}}\right)\end{aligned}$$

Kernel scales $\lambda_\star$ were fixed, as trust regions for $\mathcal{V}$ must not depend on unstratified explored samples. Lastly, to determine the next best action parameters, we need an information criterion to optimize. While Bottero et al. [39] recently proposed a mutual information criterion on the discriminative function, we use the simpler differential entropy of the onset characteristics Gaussian Process, which can be maximized simply through the posterior predicted variance σ_ψ. Finally, we iteratively select the next best pluck to execute for a given string and direction as

$$\begin{aligned}\phi^* = \arg\max_{\phi\in\Phi} H[g|X^s_{\Psi}(\phi)];\ g \sim GP(rbf(\lambda_s)) \\ s.t.\ \mathcal{V}(s,\phi) > \alpha\end{aligned}$$

While we could optimize over $s \in \mathcal{S}$ as well, this would increase the computational burden per string and would not consider time for transit motions. Instead, we apply the exact sampling distribution used in the geometry exploration. To accelerate exploration at the cost of slightly more invalid plucking attempts, we accept a lower validity threshold $\alpha = 0.7$ during active exploration and solve the optimization through Monte Carlo methods. An exemplary $\langle V, D \rangle$ model and the underlying dataset are illustrated in figures 4 and 5.

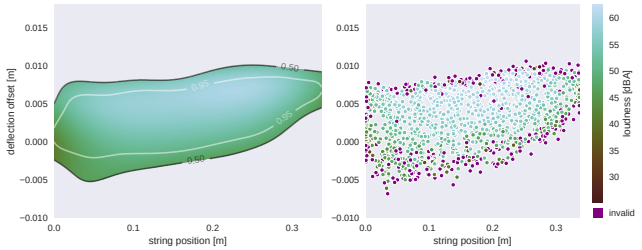

Fig. 4. (left) Exemplary pluck dynamics $\langle \mathcal{V}, \mathcal{D} \rangle$ fit for 1500 explored F♯4 plucks. Contours indicate $\Phi^{0.5}_s$ and $\Phi^{0.95}_s$. (right) All associated explored plucks, including those evaluated as invalid (highlighted in purple).

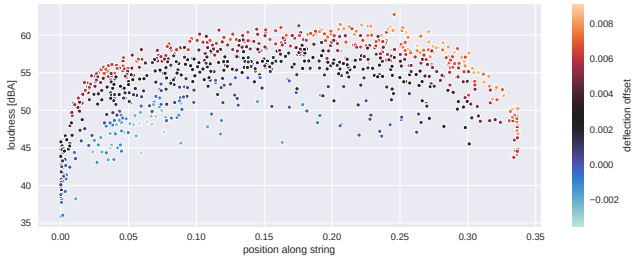

Fig. 5. Dynamic range of explored pluck motions for F♯4 along the string. As expected, plucks at the center of the string produce significantly louder onsets with similar deflection offsets.

V. Experiments

As an extensive evaluation of all required system components is beyond the scope of this paper, we focus on evaluating the higher-level exploration stages. Typical error conditions of the lower-level modules include (*a*) soft onsets being missed or misclassified when strong overtones are present or the audio-tactile modalities are misaligned, (*b*) false positive pluck detections when a plectrum slides along the string, (*c*) temporal drift between the robot sensor data and the audio channels due to drifting clocks. We mitigate this last problem through occasional manual recalibration.

To illustrate the behavior of the exploration stages, we consider several representative experiments, each limited to one finger and individual strings.

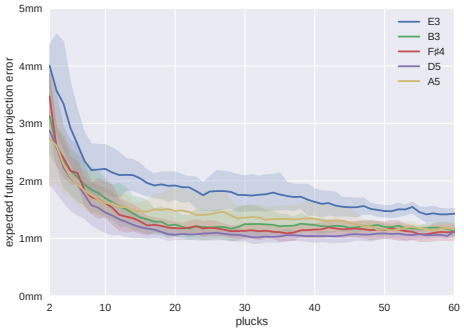

Fig. 6. Expected Euclidean model error between kinematic onset projections p and their projection $m(p)$ onto the current RANSAC line fit as the number of attempted plucks increases. Expectations are approximated through 30 following plucks at each point. Five episodes were collected per string.

