

Issues and Challenges in Vision-Based Mid-Air Musical Instruments: Control Estimation in an Optical Theremin

Mario Sguario, Nicola Garau, Nicola Conci, and Francesco Ardan Dal Rí

DISI — Department of Information Engineering and Computer Science

University of Trento

Italy

mario.sguario@studenti.unitn.it, nicola.garau@unitn.it, nicola.conci@unitn.it, francesco.dalri-2@unitn.it

Abstract—Mid-air musical instruments, the Theremin arguably being the most relevant example, rely on continuous, contact-free performer hand gestures and are highly sensitive to small changes in hand position, making them a compelling target for fine-grained vision-based tracking. This paper investigates the issues and challenges involved in estimating Theremin control signals (pitch and volume) from the sole visual hand-joint trajectories captured using a) monocular vision, b) stereo vision, and c) motion capture. Through a series of controlled experiments, we evaluate practical benchmarks aimed at isolating technical issues and potential failure cases in vision-based mid-air instrument tracking. We analyze how sensing modality, temporal resolution, real-time deployment constraints, and take-to-take variability can affect online control estimation, discussing methodological considerations for improving visual tracking under real-world conditions.

Index Terms—Optical Theremin, Control estimation, Computer Vision, Hand tracking

I. INTRODUCTION

Recent advances in computer vision and deep learning have made body and hand tracking increasingly accessible for interactive systems. Well-known frameworks such as OpenPose [1] and MediaPipe [2] demonstrate that human pose and hand landmarks can be efficiently estimated from ordinary video streams with sufficient robustness for many real-time or near-real-time applications. In particular, hand gestures have become a prominent control in musical instruments, with MediaPipe constituting a common practical choice for camera-based applications due to its ability to provide real-time hand landmarks from consumer RGB cameras [3], [4]. More broadly, these advances have sparked interest in optical instruments, where performers' motion can be translated into sound without physical contact [5], [6]. Accordingly, such developments are also relevant to the Internet of Sounds (IoS), where musical interaction may involve networks of sensors and musical objects [7].

In literature related to Digital Musical Instruments (DMI), interaction strongly depends on the mapping between gesture and resulting sound, where tracking error can be negligible in one mapping and musically disruptive in another [8]. For mid-air instruments, this issue is exacerbated by the absence of

physical contact or tactile anchors to stabilize the performer's gestures. As a result, latency, jitter, ambiguity, drift, and other tracking issues can become part of the musical behavior of the instrument itself [8], [9]. From a networked-musical-systems perspective, similar issues can also affect the reliability of continuous control streams [10]. In this sense, the Theremin is a particularly demanding case, as its playability strictly depends on extremely precise spatiotemporal mapping [11]. In addition to being one of the early examples of electronic instruments, it is still considered the most prominent archetype of mid-air musical instruments, as it is played through hand motion around two electromagnetic fields, controlling pitch and volume through the performer's proximity to two antennas. Unlike several gesture instruments that can quantize notes or trigger discrete events, analog Theremins¹ directly expose subtle spatial drifts as pitch instability, volume artifacts, or hampered playability, requiring performers to maintain fine-grained control without any reference points, which contributes to the instrument's well-known playing difficulty. Indeed, previous work has already shown that free-gesture instruments can benefit from visual guides, mixed-reality augmentation, optical sensing, or simplified mappings [12]–[14].

Building on the aforementioned technological advances, a substantial body of related work has proposed many versions of digital, optical, or depth-sensor-based Theremin-like instruments, showing a growing interest in mid-air musical interactions [15]–[18]. Yet, many of such works mostly tackle the problem from the point of view of interface/interaction design, learnability, or accessibility; at the current state, more restrictive tasks related to reconstructing precise control signals from hand motion remain a relevant open problem.

In this paper, we address this gap by investigating the issues and challenges of vision-based Theremin control estimation under real-time constraints. We track pitch and volume Control Voltages (CVs) while simultaneously recording hand motion with three sensing configurations: monocular camera tracking

¹Recent digital or digitally-controlled Theremins, such as the Moog Claravox or the D-Lev, allow field quantization; however, in this paper, we consider the more traditional analog models, which do not have this capability.

only with MediaPipe landmarks, stereo-camera tracking with both MediaPipe and depth sensing, and marker-based motion capture. A compact recurrent model is trained to estimate pitch and volume CV signals from the corresponding hand trajectories. Throughout a series of experiments, we compare the behavior of these configurations and analyze how practical factors affect the estimation process. The present study is intentionally restricted to a single experienced performer and a fixed instrument, calibration, and sensing setup; accordingly, the experiments are intended to characterize the technical feasibility and limitations of control reconstruction within a controlled configuration, rather than cross-performer or cross-setup generalization. The main contribution of this work thus lies in a measurement-oriented account of optical Theremin emulation: which sensing conditions support usable CV reconstruction, which artifacts degrade control, and what compromises emerge between low-cost deployment and high-precision tracking.

Specifically, this paper thus contributes a comparative benchmark of lightweight real-time regressor pipelines tailored to pitch and volume CV estimation, and an analysis of how sensing modality, temporal context, latency, and tracking artifacts affect the feasibility of vision-based mid-air musical control.

II. BACKGROUND

A. Hand-joint tracking in computer vision

Hand joint tracking has evolved from image-based 2D hand keypoints estimation to temporally stable 3D hand pose and shape estimation. Early RGB-based methods showed that reliable supervision could be obtained without full manual annotation or with synthetic data [19], [20].

MediaPipe Hands [2] combined palm detection with real-time hand landmark regression in an efficient on-device pipeline. While designed for deployment rather than as a pure research benchmark, it became a reference point for real-time hand tracking in applications. In parallel, parametric hand models such as MANO [21], introduced in the context of embodied hand and body modeling, provided a compact representation of hand articulation and shape, enabling optimization-based fitting, mesh recovery, and physically coherent tracking. Datasets such as FreiHAND [22] further supported this shift by providing real RGB images with 3D hand pose and shape annotations, making monocular hand reconstruction more comparable across methods.

Recent work focuses less on isolated hand pose and more on robustness, temporal consistency, and interaction. HOT3D [23] targets 3D hand and object tracking from egocentric multi-view videos, reflecting the growing importance of AR/VR and embodied perception settings; Ren et al. [24] explicitly model temporal dynamics for sequential 3D hand pose estimation, addressing typical failure cases such as fast motion, self-occlusion, and depth ambiguity; UniHOPE [25] unifies hand-only and hand-object pose estimation, making the representation more suitable for interaction-heavy scenes. Novel hybrid representations [26] further combine image evidence with

structural pose cues to improve 3D hand pose and shape recovery.

