INTERFACE

royalsocietypublishing.org/rsif

Research

Cite this article: Azadjou H, Marjaninejad A, Valero-Cuevas FJ. 2026 Perception in action: a robotic system that can teach itself to melodiously play music by ear. *J. R. Soc. Interface* **23**: 20250909.
https://doi.org/10.1098/rsif.2025.0909

Received: 15 September 2025
Accepted: 18 February 2026

Subject Category:
Life Sciences–Engineering interface

Subject Areas:
biomechanics, biomedical engineering

Keywords:
perceptual learning, convolutional neural networks, multi-layer perceptron, perception–action loop, deep learning, tendon-driven robot

Author for correspondence:
Francisco J. Valero-Cuevas
e-mail: valero@usc.edu

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8469590.

THE ROYAL SOCIETY PUBLISHING

Perception in action: a robotic system that can teach itself to melodiously play music by ear

Hesam Azadjou[1], Ali Marjaninejad[1] and Francisco J. Valero-Cuevas[1,2]

[1]Alfred E. Mann Department of Biomedical Engineering, and [2]Division of Biokinesiology and Physical Therapy, University of Southern California, Los Angeles, CA, USA

HA, 0000-0002-9596-5539; AM, 0000-0001-6943-151X; FJV-C, 0000-0002-2611-7923

Learning to imitate nuanced motor behaviour by perceiving someone's actions is a human ability central to non-verbal communication, interaction and collaboration. Despite extensive psychophysics research, robotic systems do not yet seamlessly bridge the gap between perception and action. We demonstrate a perceptual learning algorithm that can replicate a melody after hearing it once by using a physical robotic hand on a keyboard. Importantly, this system (the Musician Hand) can do this after only 2 min of random 'motor babbling', and driven only by its own experience, with minimal pre-defined features of musical notes. Experiments with three melodies showed that our system can play-by-ear at a level comparable with four trained pianists, and better than five novices. This demonstrates how perception can seamlessly drive nuanced robotic action based only on its own limited experience. This demonstration of a perception-driven system paves the way for human-friendly and intuitive systems for entertainment, collaboration and physical assistance/augmentation.

1. Introduction

Traditional robotics relies on control theory, pre-programmed actions, trial-and-error reinforcement learning or precise models of the physical system, task and/or environment. Alternatively, data-driven or empirical approaches often use imitation/demonstration of a task, trial-and-error in hardware or extensive training in simulation [1–10]. By contrast, animals are the best examples we have of motor performance. In them, motor learning, behaviour and adaptation are intricately intertwined with perception, and are the foundation of human development and artistic expression. This fundamental notion is well-established in psychological and psycho-physical studies in humans and animals [11–17]. Thus, we and others have argued that advances in machine learning should be complemented by biological strategies to achieve effective learning, performance and adaptation in the physical world [18,19].

Perception, in particular, plays a critical role in shaping the coordination, stability and organization of motor actions [15,20–31]. Perception fundamentally contextualizes and attributes meaning to sensations to produce meaningful actions. It is a process built upon experience and interactions with the physical world that creates a *Sensorimotor Gestalt* (an organized whole that is perceived as being more than the sum of its individual components) of compatible sensory sets and motor actions (i.e. the perception–action link) [32–36]. These ideas have been proposed as potential foundations of identity, agency and self for robots [37]. Therefore, we developed a *perceptual learning* algorithm that imitates the psychophysical ability of humans to convert sounds into actions, and tested it on a *Musician Hand*.

While traditional engineering methods, including reinforcement learning, have demonstrated success in piano playing [38–40], our approach focuses on

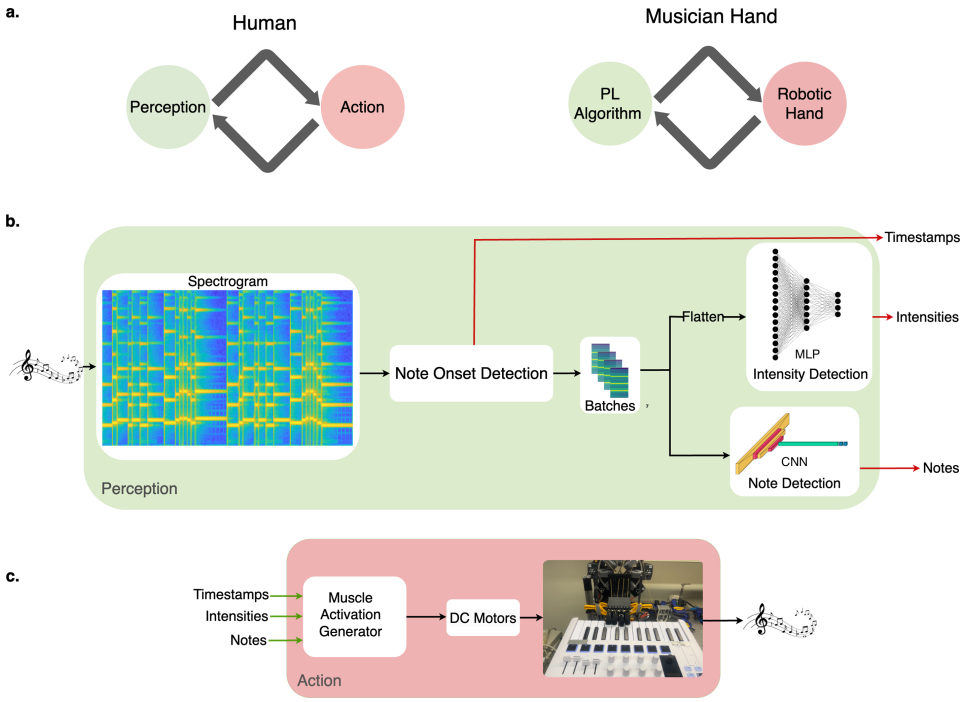

Figure 1. Overview of the Musician Hand. (a) The perception–action loop is a fundamental principle of biological behaviour, especially in artistic expression. (b) Our algorithm takes a melody (sensed as a spectrogram) and extracts batches of 'percepts' (sequences of notes and their intensity) on the basis of prior 'motor babbling' experience (see figure 2). (c) The melody is replicated by playing the notes on an actual keyboard by the four fingers of a tendon-driven robotic hand.

using perceptual representations to reproduce melodies with nuanced timing and dynamics after exposure to a single example (a version of one-shot learning [41]). Unlike traditional robotics approaches that rely on error signals (e.g. trial-and-error learning or error-driven control [10,42–47]), the perceptual learning algorithm melodiously replicates melodies in a feed-forward way by leveraging knowledge acquired solely during its own 'motor babbling' phase.

We show the performance of the Musician Hand (i.e. our perceptual learning algorithm coupled to a robotic hand) to be better than human novice players, and comparable with trained pianists who were asked to play the same melodies by ear—as per quantitative measures of precision and recall, and qualitative scores and rankings by blinded musical experts. This demonstrates that our perceptual end-to-end pipeline (the Musician Hand) succeeds at emulating the human ability to play-by-ear.

2. Methods

The Musician Hand consists of a perceptual learning algorithm coupled to a robotic hand, as described below.

2.1. Hardware

The tendon-driven, four-finger hand (figure 1) was designed with a lightweight architecture optimized for additive manufacturing (three-dimensional printing). The modular fingers and the hand palm were engineered for easy interchangeability, a critical feature given the potential for part failures. Failures arose when mechanical stresses surpassed the tensile strength of components or when motor temperatures exceeded the glass transition point of poly-lactic acid, causing the three-dimensional-printed motor mounts to warp under load. To ensure stability, hard stops at the palm restricted joint angles, preventing the collapse of the end effectors' feasible force spaces [48]. This design helps in avoiding singularities in the cost landscape, ensuring continuous functionality. We utilized four identical springs to passively return the fingers to a neutral position once the DC motor returned to its minimum activation level, inspired by muscle tone to improve responsiveness to sudden force changes. The electromechanical system for sensing and tendon actuation leveraged extensive prior work on locomotion and movement in tendon-driven systems [10,49]. Each finger's inelastic nylon tendon was actuated by Faulhaber® 2342-S024-CR gear-free brushed DC motors. The gear-free design, coupled with a small tendon coil on each motor shaft, ensured back-drivability. Motor torque was modulated using voltage commands sent to Western Servo Design LDU-S1 linear current amplifiers, which controlled the DC current.