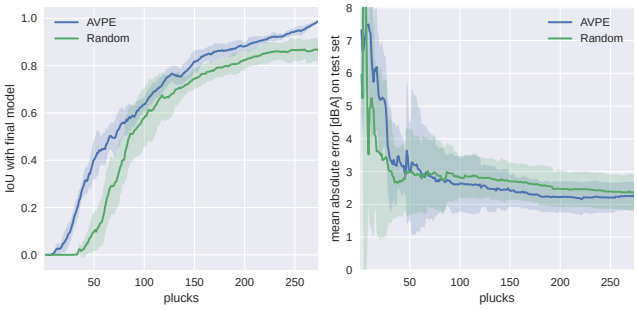

Fig. 7. (left) IoU of valid action space volumes $\Phi_s^{0.75}$ after each new pluck and the final valid volume. Volumes were estimated through Monte Carlo integration. (right) Mean absolute error of $\mathcal{D}$ on test set over number of recorded plucks. Both cases compare AVPE and a hypothetical random baseline.

A. Geometry Exploration

To quantify the results of the string reconstruction stage, we repeatedly initialized single strings through two kinesthetic demonstrations with varied instrument positions and let the system explore each string through 90 plucks. We analyze the expected Euclidean projection error for the current string model m_i: $E[\|m_i(p) - p\|_2]$ and estimate it through the next 30 successful plucks. Figure 6 illustrates the results of five strings with five episodes each. All models converge to an expected projection error of about $1\,\mathrm{mm}$, corresponding to the limitations of the PR2 robot and the unmodelled deflection of plectrum and string. Almost all episodes converge between 10 and 20 explorative plucks, where the exact behavior does not strictly depend on the length of the string. We sample each string at least 15 times during the complete geometric exploration stage to balance the best achievable precision and swift exploration.

B. Pluck Exploration

To evaluate our active valid pluck exploration process, we investigate the growth and evolution of the valid parameter set throughout the exploration, using an intersection over union (IoU) measure with the corresponding final valid parameter set. As shown in figure 7, a significant portion of the final set is approximated within the first 70 samples where IoU exceeds 0.5. Progress near the border regions

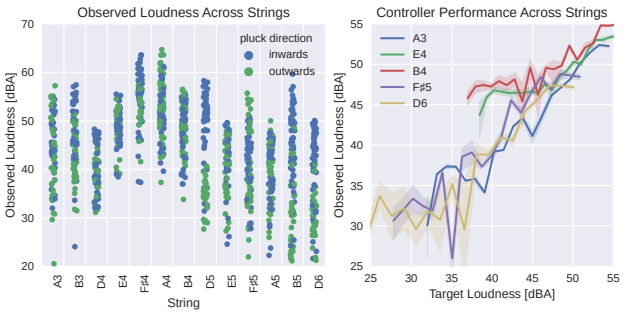

Fig. 8. (left) loudness distribution of evaluated plucks for 13 strings after 1300 plucks of autonomous exploration. All reconstructed strings could be plucked within a dynamic range of at least 20 dBA, corresponding to a four- to eightfold subjective doubling of the loudness. (right) Performance curve of predicted plucks for a target loudness.

takes significantly more attempted plucks. A hypothetical baseline exploration, which integrates random samples from the same dataset, performs significantly worse, because it does not systematically grow the initial valid set. As it is not bound by the unknown validity of sampled plucks, it provides global information on $\mathcal{D}$ with less sampled plucks, but the mean absolute test set error still falls below 3 dBA for both policies within 50 plucks.

C. System-Level Integration

To evaluate whole-system integration, we initialized estimates for the 13 upper strings of the instrument and let the system explore for two hours, automatically switching to the second exploration stage after 200 successful plucks. Figure 8 (left) shows the resulting loudness distribution observed during exploration. The significant shift in measured loudness between center notes appears as an artifact of the instrument's resonance and the contact microphone.

Lastly, we can optimize $\mathcal{D}$ within $\Phi^{0.95}$ through Monte Carlo methods to predict parameters for target loudness values within the explored range and evaluate them through execution. As shown in figure 8 (right), response predictions are usually accurate within 3 dB, but flatten out where weak onsets cannot be reproduced sufficiently well.

VI. Conclusion

We proposed a two-stage exploration approach that allows a robot to reconstruct the geometry of chordophones and learn associated plucking motions for expressive play. Our PR2 robot successfully explored the doubly-curved geometry of a Chinese Guzheng and characterized motion primitives for plucking with desired note onsets with a small margin of error. We plan to investigate different plucking techniques using multiple fingers in the future.