B. Vision-based mid-air musical instruments

Mid-air musical instruments use body, hand, or finger motion as musical input without any physical contact with the hardware interface. This design space has been explored in a variety of related fields, such as accessibility [27], immersive interaction [28], networked performance [3], and novel musical expression [29]. Removing contact, however, causes the loss of physical or tactile stabilization, making precise control dependent on sensing quality, mapping design, and immediate feedback. Because of this, several systems are deliberately designed to reduce the required precision to be interacted with. For instance, Lages *et al.* avoid absolute note placement in a 3D space by using interval gestures [30]; Cocchiara and Turchet target musically-untrained people, promoting simple collaborative music-making through webcam-based finger tracking over virtuosic control [3].

The limits of mid-air interaction also extend beyond technological means: unsupported motion introduces fatigue and occlusion problems. Cui and Sourin indeed discuss these factors in precision-oriented mid-air 3D modeling, observing that optical tracking must compensate for both sensor limitations and intrinsic hand-motion instability [31]. Additionally, these issues can become musically relevant when the tracked signal is closely mapped to fine-grained continuous sound parameters, as in the case of the Theremin [32]. Recent autoethnographic work on Theremin learning further supports this view, showing that touchless performance strongly relies on auditory feedback, proprioception, and mental imagery [11]. Latency is also critical: in a gesture-controlled instrument, Mäki-Patola and Hämäläinen found a just-noticeable latency range around 30 ms at max, with detectability depending on playing style and expertise [33].

Taken together, these works show that mid-air musical control is shaped by a tension between expressiveness and precision, making vision-based control estimation a very constrained problem where small tracking errors, instability, or latency can highly affect the resulting sound and playability.

C. Digital, optical, and other Theremin-inspired instruments

Theremin-inspired systems can be grouped into three main categories: augmentation of the physical instrument, reinterpretation or adaptation through alternative sensing, and fully digital/optical emulation.

Augmentation systems still preserve real Theremins while adding feedback or guidance. Relevant examples include Hasan *et al.*'s Termenova, which adopts capacitive sensing and laser guides as visual references for free-gesture performance [12]; Johnson and colleagues' VRMin [14] and MR:emin [13], which both similarly augment a physical Theremin with mixed-reality cues for tutoring, showing that visual feedback can improve practicing effectiveness and reduce errors; or Xiao *et al.*'s T-Voks [34] expands sonic palette via a force-sensing resistor on the left hand.

Optical and digital Theremins replace the original sensing mechanism with cameras or depth/light sensors. For instance, Dos Santos and Salazar propose a low-cost optical implementation including audio synthesis, graphical feedback, frequency-distance linearization, and pitch-error evaluation [15]; Sourin and Xing exploit a Leap Motion² controller in a system completed with real-time synthesis and visual feedback, reporting learnability and responsiveness while also identifying tracking precision, smoothing, and objective latency/pitch evaluation as limitations [17], [18]; Usabaev *et al.* design a virtual Theremin for film-scoring interaction, using web technologies to make the instrument accessible within a playful and approximate interaction context [16].

Finally, other related work moves away from optical vision. For instance, a hybrid digital Theremin introduced by García-Guzmán and colleagues retains antenna-based control while adding digital synthesis and calibration flexibility [35]; Bello *et al.* repurposed an OpenTheremin³ board as a capacitive sensing device for posture and gesture recognition, embedding conductive textile antennas into a loose garment [36]; Hermawan replaced the Theremin’s electromagnetic principle with a room-scale wireless network of ultrasound sensors, mapping body distance and movement to MIDI messages [37]; more recently, Liu *et al.* propose metasurface sensing for fine dual-hand perception and interference-resistant Theremin-like interaction [38].

To the best of our knowledge, although Theremin-inspired instruments have been widely reimaged, an accurate vision-based reconstruction of pitch and volume controls remains overall underexplored, motivating the scope of this work.

III. EXPERIMENTAL FRAMEWORK

The experimental setup was designed to record the performer’s hand movements and the corresponding CVs under different sensing conditions. Since the two hands perform different functions and separate gestures, pitch and volume settings were recorded individually.

Hand motion was captured using three sensing modalities: monocular vision, stereo vision, and motion capture.

A. Sensing Setup

The vision-based setup consisted of:

- A portable webcam (Logitech C905), 640x480 30fps;
- Two stereo camera (StereoLabs ZED 2i), 2560x720 30fps;
- An OptiTrack motion capture system (8 PrimeX 22, 6.9mm focal length), 360 fps.

The cameras were positioned at a fixed distance from the Theremin, determined by preliminary tests, covering the full gesture range and preserving a clear viewpoint of the hands in their traditional technique position.

For motion capture, reflective markers were placed directly on the performer’s hands. Initially, we tried simple solutions

such as adhesive supports and marker gloves. However, these were not optimal, both in terms of stability during performance, and in terms of interfering with MediaPipe hand tracking as they strongly alter the visual appearance of the hands. Therefore, to ensure stable placement, we designed a series of 3D-printed ring supports holding reflective markers on top. Several marker configurations were tested, according to the typical movements of traditional 8-position Theremin technique [39]. A one-marker-per-joint configuration proved unreliable, as it produced frequent marker displacement, occlusion, tracking loss, and identity confusion. Therefore, we adopted a reduced configuration with six markers on the right hand (pitch) and three markers on the left hand (volume). Marker locations were selected to capture the characteristic playing positions while reducing redundancy on fingers whose movements are strongly constrained. Their asymmetry follows the distinct roles of the two hands: the right hand articulates across a subtle and fine-grained pitch field, so its configuration must be resolved in detail, whereas the left one mostly operates through simpler opening/closing gestures that three markers already capture — Fig. 1.



Fig. 1: The displacements of the marker rings on the left hand (volume) and right hand (pitch). This configuration is optimized according to the traditional playing technique.

Three additional markers were placed on each Theremin antenna and treated as rigid bodies, allowing the antenna positions to be used as spatial references — Fig. 2.

The pitch and volume control voltages were acquired separately. This was necessary because the two control signals cannot be reliably recovered from the audio output alone: when the volume is low, pitch information becomes indeed partially or fully unretrievable. Therefore, we rely on a Moog Etherwave⁴ model with individual CV outputs. CV acquisition was performed using a Monome Crow⁵ Eurorack module running a custom Lua script. The script converts the incoming nominal voltage ranges, $[-2.5V, +4.5V]$ for pitch and $[0V, +10V]$ for volume, and normalizes both signals to the $[0, 1]$ interval. The normalized CV streams for each take were sent to a dedicated Python thread and stored as .csv files.