2.2. Dynamic control of tendon-driven fingers

Our system controls four tendon-driven fingers to play piano keys using four DC motors. Each motor pulls a tendon, adjusting the corresponding finger's position and velocity. The control strategy dynamically activates the motors based on the output of an artificial neural network (ANN) that classifies musical notes and their intensities at different time points. Each finger F_i *is* modelled

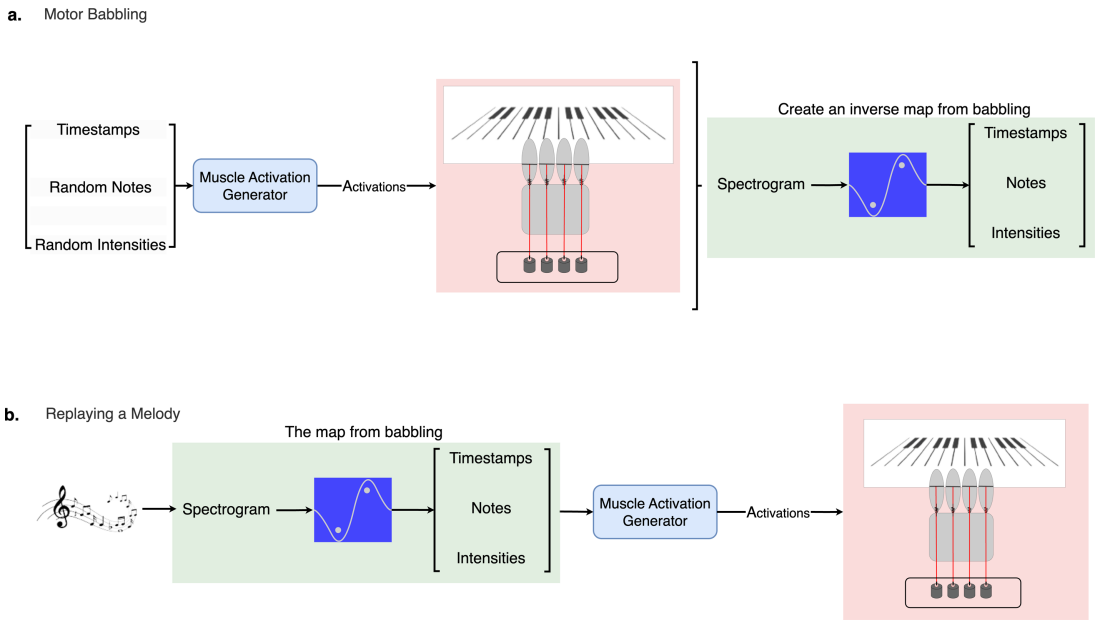

Figure 2. Motor babbling and how it is used to play a melody 'by ear.' (a) The algorithm starts from a naive state and, like a novice human, randomly explores the mapping from finger actions to percepts (i.e. note sequences and intensities). We implemented the motor babbling by recording 2 min of random sequences of individual finger actions, each lasting 500 ms. We used these input–output pairings to train artificial neural networks (ANNs) to detect percepts, to then produce finger actions that replicate the melody. (b) This allows the *Musician Hand* to play any arbitrary melody that includes the notes experienced on the keyboard.

as a second-order system:

$$\ddot{q}_i = -I(q_i)^{-1}c(q_i,\dot{q}_i) + b\dot{q}_i + I(q_i)^{-1}T, \tag{2.1}$$

where q_i, $\dot{q}_i$ and $\ddot{q}_i$ are the joint angle and its first and second derivatives, respectively, in finger i, I is the inertia, c is the Coriolis and centripetal force, b is the joint friction coefficient, k is the stiffness and $T_i(t)$ *is the* torque in finger i generated by M_i. The musculotendon forces (represented here as cables pulled by the motors) are subsequently correlated with the vector of applied joint torques:

$$T_i(t) = r(q_i)F_0 a_i, \tag{2.2}$$

where r is the moment arm of each finger, F_0 is the maximal force exerted by each motor and a is the normalized actuation of the motor. In order to predict the actuation values of each motor to play the melody, we use a mapping that connects a melody's spectrogram to actuation time series:

$$A(t) = [a_1(t), a_2(t), a_3(t), a_4(t)] = \psi[X(t,f)], \tag{2.3}$$

where ψ is a mapping consisting of an energy function, a convolutional neural network (CNN), a multi-layer perceptron (MLP) and a motor activation generator that uses $X(t,f)$, the short-time Fourier transform (STFT) of the input audio signal:

$$\left[x_p(t,f), t_p\right] = E[X(t,f)], \tag{2.4}$$

$$N_p = CNN\left[x_p(t,f)\right], \tag{2.5}$$

$$I_p = MLP\left[x_p(t,f)\right], \tag{2.6}$$

$$A(t) = G\left[N_p, I_p, t_p\right], \tag{2.7}$$

where E is the energy function that extracts the timestamps for the note onsets (t_p) and batches of spectrograms that start from the timestamps (x_p). The CNN and MLP then use these batches to extract notes and their intensities (N_p,I_p) and the motor function (G) uses an exponential approximation to muscle activation–contraction dynamics [50–52] (see figure 3) to generate a motor activation sequence that leads the Musician Hand to better approximate melodious playing as a human hand would. We used four different intensity plateaus, K_{1-4}, and time constants, τ_{1-4}. This produced four different key-press mechanics leading to different note onset, loudness and decay. The mapping adjusts the voltage applied to each motor, ensuring the fingers achieve the desired movements and note intensities in real time, as dictated by the percepts from the ANN.

2.3. Time–frequency representation of melodies

We employed the STFT to extract spectrograms from the audio signals of the melodies under investigation. The STFT is a widely utilized signal processing technique for analysing the frequency content of non-stationary signals over short time intervals. First, each audio signal $x(t)$ was divided into short overlapping segments. Let $x_n(t)$ represent the nth segment of the audio signal. The

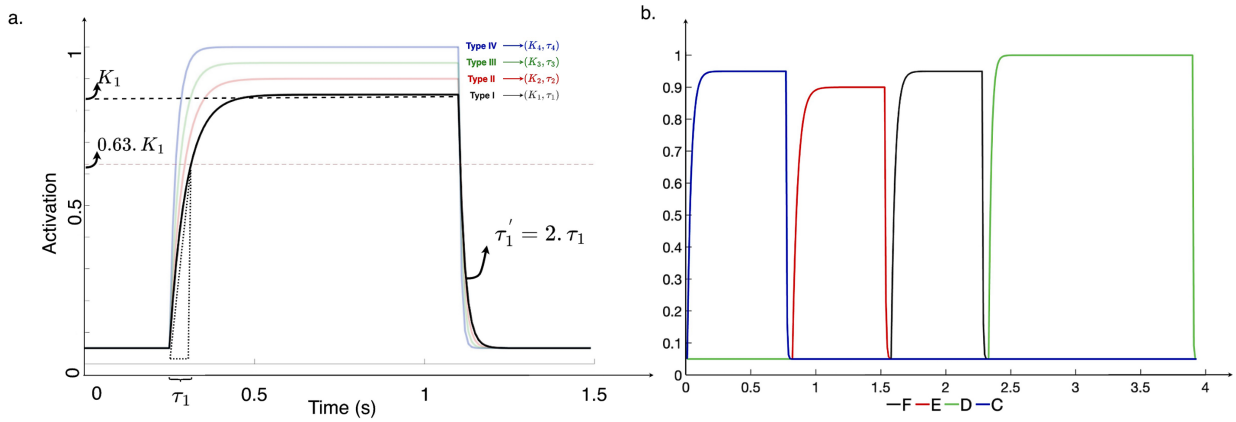

Figure 3. (a) The four motor activation profiles used by the Musician Hand: we pre-programmed four activation profiles to replay the melodies. The motor function uses an exponential approximation to muscle activation–contraction dynamics [50–52]. Each profile started at 5% of the maximum activation (to replicate muscle tone, and maintain posture and readiness for action) and reached four different intensity plateaus, K_{1-4}, and time constants, τ_{1-4}. This produced four different key-press mechanics leading to different note onset, loudness and decay. (b) Motor activation sequences: in this figure, we see an example of the activation sequence sent to the motors to pull the tendons and press the keys.

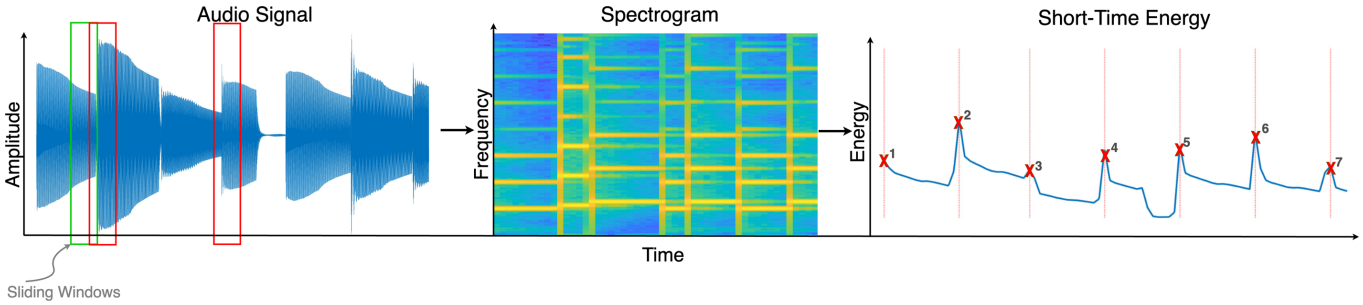

Figure 4. Signal processing transforms the percept of a melody into signals that can train the ANNs. As shown in figure 1*b*, we extract the time–frequency representation (spectrogram) of the audio signals using sliding windows that compute the Fourier transform of the signal in short overlapping time intervals. As STFT computes the signal's power in different frequencies for short time intervals, we integrate the spectrogram on the frequency axis to compute the short-time energy of the audio signals that represent the notes onsets as their peaks (the red windows that cover the second and the third note's onsets are the second and third peaks in the short-time energy).

Fourier transform was then applied to each segment to transform it from the time domain to the frequency domain. This process is mathematically defined as:

$$\mathrm{STFT}\{x(t)\}(t,f) = X(t,f) = \int_{-\infty}^{\infty} x(\tau)\, w(t-\tau)\, e^{-j2\pi f\tau}\, d\tau, \tag{2.8}$$

where $w(t)$ is a window function that is non-zero for a short duration and $X(t,f)$ represents the STFT of the signal $x(t)$, giving the frequency content of the signal at time t. Repeating this process for all segments and stacking the resulting frequency spectra over time, we extracted a time–frequency representation of the melody's audio signal, commonly known as a spectrogram. The spectrogram $S(t,f)$ is given by:

$$S(t,f) = \left|X(t,f)\right|^2. \tag{2.9}$$

This spectrogram provided valuable insight into the distribution of frequencies present in the melody at different points in time. We used the spectrogram to extract the notes and their intensities to automate the Musician Hand (figure 4).