The general active valid exploration paradigm using motion primitives is not bound to musical instruments and we will also investigate transfer to other domains. Lastly, the system model generated by the exploration could be exploited in simulation environments to provide a safe training ground for reinforcement learning with a reduced reality gap.

All source code and discussed datasets can be found at https://github.com/TAMS-Group/chordophone_exploration.

REFERENCES

[1] Mason Bretan and Gil Weinberg. "A Survey of Robotic Musicianship". In: *Commun. ACM* 59.5 (Apr. 2016), pp. 100–109. DOI: 10.1145/2818994.

[2] Jim Murphy et al. "Expressive Robotic Guitars: Developments in Musical Robotics for Chordophones". In: *Computer Music Journal* 39.1 (2015), pp. 59–73. DOI: 10.1162/COMJ_a_00285.

[3] Takumi Ogata and Gil Weinberg. "Robotically augmented electric guitar for shared control." In: *NIME*. 2017, pp. 487–488.

[4] Gil Weinberg et al. *Robotic Musicianship: Embodied Artificial Creativity and Mechatronic Musical Expression*. Springer, 2020, p. 270. DOI: 10.1007/978-3-030-38930-7.

[5] C. Waltham. "An Acoustical Comparison of East Asian and Western String Instruments". In: *Proceedings Intl. Symposium on Musical Acoustics (ISMA 2014), Le Mans, France*. 2014, pp. 375–380. URL: http://conforg.fr/isma2014/.

[6] Zixuan Fu. "A Brief Analysis of the Performance Skills and Treatment of Guzheng's". In: *Proceedings of the 2021 3rd International Conference on Literature, Art and Human Development (ICLAHD 2021)*. 2021, pp. 733–741. DOI: https://doi.org/10.2991/assehr.k.211120.134.

[7] Hailei Ding et al. "Automatic Recognition of Basic Guzheng Fingering Techniques". In: *Proceedings of the 8th Conference on Sound and Music Technology*. Springer Singapore, 2021, pp. 66–77. DOI: 10.1007/978-981-16-1649-5_6.

[8] Ajay Kapur. "A History of robotic Musical Instruments". In: *Proceedings ICMC*. 2005.

[9] Tarek M. Sobh, Bei Wang, and Kurt W. Coble. "Experimental Robot Musicians". In: *J. Intell. Robotic Syst.* 38.2 (2003), pp. 197–212. DOI: 10.1023/A:1027319831986.

[10] Ada Zhang, Mark Malhotra, and Yoky Matsuoka. "Musical piano performance by the ACT Hand". In: *2011 IEEE international conference on robotics and automation*. IEEE. 2011, pp. 3536–3541.

[11] J. A. E. Hughes, P. Maiolino, and F. Iida. "An anthropomorphic soft skeleton hand exploiting conditional models for piano playing". In: *Science Robotics* 3.25 (2018), eaau3098. DOI: 10.1126/scirobotics.aau3098.

[12] Ting Fei et al. "Performance control system of dulcimer music-playing robot". In: *2017 11th Asian Control Conference (ASCC)*. 2017, pp. 1345–1350. DOI: 10.1109/ASCC.2017.8287367.

[13] Deng Xiaowei et al. "Simulation Analysis of Vibro-Acoustic Characteristics of Traditional Guzheng". In: *Journal of Shanghai Jiaotong University* 50.02, 300 (2016), p. 300.

[14] Enda Zhang et al. *An Efficient Modal-based Approach Towards Guzheng Sound Synthesis*. 2019. DOI: 10.48550/ARXIV.1910.05447.

[15] Przemyslaw Mazurek and Dorota Oszutowska-Mazurek. "String Plucking and Touching Sensing using Transmissive Optical Sensors for Guzheng". In: *16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. 2020, pp. 1143–1149. DOI: 10.1109/ICARCV50220.2020.9305480.

[16] Delphine Chadefaux et al. "Harp plucking robotic finger". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012, pp. 4886–4891. DOI: 10.1109/IROS.2012.6385720.

[17] Chris Duxbury and Mark Sandler and Mike Davies. "A Hybrid Approach to Musical Note Onset Detection". In: *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02), Hamburg, Germany*. Sept. 2002, pp. 33–38.