A thread barrier was used to start the acquisition streams jointly, while each stream was then processed according to its

²<https://www.leapmotiontechnology.com/>

³<https://www.gaudi.ch/OpenTheremin/>

⁴<https://www.moogmusic.com/synthesizers/etherwave-theremin/>

⁵<https://monome.org/docs/crow/>



Fig. 2: The Theremin setup during a take. Antennas and other reflective parts were covered to avoid visual interference.

native sampling rate. CV samples were associated with the closest-available visual frame, after resampling.

B. Recording Protocol

Before each recording session, the Theremin was calibrated to obtain a comfortable and musically usable playing range, consistent with a conventional performance setup.

For each hand, we recorded multiple takes combining visual streams, motion-capture data, and the corresponding CV signal. Individual takes were recorded by the last author of this paper, who is a trained Thereminist with over 10 years of experience. The protocol included a selection of common controlled gestures derived from traditional Theremin technique, together with short free-play recordings.

For the pitch hand, the recorded takes were:

- *Slow-Glissando*: one-octave continuous glissando, positions 1-8, 60 bpm, 10 repetitions.
- *Fast-Glissando*: one-octave continuous glissando, positions 1-8, 120 bpm, 20 repetitions.
- *Slow-Scale*: ascending/descending major-scales, positions 1-8, 60 bpm, 10 repetitions.
- *Fast-Scale*: ascending/descending major-scales, positions 1-8, 120 bpm, 20 repetitions.
- *Free Play*: approximately 3 minutes of unconstrained playing.

The pitch exercises were performed in the middle register of the instrument, between G3 and B4.

For the volume hand, the recorded takes were:

- *Slow-Arc*: full-range volume arcs, 60 bpm, 10 repetitions.
- *Fast-Arc*: full-range volume arcs, 120 bpm, 20 repetitions.

- *Staccato*: fast articulations, 90 bpm, one articulation per beat, 50 repetitions.
- *Free Play*: approximately 2 minutes of unconstrained playing.

The protocol covers smooth continuous gestures, discrete pitch movements, abrupt volume articulations, and unconstrained performance material. In total, we collected about 16'50" of recordings; consistent with previous literature, e.g., [40], this constitutes a sufficient amount to fit low-capacity models such as the ones we used (see Sec. III-D).

C. Data Preprocessing

The aligned raw recordings underwent specific processing steps to end up with clean joint data, shown in Fig. 3.

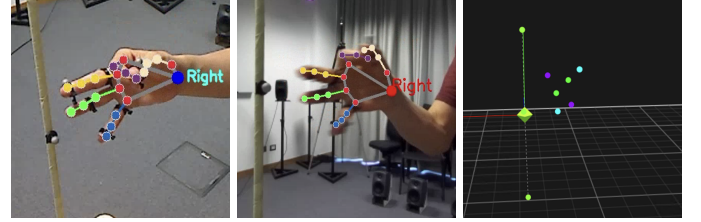


Fig. 3: Examples of rendered joints data (pitch hand) from the webcam (left); ZED (middle); and MoCap (right).

Webcam and ZED videos are processed frame-by-frame using MediaPipe to detect hand landmarks. For the ZED pipeline, three-dimensional positions are derived by first obtaining the wrist's metric depth from the ZED point cloud and then computing the remaining landmarks' positions relative to the wrist via the perspective projection of their MediaPipe relative 3D coordinates.

In both vision settings, MediaPipe occasionally produced artifacts and hallucinations, mainly spurious hand detections or left/right label swaps. We therefore retained the most label-consistent hand across frames, corrected local swaps when the non-target wrist matched the target-hand trajectory, and, for the ZED pipeline, applied filtering to suppress residual spikes before saving the cleaned features as arrays of shape $(T, 63)$.

The OptiTrack motion capture data are initially processed in Motive to create rigid bodies for the antennas and cameras, along with a marker set for each hand to ensure proper identification in later processing, lost markers were manually labeled using the label tool to achieve a detection rate $\geq 96\%$ for each of them. The cleaned data are downsampled from 360 fps to 60 fps via block averaging, then saved as arrays of shape $(T, 18)$ and $(T, 9)$ for the pitch and volume hands respectively.

Table I reports the proportion of samples a) detected by each sensing modality; b) discarded by artifact rejection; and c) kept into the feature arrays. Notably, the webcam resolves the hand with three times fewer pixels than the ZED, leaving the webcam pitch models fitted on little more than half of the recorded material.

TABLE I: Sample coverage per sensing modality. Percentages of all coordinate samples (frames $\times$ features), aggregated over the five pitch and four volume takes.

Modality	CV	Detected (%)	Removed (%)	Valid (%)
Webcam	Pitch	55.31	0.00	55.31
	Volume	79.58	0.04	79.54
ZED	Pitch	98.99	0.10	98.89
	Volume	88.21	1.51	86.70
MoCap	Pitch	99.61	0.00	99.61
	Volume	99.80	0.00	99.80

D. Architecture and Training Pipeline

The regression task considered here maps short temporal windows of sequential hand motion to a scalar CV value. Since hand movements are constrained by the biomechanics and inertia of the arm/hand system, trajectories are expected to be generally smooth and locally predictable [41], [42]. Moreover, the eventual deployment of these models is intended for real-time applications: therefore, we selected a lightweight LSTM-based regressor [43] as a simple recurrent baseline, capable of exploiting short-term temporal context with limited computational overhead.

The network first applies a frame-wise coordinate encoder composed of two fully-connected layers with SiLU activations, layer normalization, and dropout. The encoded sequence is then processed by a single-layer LSTM, and the final hidden state is passed to a small regression head that outputs a single CV value. In the current configuration, models are quite tiny, ranging from 60593 trainable parameters (63-inputs MediaPipe) to 53618 (9-inputs MoCap volume). The same architecture is used for both frame-wise and sequence-wise conditions: for length $L = 1$, the model receives a single frame, while for $L = 5$ it receives five consecutive frames and predicts the CV associated with the last frame. Thus, for an input sequence $X_{t-L+1:t}$, the target is y_t .