2.4. Energy function of melodies

We integrated the spectrograms along the frequency axis to derive the energy function of each melody. Let $S(t,f)$ represent the spectrogram of the audio signal, where t denotes time and f denotes frequency. The energy function $E(t)$ at time t is obtained by integrating $S(t,f)$ over all frequencies:

$$E(t) = \int_{f_{\min}}^{f_{\max}} S(t,f)\, df. \tag{2.10}$$

This integration process aggregates the spectral energy across different frequency bands, providing a comprehensive measure of the overall energy content of the audio signal at each time step t. By finding the peaks of these energy functions [53], we identified significant increases in energy corresponding to the onsets of musical notes within the melody. Importantly, this is not a note onset detection process manufactured by us, but rather a direct analogy to the auditory temporal processing system that dissociates

sound-onset versus sound-offset (gap-detection) in mammals [54]. Let $\mathcal{P}$ be the set of time indices where $E(t)$ has local maxima. These peaks served as reliable indicators of note onsets, allowing for precise localization of the timing at which each note began:

$$\mathcal{P} = \{t_i \mid E(t_i) > E(t) \text{ for all } t \in (t_i - \epsilon, t_i + \epsilon), \epsilon > 0\}, \tag{2.11}$$

where ϵ is a small positive value used to define the neighbourhood around each peak. Thus, the times $t_i \in \mathcal{P}$ *indicate* the note onsets within the melody (figure 4).

2.5. Babbling and training the inverse mapping

Since the system has no prior information on its dynamics, the physics of the environment, topology or structure, it begins by exploring in a general sense through the execution of random control sequences to the motors, a process we refer to as motor babbling. To train the Musician Hand to replay a melody from scratch, we employed motor babbling to learn how to map sensory information (sound) to motor actions (notes played). A run from a naive state started with 2 min of babbling, where a sequence of pseudo-random motor activations (using activation profiles that use an exponential approximation to muscle activation–contraction dynamics, figure 3) were distributed across the four motors controlling the tendons. Each motor action lasted 500 ms and pressed keys at random with varying forces to produce notes of different intensities, figure 2*a*.

2.6. Motor babbling

During this phase, the system executes random control sequences and collects the resulting melody played by the Musician Hand. The MLP and CNN used in the inverse mapping are then trained with this input–output pair to create an inverse map between the system inputs (motor activation levels) and desired system outputs (melody's spectrogram).

2.7. Random activation values for motor babbling

The motor activations (control sequences) during motor babbling were generated by employing two pseudo-random number generators with a uniform distribution. The first generator randomly determines the activation order for each of the four motors (and, consequently, the keys). By contrast, the second generator selects the activation level between 75% and 100%.

2.8. Structure of the artificial neural network

Two neural network architectures were developed to extract the musical percepts from the spectrogram of the melodies: a CNN and an MLP. The CNN architecture consists of two convolutional layers with 32 and 64 filters, each followed by a max-pooling layer for downsampling. Subsequently, the flattened data are passed through two dense layers, each with 64 neurons and rectified linear unit (ReLU) activation functions, before reaching the output layer with four neurons and a softmax activation function. By contrast, the MLP architecture comprises a single dense layer with 32 neurons and ReLU activation functions, followed by the output layer with four neurons and softmax activation. Both models were compiled using the Adam optimizer and categorical cross-entropy loss function. The CNN model was trained for five epochs, while the MLP model was trained for 20 epochs, with validation data used to monitor accuracy. The architectures were implemented using the TensorFlow and Keras libraries in Python. In addition, before training, the spectrogram data were normalized by dividing by the maximum absolute value using NumPy. We used ANNs (CNNs used in parallel with MLPs in this study) to detect musical notes and their intensities from spectrograms, as they offer a powerful approach even in challenging conditions involving noise and artefacts [55,56]. Real-world audio often contains background noise, recording flaws and overlapping harmonics, which can complicate accurate frequency detection. ANNs distinguish signals from noise by learning spatial and temporal patterns in spectrograms, such as harmonics and overtones, that are not explicitly addressed in frequency-based analyses [57,58]. They are also more robust to timbre and dynamic articulation variations, enabling effective note detection across diverse instruments and playing styles [59]. They use harmonic content, combining data across multiple frequencies to emulate human auditory perception.

2.9. Motor activation profile

We pre-programmed four activation profiles to replay the melodies. The motor function uses an exponential approximation to muscle activation–contraction dynamics. Each profile started at 5% of the maximum activation (to replicate muscle tone, and maintain posture and readiness for action) and reached four different intensity plateaus, K_{1-4}, and time constants, τ_{1-4}. This produced four different key-press mechanics leading to different note onset, loudness and decay. In figure 3, we see an example of the activation sequence sent to the motors to pull the tendons and press the keys.

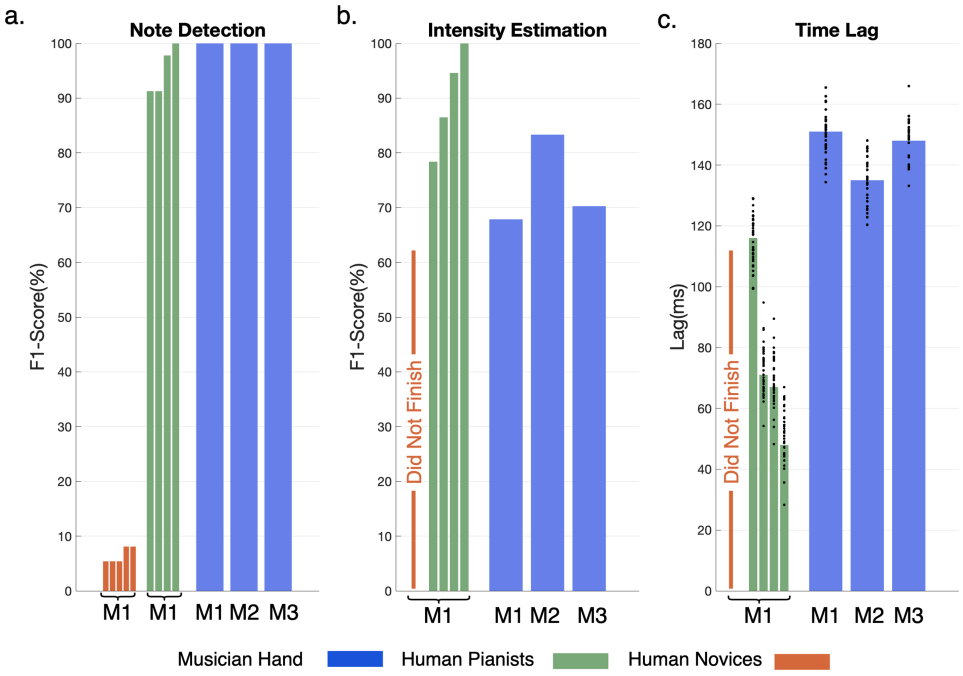

Figure 5. Quantitative results: (a) Note detection (F1 score for the notes played versus the notes in the score): while five human novices stumbled to play melody 1 (M1) by ear (red bars), four human pianists performed well (green bars), and our *Musician Hand* had perfect note detection for a total of three melodies (blue bars, M1–M3). (b) Intensity estimation (F1 score for the key-press intensity): it was scoreless for the human novices (i.e. they failed to replicate the melody). Human pianists scored between 78% and 100%, while the *Musician Hand* scored between 68% and 84%. (c) Onset time lag (smaller is better) quantifies the replication of note onset and reflected rhythm across the melody. Human pianists scored best at below 120 ms. The *Musician Hand* scored between 120 and 170 ms depending on the melody.

2.10. Evaluation metric

To evaluate the accuracy of note detection and intensity estimation, we used the F1 score, a standard metric in machine learning that balances precision and recall in classification tasks. The F1 score is defined as

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{2.12}$$

where *Precision* represents the proportion of correctly identified notes among all detected notes, and *Recall* represents the proportion of correctly identified notes among all target notes. This metric quantifies both the correctness of replayed notes and their proximity in intensity to the original target notes. In addition, we measured the time difference of each note, which captures the temporal lag or advance relative to the onset of the target notes. These measures together provide a comprehensive assessment of performance, as illustrated in figure 5.

2.11. Similarity measurement

We evaluated the similarity of the recordings in two dimensions. First, perceptual similarity was measured using the structural similarity index metric (SSIM) of the spectrograms. SSIM is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2.13}$$

where μ_x and μ_y are the mean values, σ_x^2 and σ_y^2 are the variances and σ_{xy} is the covariance of the spectrograms x and y. Constants C_1 and C_2 are used to stabilize the division. The SSIM of the spectrogram serves as a perception-based measurement, assessing the perceived alteration in structural data. It integrates critical perceptual factors like luminance and contrast masking to measure the similarity between the time–frequency representations of melodies derived from their spectrograms [60].

Second, aural similarity was evaluated through mel-frequency cepstral coefficients (MFCCs) cosine similarity. The MFCCs are computed by first taking the discrete cosine transform of the log power spectrum on a nonlinear mel scale of frequency. The kth MFCC is given by:

$$\text{MFCC}_k = \sum_{n=1}^{N} \log(X_n) \cos\left[k(n - 0.5)\frac{\pi}{N}\right], \tag{2.14}$$

where X_n is the power spectrum of the nth mel-frequency band, and N is the total number of mel-frequency bands. The cosine similarity between two MFCC vectors A and B is then given by:

$$\text{Cosine similarity}(A, B) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{2.15}$$

where $A \cdot B$ is the dot product of vectors A and B, and $\|A\|$ and $\|B\|$ are the magnitudes of the vectors. We extracted 25 MFCCs as all performance metrics in the study by Hasan *et al.* which gave the best score [61].