[18] Ruohua Zhou and Joshua D. Reiss. *Music Onset Detection, Chapter 12 in Machine Audition: Principles, Algorithms and Systems*. Ed. by W. Wang. July 2010. DOI: 10.4018/978-1-61520-919-4.ch012.

[19] Mounir, M. and Karsmakers, P. and van Waterschoot, T. "Musical note onset detection based on a spectral sparsity measure". In: *J AUDIO SPEECH MUSIC PROC. 2021* (July 2021), p. 30. DOI: 10.1186/s13636-021-00214-7.

[20] Mina Mounir, Peter Karsmakers, and Toon van Waterschoot. "CNN-based Note Onset Detection using Synthetic Data Augmentation". In: *2020 28th European Signal Processing Conference (EUSIPCO)*. 2021, pp. 171–175. DOI: 10.23919/Eusipco47968.2020.9287621.

[21] Ziyi Wang and Yin Cao. "An On-line Algorithm for Music-to-Score Alignment of Guzheng Performance". In: *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. 2018, pp. 1–5. DOI: 10.1109/ICDSP.2018.8631834.

[22] Gerelmaa Byambatsogt, Lodoiravsal Choimaa, and Gou Koutaki. "Guitar Chord Sensing and Recognition Using Multi-Task Learning and Physical Data Augmentation with Robotics". In: *Sensors* 20.21 (2020). DOI: 10.3390/s20216077.

[23] Qingyang Xi et al. "Guitarset: A dataset for guitar transcription". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*. 2018, pp. 453–460.

[24] Jiali Chen et al. "A Music-driven Deep Generative Adversarial Model for Guzheng Playing Animation". In: *IEEE Transactions on Visualization and Computer Graphics* (2021). DOI: 10.1109/TVCG.2021.3115902.

[25] Huazhe Xu et al. "Towards Learning to Play Piano with Dexterous Hands and Touch". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022, pp. 10410–10416. DOI: 10.1109/IROS47612.2022.9981221.

[26] Kevin Zakka et al. “RoboPianist: A Benchmark for High-Dimensional Robot Control”. In: *arXiv preprint arXiv:2304.04150* (2023).

[27] Jens Schreiter et al. “Safe exploration for active learning with Gaussian processes”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III 15*. Springer. 2015, pp. 133–149.

[28] Cen-You Li, Barbara Rakitsch, and Christoph Zimmer. “Safe active learning for multi-output gaussian processes”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 4512–4551.

[29] Judith C Brown and Miller S Puckette. “An efficient algorithm for the calculation of a constant Q transform”. In: *The Journal of the Acoustical Society of America* 92.5 (1992), pp. 2698–2701.

[30] Christian Schörkhuber and Anssi Klapuri. “Constant-Q transform toolbox for music processing”. In: *7th sound and music computing conference, Barcelona, Spain*. 2010, pp. 3–64.

[31] Sebastian Böck and Gerhard Widmer. “Maximum filter vibrato suppression for onset detection”. In: *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*. Vol. 7. 2013, p. 4.

[32] Jong Wook Kim et al. “Crepe: A Convolutional Representation for Pitch Estimation”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 161–165. DOI: 10.1109/ICASSP.2018.8461329.

[33] Nicholas Wettels, Jeremy A Fishel, and Gerald E Loeb. “Multimodal tactile sensor”. In: *The Human Hand as an Inspiration for Robot Hand Development*. Springer, 2014, pp. 405–429.

[34] Martin A. Fischler and Robert C. Bolles. “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”. In: *Commun. ACM* 24.6 (June 1981), pp. 381–395. DOI: 10.1145/358669.358692.

[35] Lars Berscheid and Torsten Kröger. “Jerk-limited Real-time Trajectory Generation with Arbitrary Target States”. In: *Robotics: Science and Systems XVII* (2021).

[36] Philipp Ruppel et al. “Cost Functions to Specify Full-Body Motion and Multi-Goal Manipulation Tasks”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 3152–3159. DOI: 10.1109/ICRA.2018.8460799.

[37] David Coleman et al. “Reducing the Barrier to Entry of Complex Robotic Software: a MoveIt! Case Study”. In: *Journal of Software Engineering for Robotics* 5.1 (May 2014), pp. 3–16. URL: http://moveit.ros.org.

[38] Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*. Vol. 1. Springer, 2006.

[39] Alessandro Bottero et al. “Information-Theoretic Safe Exploration with Gaussian Processes”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 30707–30719.