Models were trained by minimizing Mean-Squared Error (MSE) using the AdamW optimizer, with an initial learning rate of 10^{-3} and weight decay of 10^{-4} . Training was run for 50 epochs with a batch size of 64. A learning rate scheduler was used to decrease the learning rate by a factor of 0.75 when the monitored loss stopped improving, with a patience of 5 epochs and a minimum learning rate of 10^{-6} . Gradients were clipped with a maximum norm of 1.0 for training stability. All experiments were run on a single NVIDIA GeForce RTX 4090 GPU. To avoid temporal leakage, all evaluations used blocked temporal 5-fold cross-validation, with 80%/20% train/test assignment performed before windowing. Folds were constructed as contiguous temporal blocks within each recording category so that neighboring frames from the same temporal segment could not appear in both training and test sets. This provides a more stable estimate of generalization performance, especially in the sequential setting. Still, this evaluation measures generalization to unseen

temporal segments within the same performer, calibration, and sensing setup: it should therefore not be interpreted as evidence of cross-performer or cross-configuration generalization. We also include a simple MLP-based baseline using the same sequence lengths (here, $L = 5$ windows are flattened) to compare against the LSTM under matched temporal inputs, but without recurrent modeling.

IV. RESULTS

A. Quantitative Performance Benchmark

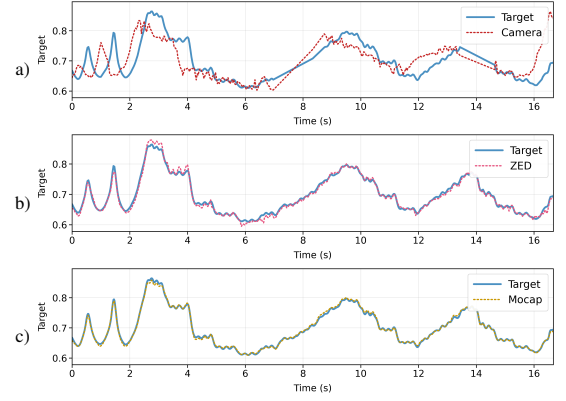


Fig. 4: Example of 500 sequential frames of target + reconstructed pitch CV using a) Webcam_5f, b) ZED_5f, and c) MoCap_5f models trained on the whole respective corpora.

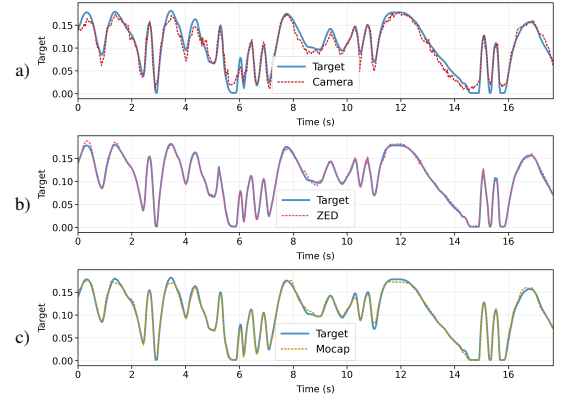


Fig. 5: Example of 500 sequential frames of target + reconstructed volume CV using a) Webcam_5f, b) ZED_5f, and c) MoCap_5f models trained on the whole respective corpora.

Albeit our experiments show that precise CV reconstruction and accurate spatial mapping from hand/joints trajectories is feasible, this task is strongly conditioned by both the hardware and the type of gesture. Representative examples of the reconstructed pitch and volume trajectories are shown in Figs. 4 and 5.

Slow and regular trajectories, such as glissandi and volume arcs, are reconstructed accurately by all configurations: this suggests that basic spatial relations are learnable even from simple, monocular landmarks. However, differences in accuracy between different modalities become more pronounced

TABLE II: Theremin control voltage prediction accuracy: 5-fold cross-validation test results, grouped by stimulus category, sensor type, and sequence length. MSE values are reported in 10^{-5} units. Aggregated baselines (in bold) are also included.

CV	Take	Webcam				ZED				MoCap			
		Seq Len = 1		Seq Len = 5		Seq Len = 1		Seq Len = 5		Seq Len = 1		Seq Len = 5	
		MSE	R ²	MSE	R ²	MSE	R ²	MSE	R ²	MSE	R ²	MSE	R ²
Pitch	S_G	6.33 ± 2.90	0.993 ± 0.004	5.41 ± 1.97	0.994 ± 0.002	2.95 ± 0.24	0.996 ± 0.000	2.21 ± 0.82	0.997 ± 0.001	0.92 ± 0.15	0.999 ± 0.000	0.80 ± 0.27	0.999 ± 0.000
	F_G	29.37 ± 4.30	0.963 ± 0.004	27.20 ± 5.30	0.966 ± 0.005	12.55 ± 15.80	0.978 ± 0.027	4.87 ± 1.80	0.991 ± 0.003	1.35 ± 0.36	0.998 ± 0.001	1.20 ± 0.19	0.998 ± 0.000
	S_Sc	31.17 ± 4.29	0.927 ± 0.009	24.77 ± 6.36	0.942 ± 0.014	2.70 ± 0.26	0.994 ± 0.001	2.64 ± 0.34	0.994 ± 0.001	0.88 ± 0.15	0.998 ± 0.000	0.56 ± 0.10	0.999 ± 0.000
	F_Sc	39.92 ± 4.67	0.891 ± 0.018	37.15 ± 13.77	0.896 ± 0.051	2.84 ± 0.42	0.991 ± 0.001	3.03 ± 0.59	0.990 ± 0.002	1.97 ± 0.50	0.994 ± 0.002	0.59 ± 0.10	0.998 ± 0.000
	FP	137.23 ± 19.24	0.891 ± 0.014	98.87 ± 31.43	0.923 ± 0.019	8.08 ± 0.95	0.993 ± 0.001	6.63 ± 1.18	0.994 ± 0.001	6.34 ± 0.47	0.995 ± 0.000	2.92 ± 0.29	0.997 ± 0.000
	Agg	61.87 ± 3.09	0.931 ± 0.004	38.03 ± 5.32	0.958 ± 0.005	6.08 ± 0.39	0.991 ± 0.001	4.03 ± 0.65	0.994 ± 0.001	2.18 ± 0.09	0.997 ± 0.000	0.97 ± 0.07	0.999 ± 0.000
	B_Agg	86.93 ± 4.00	0.904 ± 0.006	88.55 ± 6.58	0.902 ± 0.008	12.33 ± 1.35	0.981 ± 0.002	12.92 ± 1.61	0.980 ± 0.002	7.52 ± 1.03	0.989 ± 0.002	6.77 ± 1.54	0.990 ± 0.002
Volume	S_A	33.00 ± 11.94	0.950 ± 0.019	12.04 ± 4.56	0.982 ± 0.006	1.24 ± 0.40	0.998 ± 0.001	1.45 ± 0.98	0.998 ± 0.002	1.77 ± 0.43	0.997 ± 0.001	1.16 ± 0.31	0.998 ± 0.001
	F_A	32.55 ± 7.28	0.945 ± 0.013	16.20 ± 5.11	0.973 ± 0.009	1.64 ± 0.35	0.997 ± 0.001	1.56 ± 0.85	0.997 ± 0.001	5.13 ± 0.66	0.991 ± 0.001	1.52 ± 0.39	0.997 ± 0.001
	ST	89.02 ± 13.45	0.566 ± 0.052	41.19 ± 8.45	0.800 ± 0.040	1.82 ± 0.20	0.981 ± 0.003	2.81 ± 1.16	0.972 ± 0.009	13.67 ± 2.70	0.935 ± 0.014	1.13 ± 0.18	0.995 ± 0.001
	FP	13.97 ± 2.55	0.959 ± 0.007	9.14 ± 1.31	0.973 ± 0.004	2.24 ± 0.30	0.994 ± 0.001	2.46 ± 1.03	0.993 ± 0.003	14.49 ± 1.93	0.961 ± 0.005	1.50 ± 0.13	0.996 ± 0.000
	Agg	62.26 ± 5.01	0.877 ± 0.012	24.38 ± 2.90	0.952 ± 0.005	2.76 ± 0.26	0.994 ± 0.001	2.36 ± 0.29	0.995 ± 0.001	10.04 ± 0.29	0.980 ± 0.000	1.38 ± 0.10	0.997 ± 0.000
	B_Agg	121.42 ± 6.38	0.761 ± 0.014	111.49 ± 10.19	0.781 ± 0.020	6.11 ± 1.19	0.987 ± 0.002	5.68 ± 0.61	0.988 ± 0.001	18.40 ± 1.64	0.963 ± 0.003	3.65 ± 0.45	0.993 ± 0.001