Finally, the rhythmic similarity metric which focuses on the presence and timing of onsets [62] was calculated. Onsets represent abrupt changes or transients in the audio waveform corresponding to beats or rhythmic events. Onsets are extracted from the audio signals using the energy function. Let the onset times for the original and generated signals be denoted as:

$$\mathbf{T}_{\text{original}} = [t_1, t_2, \ldots, t_{N_o}], \quad \mathbf{T}_{\text{generated}} = [\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_{N_g}],$$

where t_i and $\hat{t}_i$ are the onset times in seconds. The rhythmic similarity is measured using cross-correlation to account for timing differences between the two signals. The cross-correlation is defined as:

$$C(\tau) = \sum_{i=1}^{N_o} \sum_{j=1}^{N_g} \delta(t_i - \hat{t}_j - \tau),$$

where τ is the time lag, and $\delta(\cdot)$ is the Dirac delta function. The rhythmic similarity score is the maximum cross-correlation:

$$\text{Rhythmic similarity} = \max_{\tau} C(\tau).$$

2.12. Blind ranking by human critics

To evaluate the melodic replications, two professional composers (RT and TKM) independently conducted a blind assessment. They ranked the replications according to their perceived similarity to the target melody 1, without access to any identifying labels.

2.13. Study participants

We recruited nine participants for the study, including four trained pianists and five novices. All participants gave their informed written consent to participate in this study. The procedures were approved by the USC IRB protocol USC IRB: HS-17-00304. The five novices had no prior experience with piano playing. They were confirmed to have normal pitch perception through a tone-deafness screening, which involved identifying differences in notes played on piano keys. The group of four pianists consisted of three professionals and one amateur, all with prior training and experience in piano performance.

2.14. Experimental set-up

In this study, we used a four-finger tendon-driven hand, and a DC motor akin to a muscle which provided the mechanical power to the joints. Each finger is driven by an individual tendon actuated by a DC brushless motor. The tendons are designed to passively return to their initial positions with the aid of springs. At the same time, the connection between the muscles and joints, analogous to the tendons in biological systems, is represented by a string referred to as a tendon within our robotic hand. Our modular tendon-driven robotic hand, equipped with DC motors, was managed by a custom data acquisition (DAQ) and control programme on a DAQ-capable computer with the NI-DAQmx driver. Operating on a machine learning-capable computer, our perceptual learning pipeline utilized RT-Bridge over ethernet to transmit commands and receive sensory data from the DAQ programme with a sub-millisecond round-trip latency. To play and record the melodies, we used a musical instrument digital interface (MIDI) device connected to a computer that runs GarageBand, and then the recorded files were exported as wave files to be used by the algorithm in Python.

2.15. Human experiments

Human participants were allowed to babble for 5 min on the relevant keys (without shifting their fingers) and had 3 min to practice melody 1 before being given 1 min to play the 29 s melody (i.e. 'final performance'). To eliminate the need to memorize melody 1, they could navigate melody 1 forward and backward during the practice and performance periods.

2.16. Melodies

For our study, three original melodies were meticulously composed for this research (figure 6). These melodies were deliberately crafted to utilize only four distinct keys: C4, D4, E4 and F4. Each melody was performed at a consistent tempo of 90 beats per minute, executed by our two skilled composers on a Steinway grand piano. This selection of keys and tempo ensured a controlled and uniform musical environment, facilitating precise analysis and comparison within our experimental framework.

3. Results

We demonstrate that the Musician Hand can replay any arbitrary four-note melody without (i) closed-loop error correction, or (ii) an explicit model of the dynamics of the tendon-driven fingers, task or the physical environment (e.g. keyboard inertia or

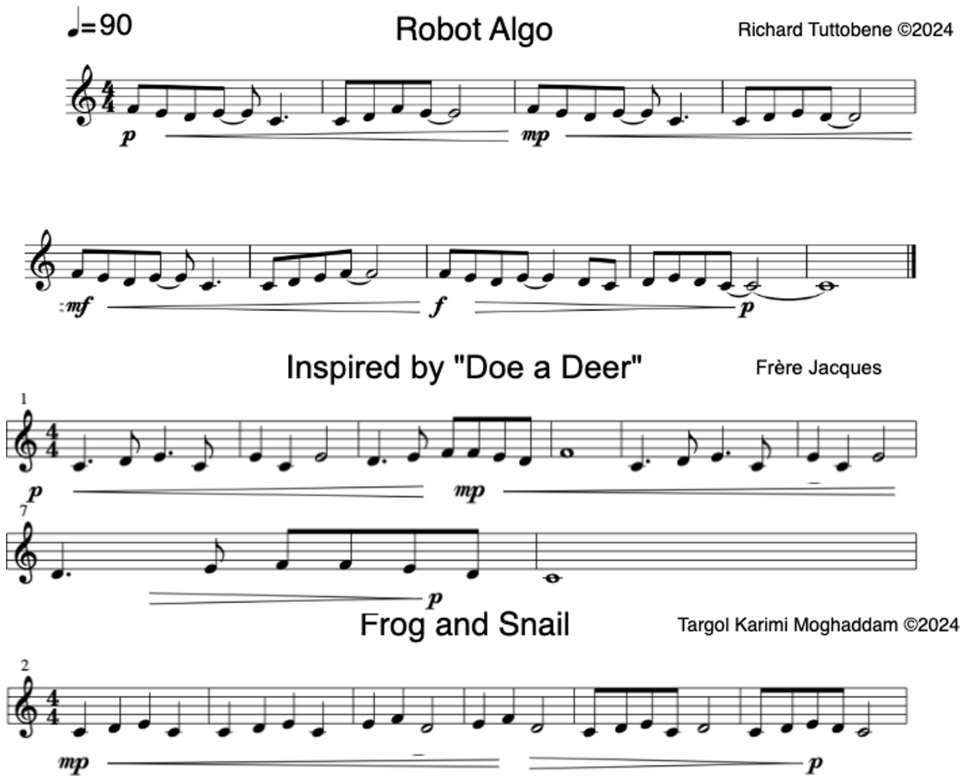

Figure 6. Scores for all three melodies: melody 1: Robot Algo, melody 2: Sleepless Nights and melody 3: Frog and Snail.

Table 1. Validation accuracy of networks on babbling data: we validated the accuracy of the ANNs for detecting correct notes and estimation of their intensities after the babbling phase.

network	cross-validated accuracy (%)
note detector network	98.4
intensity estimator network	86.2

contact dynamics). We also show how the Musician Hand outperforms novice participants as it replays a melody at a level comparable with that of trained human pianists. This system is inspired by a fundamental principle in sensorimotor neuroscience: the perception–action loop [15,20,32]. To implement the perception–action loop, the system receives continuous sensations (audio signal) and converts them to spectrograms (percepts) that are associated with percepts generated by its prior motor-actions. It then uses those percepts to create a set of motor actions that replicate the audio signal (musical sounds to a listener) of that particular melody.

3.1. Performance of perceptual networks

The validation results of the networks to detect notes and their intensity trained on babbling data (figure 1*b*) are listed in table 1. The detection of note onset was automated via identification of peaks in the short-time energy function of the spectrogram (similar to gap detection in mammals [54]).

3.2. Comparison of musical performance of melody 1 between the Musician Hand and human participants

We evaluated the performance of both the Musician Hand and human participants (novice and pianists) when replicating melody 1 (which has 37 notes). We did this in both quantitative and qualitative ways to establish a baseline performance for the Musician Hand. In the next section, we used this baseline performance metrics on melody 1 to test for generalization of our results to two additional melodies—melody 2 and 3, with 28 and 24 notes each, respectively, which were not played by humans.

3.2.1. Quantitative comparison

We compared the performance of both the Musician Hand and human participants in replicating melody 1 by the F1 score. Recall that the F1 score quantifies both *Precision* and *Recall* of the notes played, equation (2.12). For note detection, the five novice human participants could only repeat the first two or three notes correctly. That is, none succeeded in playing the melody beyond the third note in the 1 min provided. This provides an F1 score between 6% and 9%, figure 5. Because they did not complete the melody, the