S = Slow; F = Fast; G = Glissando; Sc = Scale; A = Arc; ST = Staccato; FP = Free Play; Agg = Aggregate; B = MLP Baseline.

in fast and free-play material. In these conditions, webcam tracking degrades substantially, whereas the ZED and motion-capture configurations remain overall more stable. Apart from the missing depth information, this is also motivated by the quality of the data: in the case of the webcam, MediaPipe drops tracking more often than with the stereo camera.

The quantitative results reported in Table II confirm this behavior across the different recording conditions. The MLP baseline further contextualizes this result: under matched input lengths, the LSTM is generally more accurate, suggesting that the ordering of recent frames contributes useful information.

In particular, the fact that consumer webcams cannot perform reliably under free play conditions is relevant, as this — trivially — constitutes the typical use-condition of the instrument. Conversely, stereo cameras are much less affected by this transition, narrowing the gap to motion capture, suggesting that the addition of direct depth information and a more stable spatial reference is indeed an important factor for preserving realistic feedback.

Volume estimation reveals a similar issue: smooth arcs are reconstructed accurately, but the staccato take is more challenging due to its fast articulation of all fingers together, particularly — again — for webcam tracking. In this case, the strong improvement obtained by the five-frame models indicates that short-term motion context helps recover articulation dynamics, consistently with the intrinsic inertia of physical gestures [41], [42].

The MoCap yields much better results overall, the only exception being the volume in the single-frame setting. This is informative: as the volume hand was represented by only three markers, a single frame may not contain enough information to disambiguate the relation between hand configuration and CV output. Conversely, short temporal context sharply improves performance. This may suggest that a small number of spatially-precise and temporally-stable markers, placed on task-relevant joints, may be more informative than a denser but noisier landmark representation, especially when the model can exploit previous motion history. This interpretation re-

mains however provisional and confirming that would require a dedicated study varying marker density, placement, and technique.

While reported metrics are effective to evaluate models' performance, they lack interpretability from a musical perspective. Therefore, we converted the normalized test RMSE of the best models (Seq Len = 5) back to physical units — Table III. For pitch, since the considered Theremin follows a 1 V/oct scale over a range of 7V, where one volt corresponds to 1200 cents, the error e_p (in cents c) can be computed as

$$e_p[c] = e_{\text{norm}} \cdot 7 \cdot 1200 = 8400e_{\text{norm}}. \quad (1)$$

For the volume antenna, the error is instead converted back to the physical voltage:

$$e_v[\text{mV}] = 10,000e_{\text{norm}}. \quad (2)$$

It is worth noting that voltage is indeed the Theremin's volume CV output and does not correspond to a fixed acoustic level in dB, as the resulting level depends on the downstream amplification setting.

TABLE III: Physical reconstruction errors for the Seq-Len=5 models. Pitch error is reported in cents (1200 cents = 1 Octave), while volume error is reported in millivolts (mV).

Error	Webcam	ZED	MoCap
Pitch (Cents)	163.5 ± 11.1	53.2 ± 4.2	26.1 ± 0.9
Volume (mV)	155.9 ± 9.5	48.5 ± 3.0	37.1 ± 1.3

These metrics better reveal practical limitations, especially for webcam-based pitch reconstruction. For reference, the just-noticeable difference in pitch is commonly reported in the range of 5–10 cents for trained listeners [44], so the webcam error of 163.5 cents, roughly one and a half semitones, would be heard as a wrong note rather than as imprecision. The ZED and motion-capture errors remain above the threshold but stay within roughly half and a quarter of a semitone respectively, a range a performer can plausibly correct aurally while playing. Still, errors of this magnitude may indeed hinder playability, particularly for fine vibrato.

B. Spatial Comparison

The reconstructed CVs remain strongly related to the spatial displacement of the hands. Figure 6 shows the normalized mean hand coordinates along the three spatial axes overlaid on the aligned pitch CV for the three sensing modalities. In all configurations, one spatial component follows the control signal closely: the horizontal displacement of the hand reaches $|r| = 0.86$ for the webcam and $|r| = 0.93$ for both ZED and MoCap.

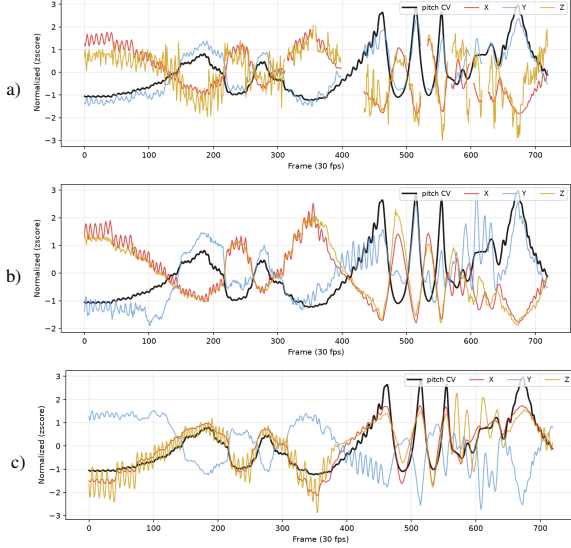


Fig. 6: Mean hand-coordinate trajectories along the x , y , and z axes for a) Webcam, b) ZED, and c) MoCap, overlaid on the simultaneously recorded pitch CV (black).

Table IV compares the recurrent models with the geometric baseline based on a single hand-reference distance. Despite the strong spatial relationship (Fig. 6), a simple distance-based predictor performs consistently worse than the LSTM across sensing configurations, for both pitch and volume. This indeed indicates that proximity alone is insufficient to explain the reconstructed CVs.

TABLE IV: Geometric baselines against the recurrent models for each sensing modality.

CV	Predictor	Webcam	ZED	MoCap
Pitch (cents)	Distance $f(d)$	363.4	196.4	125.4
	LSTM, $L = 1$	208.9	65.5	39.2
	LSTM, $L = 5$	163.5	53.2	26.1
Volume (mV)	Distance $f(d)$	498.4	69.3	194.6
	LSTM, $L = 1$	249.5	52.5	100.2
	LSTM, $L = 5$	155.9	48.5	37.1

The single-frame LSTM already improves substantially over the geometric baseline, indicating that the full hand configuration contains relevant information beyond a single proximity measure. The additional improvement obtained with five-frame sequences further shows the contribution of short-term temporal context. The difference is smaller for ZED volume,

where changes in finger aperture also shift the centroid of the tracked landmarks, causing the distance measure itself to encode part of the hand configuration.

C. Latency Analysis

Finally, Table V reports the latency breakdown for the three sensing configurations. The webcam and ZED pipelines show comparable total latencies, around 32 ms; in both cases, acquisition and visual processing account for most of the delay, while model inference contributes only about 1.35 ms. The MoCap latency is estimated at ca. 2.78 ms from the 360 Hz sampling interval plus a negligible local network transmission delay (already accounted in the table) and model inference, suggesting that the main bottleneck is indeed sensing.

TABLE V: Inference time latency breakdown (in ms). Overall system performance is camera-bound by the data stream acquisition, rather than being CPU or GPU-bound during processing and inference.

Stage	Webcam	ZED	MoCap
Data stream	19.40 ± 10.54	20.24 ± 5.08	2.78 ± 0.34
Processing	11.90 ± 3.26	11.23 ± 2.94	N/A
Model	1.35 ± 0.21	1.35 ± 0.15	0.69 ± 0.05
Total	32.65 ± 14.01	32.82 ± 8.17	3.47 ± 0.39

For IoS deployments, additional network latency would therefore accumulate on top of an already camera-bound control pipeline. In this context, performing visual processing and CV reconstruction locally — transmitting only the resulting control stream — may represent a practical solution for networked applications, reducing the amount of exchanged data [7].

D. Preliminary Evaluation

Despite not providing an exhaustive evaluation with different users and settings, we still wanted to briefly test a prototypical online version of the pipeline as a preliminary deployment check, in order to verify to which extent the instrument could behave like a real Theremin. The evaluation was carried out informally by the last author, using the same instrument, setup, and gestural configuration employed to train the models.

Thus, we implemented the full pipeline for both the ZED and MoCap sensing configurations and check that the reconstructed control signals can indeed drive an audio process in real-time. To emulate a traditional Theremin-like sound engine, we implement a simple SuperCollider⁶ script containing a sinusoidal oscillator with pitch and volume arguments, each controlled independently from the two model outputs, fed directly via Open Sound Control (OSC)⁷ protocol — compatible with common networked-audio workflows. The normalized model outputs are mapped back to the original control ranges.

⁶<https://supercollider.github.io/>

⁷<https://www.cnmat.berkeley.edu/opensoundcontrol>

These preliminary tests confirm that the proposed systems can operate as complete real-time interaction pipelines. In both sensing configurations, the reconstructed control signals proved robust for directly controlling the synthesis engine, yielding an interaction that reproduces the basic gestural paradigm of a conventional Theremin, and provides an initial indication that the learned mappings remain meaningful when embedded in an actual interactive audio application. Nevertheless, the evaluation is qualitative and limited to the original experimental setting and performer; a systematic assessment of playability, responsiveness, and generalization across performers and setups would require a dedicated user study (see Sec. V-B).

Readers interested in experimenting with the prototype can access the full pipeline (train, inference, and pretrained models), along with a short video demo of the optical Theremin being played, at the accompanying repository: <https://github.com/Sm3r/Optical-Theremin>.

V. DISCUSSION

The present results should be first positioned within existing Theremin-inspired/mid-air musical interfaces. As said, previous systems have mostly focused on augmenting the physical Theremin with visual or mixed-reality guidance, or on designing optical/digital reinterpretations that map hand position to sound in more accessible and playful ways [12]–[18]. While these works are relevant for interaction design, learning, accessibility, and musical prototyping, they do not usually treat the problem as one of reconstructing the continuous and fine-grained pitch and volume control signals produced by a trained performer on a real Theremin. This distinction is central to our contribution. An optical Theremin can be (re)designed around the limitations of the sensor: pitch may be quantized or smoothed, the interaction may be simplified, or the mapping may be redefined [15], [17]. In the present work, on the contrary, the target is *precisely* the original CV behavior. The system is therefore not evaluated only as an interface, but as an attempt to recover a demanding continuous control signal from visual motion. This arguably moves the task closer to system identification [45] than to conventional gesture recognition. The results also reinforce a pivotal point in DMI research, which is, the mapping between gesture and sound is not a neutral implementation layer, but a component through which the instrument is defined [8], [29]. In the case of the Theremin, this dependency is clearly exposed. The absence of tactile reference points, the continuous pitch field, and the reliance on auditory and proprioceptive feedback make small gestural deviations musically relevant [11], [32]. From this perspective, the three sensing configurations reveal different compromises upon mere reconstruction-quality hierarchy: the webcam pipeline is inexpensive and easy to deploy, but remains fragile when the task approaches real performance conditions. Motion capture provides the most stable measurement overall, but is intrusive, extremely expensive, and impractical for ordinary musical use. The ZED configuration occupies an intermediate position, as it narrows the gap