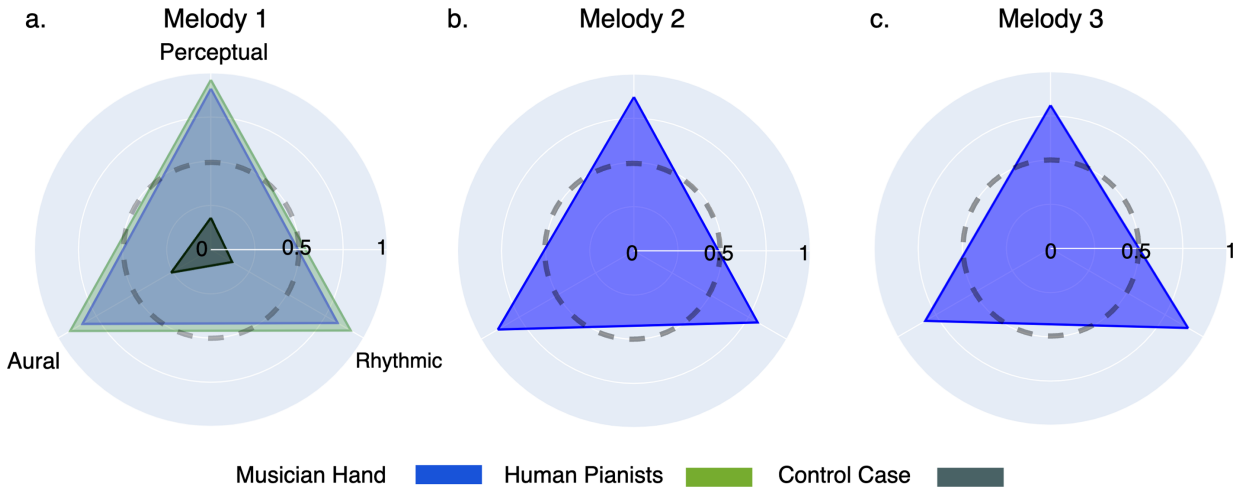

Figure 7. Qualitative results: we algorithmically assessed the similarity of the recorded melodies and the original melody along three qualitative dimensions: (a) perceptual, as per the SSIM between the spectrogram images [60]; (b) aural, as per similarity of MFCCs (approximating the human auditory system's response [63]); and (c) rhythmic, as per the alignment of rhythmic patterns between the original audio and the generated audio signals, focusing on the presence and timing of onsets [62]. For all three metrics, a value of 0.5 indicates a moderate similarity, implying that the melodies share some features but have significant differences in their spectrograms (perceptually on *y*-axis) and short-term power spectrum (aurally on *x*-axis). A value of 1 signifies complete similarity, meaning the melodies are identical in their measured characteristics. We found that the *Musician Hand* performed comparably with the four human pianists, all of whom scored greatly above the control case, which is a melody with random notes and the same length as M1.

novice humans had null results for intensity estimation and time difference (i.e. missing red bars for these two metrics in figure 5*b*,*c*).

By contrast, all human pianists successfully reproduced the entire melody (one played flawlessly, while the others made between one and four mistakes when performing the 37 notes). For note detection, this awarded them an F1 score between 92 and 100% figure 5*a*. As they were able to complete the melody 1, we were able to calculate their intensity estimation's F1 score, which fell in the range between 78% and 100%, figure 5*b*. Their average time difference was below 120 ms (i.e. the gap between notes was an average of 120 ms shorter or longer than the target melody 1, figure 5*c*).

The Musician Hand, also shown in figure 5*a*–*c*, was able to complete the playback of melody 1 (all 37 notes), which resulted in an F1 score of 100%. The intensity estimation F1 score was 68%; its time difference averaged 150 ms.

3.2.2. Qualitative comparison

We measured the similarity of the audio signals generated by the Musician Hand and the human participants to the target audio signals from the composer's (RT or TKM) own performance of each melody 2. This was done only for the human pianists (i.e. the novice humans did not complete melody 1) by using perceptual similarity (measured by the SSIM of spectrograms), aural similarity (cosine similarity of MFCCs) and rhythmic similarity (normalized cross-correlation of note onsets), as shown in figure 7. We consider these three numerical metrics as 'qualitative' because they quantify the overall resemblance between the entire target and performed audio signals. For all three metrics, a score of zero reflects no similarity (entirely different characteristics), 0.5 reflects moderate similarity (some shared features with major differences in spectrograms and short-term power spectra), and 1 reflects complete similarity (identical characteristics). To establish a baseline for melody 1, we included a control case in which the Musician Hand played 37 random notes.

For melody 1, the Musician Hand achieved similarity scores comparable with the four human pianists across all three metrics, with both consistently performing in the 0.9+ range, which was far above the control case which scored below 0.25 for all metrics figure 7*a*.

Table 2 lists the results for the blinded evaluation of the human pianists and the Musician Hand by two professional composers (RT and TKM). They independently ranked the melodic replication of the target melody 1 as per their professional experience, similar to how they rank competitors by hearing them play behind a curtain. Their rankings were in full agreement: the three professional pianists were ranked first and in the same order (B, C, D), followed by the Musician Hand ranked fourth and the least experienced pianist (A) ranked fifth. Recall that it was not possible to rank the novice pianists as they did not complete melody 1.

3.3. Generalization of performance to melodies 2 and 3

Having established the comparison of the human participants (trained and novice) versus the Musician Hand, we also evaluated the ability of the Musician Hand to replicate two additional melodies: melody 2 and melody 3 (figure 6), without providing any additional babbling or training beyond what was used for melody 1. Note that the performance of melody 1 did not provide any additional training. The Musician Hand successfully reproduced the full sequence of notes in melodies 2 and 3.

Table 2. The ranking by human expert critics: for each melody, two professional composers (i.e. expert critics) blindly evaluated how melodious the replication was by rank-ordering the recorded performances from best to worst. In particular, they evaluated the musical similarity to the target melody 1 (played by a professional pianist) as per the melodious and musical reproduction. Both expert critics unanimously agreed on the ranking order, ranking the Musician Hand fourth out of five and the only amateur fifth out of five.

ranking	rankee
1	human pianist B
2	human pianist C
3	human pianist D
4	Musician Hand
5	human pianist A

Quantitatively, the Musician Hand achieved an F1 score of 100% across both melodies (i.e. detected all notes as for melody 1). For intensity estimation, its F1 score was 84% on melody 2 and 70% on melody 3. Finally, its time difference averaged 135 ms for melody 2 and 150 ms for melody 3. These results are shown as the two rightmost blue bars in figure 5.

Qualitatively, the Musician Hand performed well, as for melody 1, as shown in figure 7. Both the spectrogram-based perceptual similarity and the MFCC-based aural similarity were in line with the values achieved for melody 1. The rhythmic similarity, based on note onset correlations, similarly demonstrated that the Musician Hand reproduced the temporal structure of the additional melodies.

Together, these quantitative and qualitative results highlight the Musician Hand's ability to replicate melodies comparably with human pianists, and its ability to generalize across different melodies.

4. Discussion

Our study demonstrates a novel end-to-end perceptual and experience-driven approach to melodious music reproduction by a robotic hand. The Musician Hand leverages perceptual learning to play-by-ear driven only by its own experience, with minimal pre-defined features of musical notes. This allows it, like a novice learning the piano by ear, to generalize to play any melody that uses the fingers and notes experienced in the past. This enabled our robotic hand to artistically perform a sensorimotor skill. As such, it serves as a proof-of-principle of an engineered implementation of an autonomous perception-driven system.

Perception fundamentally contextualizes the sensations created by one's own motor actions, and attributes meaning to them, driving subsequent relevant actions [64,65]. This process fundamentally builds upon experience and interactions with the physical world, creating a *Sensorimotor Gestalt* [37], an organized whole of compatible sensations and motor actions [15,17,32]. By leveraging the perception–action link built through this experience-driven process, the Musician Hand (a tendon-driven robotic hand that is activated by an exponential approximation to muscle activation–contraction dynamics) learns to translate audio signals of a melody directly into the motor outputs required for its melodious musical reproduction. This methodology, therefore, echoes foundational ideas of self-agency and self in a robotic system, where a self-taught perception–action system results from its own limited experience [25,32,37,66]. This proof-of-principle implementation unveils new possibilities for bridging the gap between human psychophysics and robotic autonomy for artistic and physical performance.

The Musician Hand can—after only a brief period of motor babbling—interpret the continuous sound signals of a melody into a sequence of musical percepts (i.e. sequences of notes of varying intensities) that it then executes to replicate the melody. Our algorithm indeed depends on the minimal pre-defined features of a melody being a sequence of musical notes. Thus our musically relevant algorithm is justifiably localized to this type of time-varying sequences of spectral peaks (otherwise known as notes strung into melodies in music). This enabled the Musician Hand to artistically reproduce arbitrary melodies of any duration that use the notes experienced during babbling—in a single attempt and after hearing them only once. Importantly, the algorithm's ability to 'play-by-ear' was comparable with trained human pianists and better than human novices as per quantitative and qualitative metrics (figures 5 and 7). The Musician Hand was outscored by three of the four trained pianists (the professional pianists), but outperformed an amateur pianist, table 2. We did not expect that two professional composers serving as critics would rank our first implementation of the Musician Hand higher than the performance of professional pianists with years of experience, table 2. But we were pleasantly surprised that the Musician Hand outranked an amateur pianist, and that our system's quantitative and qualitative metrics of performance (figures 5 and 7) were comparable with those of all human pianists—and better than the novices.

One difference between the Musician Hand and human participants is access to memory. The Musician Hand has access to an external representation of the melody (the spectrogram), whereas humans rely on short-term and working memory to reproduce the 29 s sequence. Musical training often enhances working memory and in this case a 29-note melody can be considered short. Nevertheless, this may have contributed to the better performance of trained pianists relative to novices—but is less of an issue when comparing the trained pianists to the Musician Hand. Our results should therefore emphasize comparison of perception-driven reproduction assuming memory resources available to trained pianists, rather than an equivalence of robotic and human musicianship.

We translate end-to-end autonomous learning architectures to the perceptual domain of experience-driven artistic performance. The value of sequences of physical actions lies in their melodic nuance. As a first example, we used piano playing as it involves

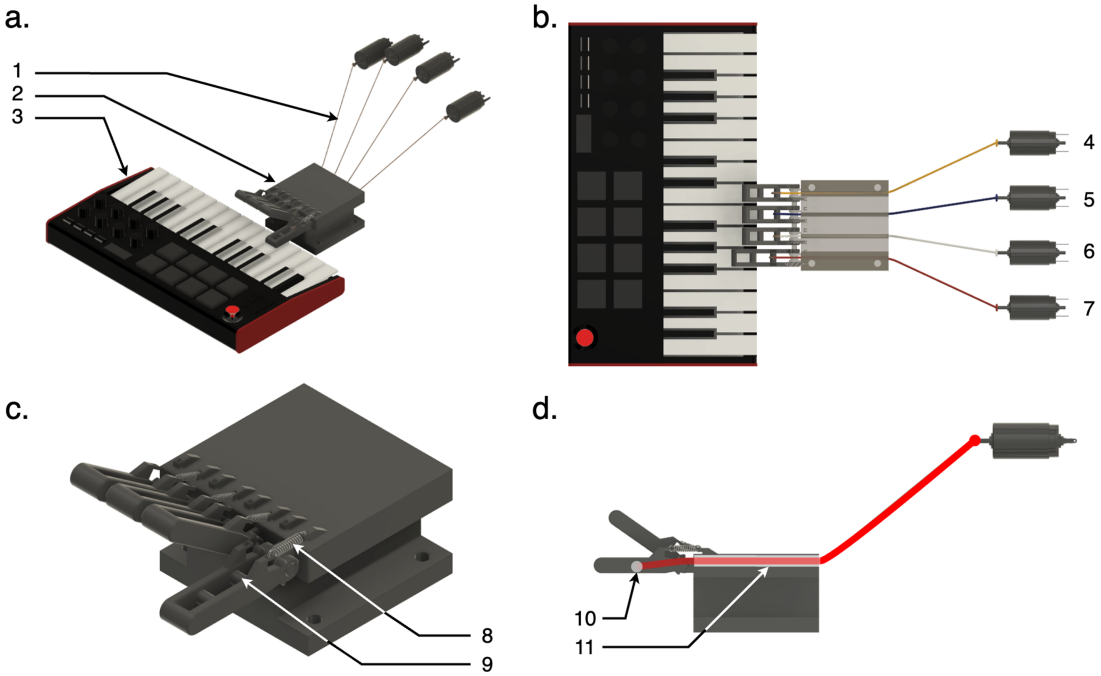

Figure 8. A tendon-driven robotic hand uses four fingers to perform motor babbling to then play melodies by ear. (a) Overview: 1. Flexor tendons pulled by DC motors. 2. Musician Hand structure. 3. Keyboard. (b) Top view of the system: 4–7. The four DC motors (Motor 0–3) used to pull the tendons. (c) Musician Hand: 8. Spring to passively extend finger. 9. Flexor tendon attachment. (d) Side view of the hand: 10. Flexor tendon attachment 11. Flexor tendon routing.

note sequences and intensities that are woven together via delicate control of the timing and dynamics of finger actions to elicit a melodious experience in the listener. Compared with the state-of-the-art in robotic piano playing, our approach is clearly different from ninteenth–twenty-first century pre-programmed pianolas that play melodies in their repertoire by activating keys according to perforated paper or metal rolls, or MIDI [67–71]. Instead, our approach is to derive melodious performance from the perception of the sound signals from unknown melodies themselves—as human pianists can. That is, as mentioned before, our approach (figures 1 and 2) is inspired by the foundations of the biological perception–action loop. Note that, as per the critical brain–body collaboration in organisms (between controller and plant in engineering terms) [19,72], we strove to make the physical properties of our Musician Hand begin to approximate the actuation and soft-tissue mechanics of the fingertips like those of the human pianists. We did this, respectively, by using a tendon-driven finger design actuated with DC motors emulating muscle contraction (figure 8) [51], and foam-covered fingertips to approximate the compression of the finger pads of human pianists [73] (figure 8). We believe that they probably contributed to the nuance of the music played by the Musician Hand. This is because they provided built-in bio-mimetic time constants of movement and impact mechanics on the velocity-sensitive keyboard we used. In addition, a foundation of perception–action is the need for a stable relationship between action and the resulting percept. Therefore, we implemented stability of input (finger action) and output (perception of music) by each finger corresponding precisely to a note. Adding a layer of learning to move the hand over the keyboard to achieve the appropriate location for the perception–action stability across octaves is not needed to provide proof-of-principle of perceptual learning. A trumpet, where each finger is obligatorily tied to a valve, is no less musical because of this mechanical limitation.

Note that the Musician Hand successfully replayed target melodies after limited prior experience (2 min of motor babbling) and after hearing them only once. This is an example of one-shot learning (by not requiring or allowing refinements as defined in the field [41]). By contrast, state-of-the-art robotic systems typically rely on extensive models, pre-programmed knowledge and feedback mechanisms during repeated attempts to refine a pre-defined known task [1,3–8,74]. Our approach extends our prior work on learning with limited experience [10] by demonstrating the utility in the perception–action loop to achieve one-shot nuanced musical performance using minimal prior information and training.

This first implementation of the Musician Hand naturally has limitations that, nevertheless, do not eclipse our proof-of-principle demonstration of perceptual learning in the physical world. First and foremost, the performance of the Musician Hand is—as in biological motor learning—a product of its own limited experience with a finite set of notes and intensities (figure 3). The experience of the Musician Hand is, by definition, a smaller set of musical perception and motor actions than the trained human pianists. This limited repertoire of experience, as in biology, naturally limits how the melodies were *perceived* and *interpreted* (i.e. heard and reproduced, respectively). Thus, the Musician Hand does not benefit from the sophisticated system humans have evolved to—during critical periods in typically developing children—create perceptual attractors to convert sound signals into phonemes or pitch [75,76]. Which, as an interesting aside, has the downside of some people having strong accents in a second language learnt later in life as they simply 'cannot hear' those new phonemes, or acquire perfect pitch.

More generally, perception–action is naturally limited by experience and the statistical priors it builds, as is emphasized in the Bayesian approach to sensorimotor function [77,78]. Thus, one source of inaccuracy in the Musician Hand could be the paucity of its musical experience during its limited motor babbling.

Regardless of these limitations, this demonstration of perceptual learning represents a departure from traditional robotics, which emphasize state-dependent control theory with specialized goals, pre-programmed behaviours or laborious constructions of internal models via extensive trial-and-error techniques. By going back to the perceptual roots of biological learning and behaviour, we demonstrate a novel end-to-end perceptual experience-driven approach for nuanced motor behaviour, with autonomous piano playing as a first demonstration.

This demonstration of the perceptual learning algorithm suggests future work can develop robotic systems that can replicate human artistic behaviour that inspires new forms of human–robot interaction and collaboration that approximate perception-driven (and at times emotionally driven [79]) strategies that humans depend on [15,20–30]. Integrating advancements in artificial intelligence and biological principles could facilitate even more seamless human–robot collaboration, driving innovation in various domains [18,19]. In addition, incorporating real-time error correction, feedback mechanisms and adaptive learning could further improve the precision and expressiveness of robotic performance. Extending perceptual learning to integrate tactile and proprioceptive inputs, for example, may enable precision tasks like surgery or craftsmanship, expanding the effect of cognitive robotic autonomy for physical function. Future work could also explore the adaptation of our perceptual learning algorithm to two movable hands that can roam across the keyboard, or its application to different types of musical instruments and performance styles to demonstrate its generalizability.

Regarding generalizability, a key element of musical performance in humans is our ability to listen to a piece of music played on one instrument and replicate it on a different one. This is a form of biological transfer learning (based on compositionality) that ANNs currently lack [18]; This is, in fact, a use of compositionality—the ability to decompose complex tasks (e.g. music from one instrument) into more elementary components that can be reused for related tasks (e.g. from another instrument) [18]—that would be particularly useful as an extension of our perceptual learning algorithm for musical performance. This fundamentally limited generalizability of ANNs comes from the fact that the training set has an inordinate influence on the setting of weights that somehow reduces the neural network's capacity to generalize across varying instrumental styles [80]. Future work could improve the generalizability of transfer learning techniques, potentially by using perception to help bridge the gap between contexts.

Finally, the perceptual learning framework introduced here, by insisting on interactions with the physical world, opens the door to other applications. For example, help an agent physically create time-varying signals that closely reproduce those of a source by maximizing perceptual similarity—such as using a generator other than a hand/keyboard (e.g. anthropomorphic robot) to match a visual signature (e.g. video).

Ethics. All procedures were approved by the University of Southern California internal review board (USC IRB: HS-17-00304), and written consent was obtained from all participants prior to participation. The study conformed to all standards the Declaration of Helsinki set, except for registration in a database.

Data accessibility. The data analysed in this study can be obtained from the corresponding author upon reasonable request. The videos and codes are included in the supplementary information.

Supplementary material is available online [81].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. H.A.: conceptualization, data curation, formal analysis, methodology, software, validation, visualization, writing—original draft, writing—review and editing; A.M.: conceptualization, methodology, validation, writing—original draft, writing—review and editing; F.V.-C.: conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work is supported in part by the NIH (Grant R21 NS113613), NSF (Grant 2113096 CRCNS US—Japan with K. Seki), DoD (Grant CDMRP MR150091 with R. Balasubramanian), DARPA (Grant L2M W911NF1820264 with A. Parker), and The Wu Tsai Human Performance Alliance at Stanford University project awarded to F.J.V.-C. And by the Provost and Research Enhancement Fellowships from the Graduate School of the University of Southern California to H.A. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, DoD, DARPA or The Wu Tsai Human Performance Alliance.

Acknowledgements. We had the privilege of collaborating with two composers, Richard Tuttobene and Targol Karimi Moghaddam, who graciously crafted three new melodies. These pieces capture melodious musical dynamics using only four adjacent notes. We also thank Jan Lao and Zeyu (Claude) Yao, who helped us build the hardware for the Musician Hand. Special thanks to Angelo Bartch-Jimenez and Majid Abbasi-Sisara who contributed to this work with their comments and insights.