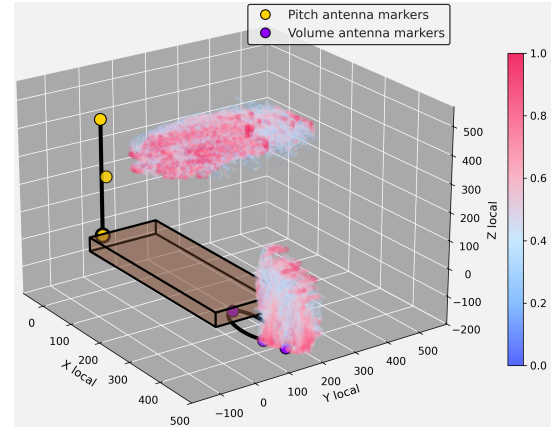


Fig. 7: Aggregated 3D hand-locus density around the Theremin antennas. The point cloud shows the spatial density of hand-marker positions across all recordings.

to MoCap-level reconstruction while remaining closer to a deployable optical setup. Finally, taken together, these results suggest that reconstruction accuracy should be considered a necessary but not sufficient criterion for evaluating optical Theremin systems, as their effectiveness as DMIs ultimately also depends on how the resulting control behavior supports musical practice and interaction [46].

A. Issues and Challenges

The main methodological issue emerging from these experiments is that the model does not learn a “universal” Theremin field, rather a relation between visible hand trajectories and CV outputs under a specific calibration and within a spatial region actually explored by the performer — Fig. 7. On the one hand, this is musically reasonable, because a playable instrument needs high accuracy under playing constraints imposed by technique and praxis [39]. On the other, it limits extrapolation, as modification in body posture, register, distance from the antennas, camera viewpoint, or technique may move the performer outside the learned space. Thus, even when held-out temporal segments are reconstructed accurately, this should not be interpreted as evidence that the model has learned a performer- or setup-independent Theremin mapping: instead, our method overfits on a specific setting, and adapting to new hardware/players would require retraining and recalibration.

A second challenge is the ambiguity of the inverse problem. Theremins do not strictly measure hand pose: they respond to the electromagnetic interaction between (mostly) the performer and the antennas. Consequently, different hand configurations may indeed correspond to similar CVs, not to mention potential edge-cases such as recalibration or environmental conditions. This ambiguity is partly controlled in the present protocol — again, we overfit on a single instrument/setting/performer/environment configuration — but it becomes relevant for free play and for any deployment involving different performers or changing setups.

A third challenge concerns the conversion of tracking artifacts into musical artifacts. Optical trackers introduce discontinuities through occlusion, label instability, depth spikes, and jitter. In Theremin-like control, these factors are immediately converted into continuous sound behavior. While these problems can be partially mitigated by careful sensor placement and heavy data preprocessing, these solutions are not always feasible from a physical (e.g., close cameras can interfere with the antennas during tracking) and temporal (e.g., real-time constraints) perspective. Therefore, tracking systems need to be calibrated under best compromise conditions. Such a compromise is also relevant whenever such a controller is used beyond a self-contained local instrument. In networked or IoS-oriented musical systems, for instance, continuous control streams are often discussed in relation to transmission latency and synchronization [10], [47]; however, our experiments suggest that, for vision-based mid-air control, part of the instability is already introduced at the sensing and tracking stages. In the current local setup, the camera-based pipelines already operate close to the perceptual threshold reported for gesture-controlled musical instruments [33]. Thus, even before considering remote transmission, the control stream may already contain timing, confidence, and calibration-dependent uncertainties that affect playability. Consequently, in IoS settings, communication latency and jitter should be considered jointly with sensing and tracking uncertainty [47].

B. Limitations and Future Work

Our experimental setting is restricted to a single performer, instrument model, and controlled setup. Although the main scope of this work is to explore whether the fine-grained behavior of a real analog Theremin can be optically reconstructed under realistic performance configurations, we still acknowledge such a constraint as a limitation. Accordingly, the reported results characterize generalization to unseen temporal segments within the same performer, instrument, calibration, and sensing configuration, rather than across different performers or setups. Indeed, further evaluation will involve different Thereminists, instrument calibrations, camera placements, and settings; this will also allow us to gather further insights on musical usages and potential deployments in real-world scenarios.

These additional evaluations will constitute the main direction for future developments of this work, together with a dedicated assessment of playability and perceptual responsiveness under different reconstruction errors and end-to-end latency conditions.

VI. CONCLUSIONS

In this paper, we investigated the issues and challenges of reconstructing Theremin pitch and volume CVs from visual hand trajectories under real-time constraints. We framed the task as a constrained control-estimation problem and compared monocular vision, stereo/depth-assisted vision, and marker-based motion capture under a shared protocol. The results show that accurate CV reconstruction is possible in controlled

conditions, but remains dependent on calibration, performer technique, spatial coverage, temporal alignment, and latency. While we provided controlled baselines over several sensing setups, findings indicate that future optical Theremin-based systems should be evaluated not only by numerical accuracy, but also by the sensing and interaction failures that affect musical playability.

ACKNOWLEDGMENTS

We would like to sincerely thank Francesco Manenti for the intuition, design, and physical production of the custom 3D-printed supports for the MoCap markers that allowed us to overcome relevant problems during simultaneous tracking.