Dedication. To my father, Ramezan, who first taught me to ask 'why'.

References

1. James S, Wohlhart P, Kalakrishnan M, Kalashnikov D, Irpan A, Ibarz J, Levine S, Hadsell R, Bousmalis K. 2019 Sim-to-real via sim-to-sim: data-efficient robotic grasping via randomized-to-canonical adaptation networks. *ArXiv:1812.07252 [cs]*. (doi:10.48550/arXiv.1812.07252)
2. Akkaya I *et al*. 2019 Solving rubik's cube with a robot hand. *arXiv Preprint arXiv:1910.07113*.
3. Krishnan S, Garg A, Liaw R, Thananjeyan B, Miller L, Pokorny FT, Goldberg K. 2019 SWIRL: a sequential windowed inverse reinforcement learning algorithm for robot tasks with delayed rewards. *Int. J. Rob. Res.* **38**, 126–145. (doi:10.1177/0278364918784350)
4. Bongard J, Zykov V, Lipson H. 2006 Resilient machines through continuous self-modeling. *Science* **314**, 1118–1121. (doi:10.1126/science.1133687)
5. Marques HG, Bharadwaj A, Iida F. 2014 From spontaneous motor activity to coordinated behaviour: a developmental model. *PLoS Comput. Biol* **10**, e1003653.
6. Manoonpong P, Geng T, Kulvicius T, Porr B, Wörgötter F. 2007 Adaptive, fast walking in a biped robot under neuronal control and learning. *PLoS Comput. Biol.* **3**, e134. (doi:10.1371/journal.pcbi.0030134)
7. Osa T, Peters J, Neumann G. 2018 Hierarchical reinforcement learning of multiple grasping strategies with human instructions. *Adv. Robot.* **32**, 955–968. (doi:10.1080/01691864.2018.1509018)

8. Lowrey K, Kolev S, Dao J, Rajeswaran A, Todorov E. 2018 Reinforcement learning for non-prehensile manipulation: transfer from simulation to physical system. In *2018 IEEE Int. Conf. on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, pp. 35–42. Brisbane, Australia: IEEE. (doi:10.1109/SIMPAR.2018.8376268)
9. Urbina-Meléndez D, Azadjou H, Valero-Cuevas FJ. 2024 Brain-body-task co-adaptation can improve autonomous learning and speed of bipedal walking. *arXiv preprint arXiv:2402.02387*.
10. Marjaninejad A, Urbina-Meléndez D, Cohn BA, Valero-Cuevas FJ. 2019 Autonomous functional movements in a tendon-driven limb via limited experience. *Nat. Mach. Intell.* **1**, 144–154. (doi:10.1038/s42256-019-0029-0)
11. Flash T. 2021 Brain representations of motion generation and perception: space-time geometries and the arts. In *Space-time geometries for motion and perception in the brain and the arts* (eds T Flash, A Berthoz), pp. 3–34. Cham, Switzerland: Springer International Publishing. (doi:10.1007/978-3-030-57227-3_1)
12. Adolph KE, Franchak JM. 2017 The development of motor behavior. *Wiley Interdiscip. Rev. Cogn. Sci.* **8**, e1430. (doi:10.1002/wcs.1430)
13. Whitall J, Clark JE. 2018 Chapter eight - a perception–action approach to understanding typical and atypical motor development. In *Advances in child development and behavior, vol. 55 of studying the perception-action system as a model system for understanding development* (ed. JM Plumert), pp. 245–272. San Diego, CA: Academic Press. (doi:10.1016/bs.acdb.2018.04.004)
14. Meer AVD, Weel FRR. 2022 Motor development: biological aspects of brain and behavior. In *Oxford research encyclopedia of psychology* (ed. O Braddick). Oxford, UK: Oxford University Press. (doi:10.1093/acrefore/9780190236557.013.903)
15. Mechsner F, Kerzel D, Knoblich G, Prinz W. 2001 Perceptual basis of bimanual coordination. *Nature* **414**, 69–73. (doi:10.1038/35102060)
16. Tomassini A, Vercillo T, Torricelli F, Morrone MC. 2018 Rhythmic motor behaviour influences perception of visual time. *Proc. R. Soc. B* **285**, 20181597. (doi:10.1098/rspb.2018.1597)
17. Fiser J. 2009 Perceptual learning and representational learning in humans and animals. *Learn. Behav.* **37**, 141–153. (doi:10.3758/LB.37.2.141)
18. Kudithipudi D *et al*. 2022 Biological underpinnings for lifelong learning machines. *Nat. Mach. Intell.* **4**, 196–210. (doi:10.1038/s42256-022-00452-0)
19. Valero-Cuevas FJ, Erwin A. 2022 Bio-robots step towards brain–body co-adaptation. *Nat. Mach. Intell.* **4**, 737–738. (doi:10.1038/s42256-022-00528-x)
20. Mechsner F. 2004 A perceptual-cognitive approach to bimanual coordination. In *Coordination dynamics: issues and trends* (eds VK Jirsa, JAS Kelso), pp. 177–195. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-540-39676-5_10)
21. Mechsner F. 2004 A psychological approach to human voluntary movement. *J. Mot. Behav.* **36**, 355–370. (doi:10.1080/00222895.2004.11007993)
22. Zimmer AC. 1990 Autonomous organization in perception and motor control (eds H Haken, M Stadler). In *Synergetics of Cognition*, pp. 332–351. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-642-48779-8_18)
23. Charles L, Chardin C, Haggard P. 2020 Evidence for metacognitive bias in perception of voluntary action. *Cognition* **194**, 104041. (doi:10.1101/423244)
24. Hamill J, Lim J, van Emmerik R. 2020 Locomotor coordination, visual perception and head stability during running. *Brain Sci.* **10**, 174. (doi:10.3390/brainsci10030174)
25. Creem-Regehr SH, Kunz BR. 2010 Perception and action. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 800–810. (doi:10.1002/wcs.82)
26. Ohson SS. 2012 The Effect of Concurrent Motor Activity on the Perception of Biological Motion. PhD diss., McMaster University.
27. Mitsumatsu H. 2009 Voluntary action affects perception of bistable motion display. *Perception* **38**, 1522–1535. (doi:10.1068/p6298)
28. Buaron B, Reznik D, Gilron R, Mukamel R. 2020 Voluntary actions modulate perception and neural representation of action-consequences in a hand-dependent manner. *Cereb. Cortex* **30**, 6097–6107. (doi:10.1093/cercor/bhaa156)
29. Gong Y, Wang Y, Wang Y. 2022 Voluntary and involuntary dynamics of perception-action processing. *Curr. Psychol.* **41**, 5343–5349. (doi:10.1007/s12144-020-01028-0)
30. Bartsch-Jimenez A, Błażkiewicz M, Azadjou H, Novotny R, Valero-Cuevas FJ. 2023 '*Fine synergies*' describe motor adaptation in people with drop foot in a way that supplements traditional '*Coarse synergies*'. *Front. Sports Act. Living* **5**, 1080170. (doi:10.3389/fspor.2023.1080170)
31. Soroushmojdehi R, Javadzadeh S, Asadi M, Sanger TD. 2025 Disentangling shared and private neural dynamics with SPIRE: a latent modeling framework for deep brain stimulation. *arXiv Preprint arXiv:2510.25023*.
32. Arbib MA. 2011 Perceptual structures and distributed motor control. In *Comprehensive physiology* (eds RJ Terjung, VB Brooks), pp. 1449–1480. Hoboken, NJ: John Wiley & Sons, Ltd. (doi:10.1002/cphy.cp010233)
33. Rizzolatti G, Arbib MA. 1998 Language within our grasp. *Trends Neurosci.* **21**, 188–194. (doi:10.1016/s0166-2236(98)01260-0)
34. MacKenzie CL, Iberall T. 1994 *The grasping hand*. Amsterdam, The Netherlands: Elsevier.
35. Niyo G, Almofeez LI, Erwin A, Valero-Cuevas FJ. 2023 An α-mn collateral to γ-mns can mitigate velocity-dependent stretch reflexes during voluntary movement: a computational study. *bioRxiv 2023–12*.
36. Asadi M, Javadzadeh S, Soroushmojdehi R, Mousavi S, Sanger TD. 2025 Bace: behavior-adaptive connectivity estimation for interpretable graphs of neural dynamics. *arXiv preprint*.
37. Berry JA, Valero-Cuevas FJ. 2020 Sensory-motor gestalt: sensation and action as the foundations of identity, agency, and self. In *The 2020 Conference on Artificial Life. Online*, pp. 130–138. Cambridge, MA, USA: MIT Press. (doi:10.1162/isal_a_00340)
38. Zakka K *et al*. 2023 Robopianist: dexterous piano playing with deep reinforcement learning. *arXiv preprint arXiv:2304.04150*.
39. Zhao Y, Chen L, Schneider J, Gao Q, Kannala J, Schölkopf B, Pajarinen J, Büchler D. 2024 RP1M: a large-scale motion dataset for piano playing with bi-manual dexterous robot hands. *arXiv Preprint arXiv:2408.11048*.
40. Qian C, Urain J, Zakka K, Peters J. 2024 Pianomime: learning a generalist, dexterous piano player from internet demonstrations. *arXiv Preprint arXiv:2407.18178*.
41. Fergus R, Li FF, Perona P, Zisserman A. 2010 Learning object categories from internet image searches. *Proc. IEEE* **98**, 1453–1466. (doi:10.1109/JPROC.2010.2048990)
42. Kobayashi H, Ozawa R. 2003 Adaptive neural network control of tendon-driven mechanisms with elastic tendons. *Automatica* **39**, 1509–1519. (doi:10.1016/S0005-1098(03)00142-0)
43. Marco A, Hennig P, Bohg J, Schaal S, Trimpe S. 2016 Automatic LQR tuning based on Gaussian process global optimization. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 270–277. Piscataway, NJ: IEEE. (doi:10.1109/ICRA.2016.7487144)
44. Takahashi K, Ogata T, Nakanishi J, Cheng G, Sugano S. 2017 Dynamic motion learning for multi-DOF flexible-joint robots using active–passive motor babbling through deep learning. *Adv. Robot.* **31**, 1002–1015. (doi:10.1080/01691864.2017.1383939)
45. Nguyen-Tuong D, Peters J, Seeger M, Scholkopf B. 2008 Learning inverse dynamics: a comparison. *Artif. Neural Netw* 13–18.
46. Gijsberts A, Metta G. 2013 Real-time model learning using incremental sparse spectrum Gaussian process regression. *Neural Netw.* **41**, 59–69. (doi:10.1016/j.neunet.2012.08.011)
47. Della Santina C, Lakatos D, Bicchi A, Albu-Schaeffer A. 2020 Using nonlinear normal modes for execution of efficient cyclic motions in articulated soft robots. In *International Symposium on Experimental Robotics*, pp. 566–575. Cham, Switzerland: Springer International Publishing.
48. Valero-Cuevas FJ. 2016 *Fundamentals of neuromechanics*. vol. 8. London, UK: Springer.
49. Jalaleddini K, Minos Niu C, Chakravarthi Raja S, Joon Sohn W, Loeb GE, Sanger TD, Valero-Cuevas FJ. 2017 Neuromorphic meets neuromechanics, part II: the role of fusimotor drive. *J. Neural Eng.* **14**, 025002. (doi:10.1088/1741-2552/aa59bd)
50. McMahon TA. 1984 *Muscles, reflexes, and locomotion*. Princeton, NJ: Princeton University Press. (doi:10.1515/9780691221540)