REFERENCES

- [1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.
- [2] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.
- [3] D. Cocchiara and L. Turchet, "Democratizing access to collaborative music making over the network using air instruments," in *Proceedings of the 17th International Audio Mostly Conference*, 2022, pp. 211–218.
- [4] Y. Yang, Z. Wang, and Z. Li, "Mugevi: A multi-functional gesture-controlled virtual instrument," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2023.
- [5] M. J. Lyons, M. Haehnel, and N. Tetsutani, "Designing, playing, and performing with a vision-based mouth interface," in *A NIME Reader: Fifteen Years of New Interfaces for Musical Expression*. Springer, 2017, pp. 107–124.
- [6] M. Lim, N. Kotsani, and P. Hartono, "Handmate: An accessible browser-based controller for web audio and midi using ai hand-tracking," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, June 2022.
- [7] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.
- [8] A. Hunt, M. M. Wanderley, and M. Paradis, "The importance of parameter mapping in electronic instrument design," *Journal of New Music Research*, vol. 32, no. 4, pp. 429–440, 2003.
- [9] D. Brown, C. Nash, and T. Mitchell, "Simple mappings, expressive movement: a qualitative investigation into the end-user mapping design of experienced mid-air musicians," *Digital Creativity*, vol. 29, no. 2-3, pp. 129–148, 2018.
- [10] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
- [11] X. Xiao and S. F. Alaoui, "Tuning in to intangibility: Reflections from my first 3 years of theremin learning," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 2649–2659.
- [12] L. Hasan, N. Yu, and J. A. Paradiso, "The termenova: A hybrid free-gesture interface," in *Proceedings of the 2002 conference on New Interfaces for Musical Expression*, 2002, pp. 1–6.
- [13] D. Johnson, D. Damian, and G. Tzanetakis, "Evaluating the effectiveness of mixed reality music instrument learning with the theremin," *Virtual Reality*, vol. 24, no. 2, pp. 303–317, 2020.
- [14] D. Johnson and G. Tzanetakis, "Vrmin: using mixed reality to augment the theremin for musical tutoring," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2017, pp. 151–156.
- [15] M. Ramos dos Santos and A. E. Coca Salazar, "Optical digital theremin with audio synthesis and graphic interface," *International Journal of Circuits, Systems and Signal Processing*, 2021.

- [16] B. Usabaev, A. Eschenbacher, and A. Brennecke, "The virtual theremin: designing an interactive digital music instrument for film scene scoring," *i-com*, vol. 21, no. 1, pp. 109–121, 2022.
- [17] A. Sourin, "Music in the air with leap motion controller," in *2019 International Conference on Cyberworlds (CW)*. IEEE, 2019, pp. 57–60.
- [18] M. Xing and A. Sourin, "Sound from the air by optical tracking in real time," in *2025 International Conference on Cyberworlds (CW)*. IEEE, 2025, pp. 63–69.
- [19] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1145–1153.
- [20] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4903–4911.
- [21] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," *ACM Transactions on Graphics*, vol. 36, no. 6, Nov. 2017.
- [22] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 813–822.
- [23] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol *et al.*, "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 7061–7071.
- [24] P. Ren, J. Wang, H. Sun, Q. Qi, X. Liu, M. Zhang, L. Zhang, J. Wang, and J. Liao, "Prior-aware dynamic temporal modeling framework for sequential 3d hand pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6476–6487.
- [25] Y. Wang, H. Xu, P.-A. Heng, and C.-W. Fu, "Unihope: A unified approach for hand-only and hand-object pose estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 231–12 241.
- [26] S. Liu, Z. Li, X. Tang, and J. Liu, "Direction-aware hybrid representation learning for 3d hand pose and shape estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 251–260.
- [27] A. Peñalba, M.-J. Valles, E. Partesotti, M.-Á. Sevillano, and R. Castañón, "Accessibility and participation in the use of an inclusive musical instrument: The case of motioncomposer," *Journal of Music, Technology & Education*, vol. 12, no. 1, pp. 79–94, 2019.
- [28] I. Hwang, H. Son, and J. R. Kim, "Airpiano: Enhancing music playing experience in virtual reality with mid-air haptic feedback," in *2017 IEEE World Haptics Conference (WHC)*, 2017, pp. 213–218.
- [29] D. Wessel and M. Wright, "Problems and prospects for intimate musical control of computers," *Computer Music Journal*, vol. 26, no. 3, pp. 11–22, 2002.
- [30] W. Lages, M. Nabiyouni, J. Tibau, and D. A. Bowman, "Interval player: Designing a virtual musical instrument using in-air gestures," in *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2015, pp. 203–204.
- [31] J. Cui and A. Sourin, "Mid-air interaction with optical tracking for 3d modeling," *Computers & Graphics*, vol. 74, pp. 1–11, 2018.
- [32] K. D. Skeldon, L. M. Reid, V. McNally, B. Dougan, and C. Fulton, "Physics of the theremin," *American Journal of Physics*, vol. 66, no. 11, pp. 945–955, 1998.
- [33] T. Mäki-Patola and P. Hämäläinen, "Latency tolerance for gesture controlled continuous sound instrument without tactile feedback," in *International Computer Music Conference (ICMC)*, 2004.
- [34] X. Xiao, G. Locqueville, C. d'Alessandro, and B. Doval, "T-voks: the singing and speaking theremin," in *NIME 2019 International Conference on New Interfaces for Musical Expression (NIME)*, 2019, pp. 110–115.
- [35] J. García-Guzmán, D. Salazar-Pérez, G. Celis-Ronzón, and J. A. Vélez-Enríquez, "Design and experimental prototype testing of a hybrid digital theremin system," in *2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 2018, pp. 1–4.
- [36] H. Bello, B. Zhou, S. Suh, L. A. Sanchez Marin, and P. Lukowicz, "Move with the theremin: Body posture and gesture recognition using the theremin in loose-garment with embedded textile cables as antennas," *Frontiers in Computer Science*, vol. 4, p. 915280, 2022.
- [37] R. Hermawan, "Dancing with theremins," Master's thesis, Department of Information Technology, Uppsala University, Nov. 2018.
- [38] C. Liu, Y. Huang, M. Zhang, Y. Ma, W. Yu, J. Wu, L. Li, Q. Ma, and T. J. Cui, "Smart theremin based on electromagnetic metasurfaces for fine dual-hand perception," *Advanced Optical Materials*, vol. 13, no. 31, p. e01899, 2025.
- [39] C. Eyck, *The Art of Playing the Theremin*. Berlin, Germany: SERVI Verlag, 2006.
- [40] S. Biswas, A. Nandy, A. K. Naskar, and R. Saw, "Mediapipe with lstm architecture for real-time hand gesture recognition," in *International Conference on Computer Vision and Image Processing*. Springer, 2023, pp. 422–431.
- [41] P. Morasso, "Spatial control of arm movements," *Experimental Brain Research*, vol. 42, no. 2, pp. 223–227, 1981.
- [42] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, "Weakly supervised 3d hand pose estimation via biomechanical constraints," in *European Conference on Computer Vision*. Springer, 2020, pp. 211–228.
- [43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [44] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Brill, 2012.
- [45] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.
- [46] S. O'Modhrain, "A framework for the evaluation of digital musical instruments," *Computer Music Journal*, vol. 35, no. 1, pp. 28–42, 2011.
- [47] L. Turchet and P. Casari, "Latency and reliability analysis of a 5g-enabled internet of musical things system," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 1228–1240, 2023.