51. Nagamori A, Laine CM, Loeb GE, Valero-Cuevas FJ. 2021 Force variability is mostly not motor noise: theoretical implications for motor control. *PLoS Comput. Biol.* **17**, e1008707. (doi:10.1371/journal.pcbi.1008707)
52. Mayfield DL, Cronin NJ, Lichtwark GA. 2023 Understanding altered contractile properties in advanced age: insights from a systematic muscle modelling approach. *Biomech. Model. Mechanobiol.* **22**, 309–337. (doi:10.1007/s10237-022-01651-9)
53. Mounir M, Karsmakers P, van Waterschoot T. 2021 Musical note onset detection based on a spectral sparsity measure. *EURASIP J. Audio. Speech. Music Process.* **2021**, 30. (doi:10.1186/s13636-021-00214-7)
54. Anderson LA, Linden JF. 2016 Mind the gap: two dissociable mechanisms of temporal processing in the auditory system. *J. Neurosci.* **36**, 1977–1995. (doi:10.1523/JNEUROSCI.1652-15.2016)
55. Schluter J, Bock S. 2014 Improved musical onset detection with convolutional neural networks. In *2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979–6983. Piscataway, NJ: IEEE. (doi:10.1109/ICASSP.2014.6854953)
56. de Jesus Guerrero-Turrubiates J, Gonzalez-Reyna SE, Ledesma-Orozco SE, Avina-Cervantes JG. 2014 Pitchestimation for musical note recognition using artificial neural networks. In *2024 Int. Conf. on electronics, communications and computers (CONIELECOMP)*, pp. 53–58. Piscataway, NJ: IEEE.
57. Lohani B, Gautam CK, Kushwaha PK, Gupta A. 2024 Deep learning approaches for enhanced audio quality through noise reduction. In *2024 Int. Conf. on Communication, Computer Sciences and Engineering (IC3SE)*, pp. 447–453. Piscataway, NJ: IEEE. (doi:10.1109/IC3SE62002.2024.10593073)
58. Gubin MV. 2018 Using convolutional neural networks to classify audio signal in noisy sound scenes. In *2018 Global Smart Industry Conference (GloSIC)*, pp. 1–6. Piscataway, NJ: IEEE. (doi:10.1109/GloSIC.2018.8570117)
59. Hernandez-Olivan C, Zay Pinilla I, Hernandez-Lopez C, Beltran JR. 2021 A comparison of deep learning methods for timbre analysis in polyphonic automatic music transcription. *Electronics* **10**, 810. (doi:10.3390/electronics10070810)
60. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. 2004 Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612. (doi:10.1109/tip.2003.819861)
61. Hasan MR, Hasan MM, Hossain MZ. 2021 How many mel-frequency cepstral coefficients to be utilized in speech recognition? A study with the Bengali language. *J. Eng.* **2021**, 817–827. (doi:10.1049/tje2.12082)
62. Cocharro D, Bernardes G, Bernardo G, Lemos C. 2021 A review of musical rhythm representation and (dis) similarity insymbolic and audio domains. In *Perspectives on music, sound and musicology: research, education and practice*, pp. 189–208. Cham, Switzerland: Springer. (doi:10.1007/978-3-030-78451-5_10)
63. Ruan P, Zheng X, Qiu Y, Hao Z. 2022 A binaural MFCC-CNN sound quality model of high-speed train. *Appl. Sci.* **12**, 12151. (doi:10.3390/app122312151)
64. Alomari M, Li F, Hogg DC, Cohn AG. 2022 Online perceptual learning and natural language acquisition for autonomous robots. *Artif. Intell.* **303**, 103637. (doi:10.1016/j.artint.2021.103637)
65. Becerra JA, Duro RJ, Monroy J. 2018 A redescriptive approach to autonomous perceptual classification in robotic cognitive architectures. In *Int. Joint Conf. on Neural Networks (IJCNN)*, pp. 1–8. Piscataway, NJ: IEEE. (doi:10.1109/IJCNN.2018.8489171)
66. Saegusa R, Metta G, Sandini G, Natale L. 2013 Developmental perception of the self and action. *IEEE Trans. Neural Netw. Learn. Syst* **25**, 183–202. (doi:10.1109/TNNLS.2013.2271793)
67. Aukkhosuwan W, Suntiamorntut W. 2022 AI system design for robotic hand to play the piano. *ASEAN Sci. Tech. Rept.* **25**, 59–68. (doi:10.55164/ajstr.v25i3.246950)
68. Hughes JAE, Maiolino P, Iida F. 2018 An anthropomorphic soft skeleton hand exploiting conditional models for piano playing. *Sci. Robot.* **3**, eaau3098. (doi:10.1126/scirobotics.aau3098)
69. Gao G, Zhong L, Zhu S, Wan M, Zhang P, Liang D, Gu J. 2022 Hierarchical optimal motion planning for piano playing robots with dexterous fingers. In *2022 IEEE Int. Conf. on Robotics and Biomimetics (ROBIO)*, pp. 357–362. Piscataway, NJ: IEEE. (doi:10.1109/ROBIO55434.2022.10011862)
70. Gao G, Zhong L, Zhang P, Huang Z, Du R, Li Y, Gu J. 2022 An intelligent piano playing algorithm applied to the humanoid robot. In *2022 12th Int. Conf. on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 55–60. Piscataway, NJ: IEEE. (doi:10.1109/CYBER55403.2022.9907577)
71. Topper A, Maloney T, Barton S, Kong X. 2019 Piano-playing robotic arm. *Worcester MA* 01609–02280.
72. Azadjou H, Błażkiewicz M, Erwin A, Valero-Cuevas FJ. 2023 Dynamical analyses show that professional archers exhibit tighter, finer and more fluid dynamical control than neophytes. *Entropy* **25**, 1414. (doi:10.3390/e25101414)
73. Müller B (ed). 2016 *Handbook of human motion*. Cham, Switzerland: Springer International Publishing.
74. OpenAI *et al*. 2019 Learning dexterous in-hand manipulation. *ArXiv:1808.00177*. (doi:10.48550/arXiv.1808.00177)
75. Dang T, Sethu V, Ambikairajah E, Epps J, Li H. 2021 Joint spatio-temporal discretisation of nonlinear active cochlear models. *ArXiv:2108.05993 [eess]*.
76. Melland P, Curtu R. 2023 Attractor-like dynamics extracted from human electrocorticographic recordings underlie computational principles of auditory bistable perception. *J. Neurosci.* **43**, 3294–3311. (doi:10.1523/JNEUROSCI.1531-22.2023)
77. Körding KP, Wolpert DM. 2004 Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247. (doi:10.1038/nature02169)
78. Lin CH, Faisal AA. 2017 The role of sensorimotor variability and computation in elderly's falls. *bioRxiv.* 196584. (doi:10.1101/196584)
79. Weidemann A, Rußwinkel N. 2021 The role of frustration in human-robot interaction - what is needed for a successful collaboration? *Front. Psychol.* **12**, 640186. (doi:10.3389/fpsyg.2021.640186)
80. Grebovic M, Filipovic L, Katnic I, Vukotic M, Popovic T. 2022 Overcoming limitations of statistical methods with artificial neural networks. In *2022 Int. Arab Conf. on Information Technology (ACIT)*, pp. 1–6. Abu Dhabi, United Arab Emirates: IEEE. (doi:10.1109/ACIT57182.2022.9994218)
81. Azadjou H, Marjaninejad A, Valero-Cuevas F. 2026 Supplementary material from: Perception in action: A robotic system that can teach itself to melodiously play music by ear. Figshare. (doi:10.6084/m9.figshare.c.8469590)