INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Lee ED, Esposito E, Cohen I. 2019 Audio cues enhance mirroring of arm motion when visual cues are scarce. *J. R. Soc. Interface* **16**: 20180903.
http://dx.doi.org/10.1098/rsif.2018.0903

Received: 1 December 2018
Accepted: 16 April 2019

Subject Category:
Life Sciences–Physics interface

Subject Areas:
computational biology, biomechanics

Keywords:
virtual reality, coordination, motion capture, mirroring, statistical learning

Author for correspondence:
Edward D. Lee
e-mail: edl56@cornell.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4486775.

THE ROYAL SOCIETY PUBLISHING

Audio cues enhance mirroring of arm motion when visual cues are scarce

Edward D. Lee, Edward Esposito and Itai Cohen

Department of Physics, Cornell University, 142 Sciences Drive, Ithaca, NY 14853, USA

EDL, 0000-0003-2075-6342

Swing in a crew boat, a good jazz riff, a fluid conversation: these tasks require extracting sensory information about how others flow in order to mimic and respond. To determine what factors influence coordination, we build an environment to manipulate incoming sensory information by combining virtual reality and motion capture. We study how people mirror the motion of a human avatar's arm as we occlude the avatar. We efficiently map the transition from successful mirroring to failure using Gaussian process regression. Then, we determine the change in behaviour when we introduce audio cues with a frequency proportional to the speed of the avatar's hand or train individuals with a practice session. Remarkably, audio cues extend the range of successful mirroring to regimes where visual information is sparse. Such cues could facilitate joint coordination when navigating visually occluded environments, improve reaction speed in human–computer interfaces or measure altered physiological states and disease.

1. Introduction

Successful coordination of human motion in a group is crucial for many tasks including dance, team sports or music ensembles [1–4]. In all these cases, it is essential that the individual extract information from the local environment [5,6] to maintain coordination with others. When the input to a sensory channel is disrupted systematically, however, how do individuals compensate for such disruption? In the context of sensorimotor integration for reaching tasks, this question has been well studied [7–9]. Here, we study the transition from coordinated to uncoordinated behaviour using an experimental apparatus that manipulates the visual and auditory fields, measures the dynamic motions of individuals, and quickly maps out performance across large regions of parameter space. To determine the relationship between available visual information and a subject's ability to mirror accurately, we asked 35 subjects to mirror the hand motions of a pre-recorded avatar while we changed the rate with and duration during which the avatar was visible. Next, we measured changes in performance when the subjects were given audio cues that mapped velocity of motion to frequency, practice training rounds or both training and audio cues. Using these data, we find that audio enhances performance at fast time scales while the combination of both audio and training affects the dynamics of coordination performance in a characteristic way that may be detectable in other experiments.

Pitch-based auditory cues provide an informative, intuitive and commonly used approach for representing kinematic and kinetic measurements in human motion [10–12]. When used as feedback, audio cues can enhance performance at motor tasks across a variety of contexts including learning a cyclic motion [10,13,14] and interpersonal coordination [2,15,16]—more generally perceptual coupling including other sensory modalities like vision and touch have been shown to enhance interpersonal coordination (see [17] for a review). In the case where subjects are trying to learn a new motion, evidence suggests that feedback based on the target motion is more effective than that based on the subject's own motion [13,15]. Along these lines, we represent the target hand motion of the avatar that the subject is mirroring using a simple proportional

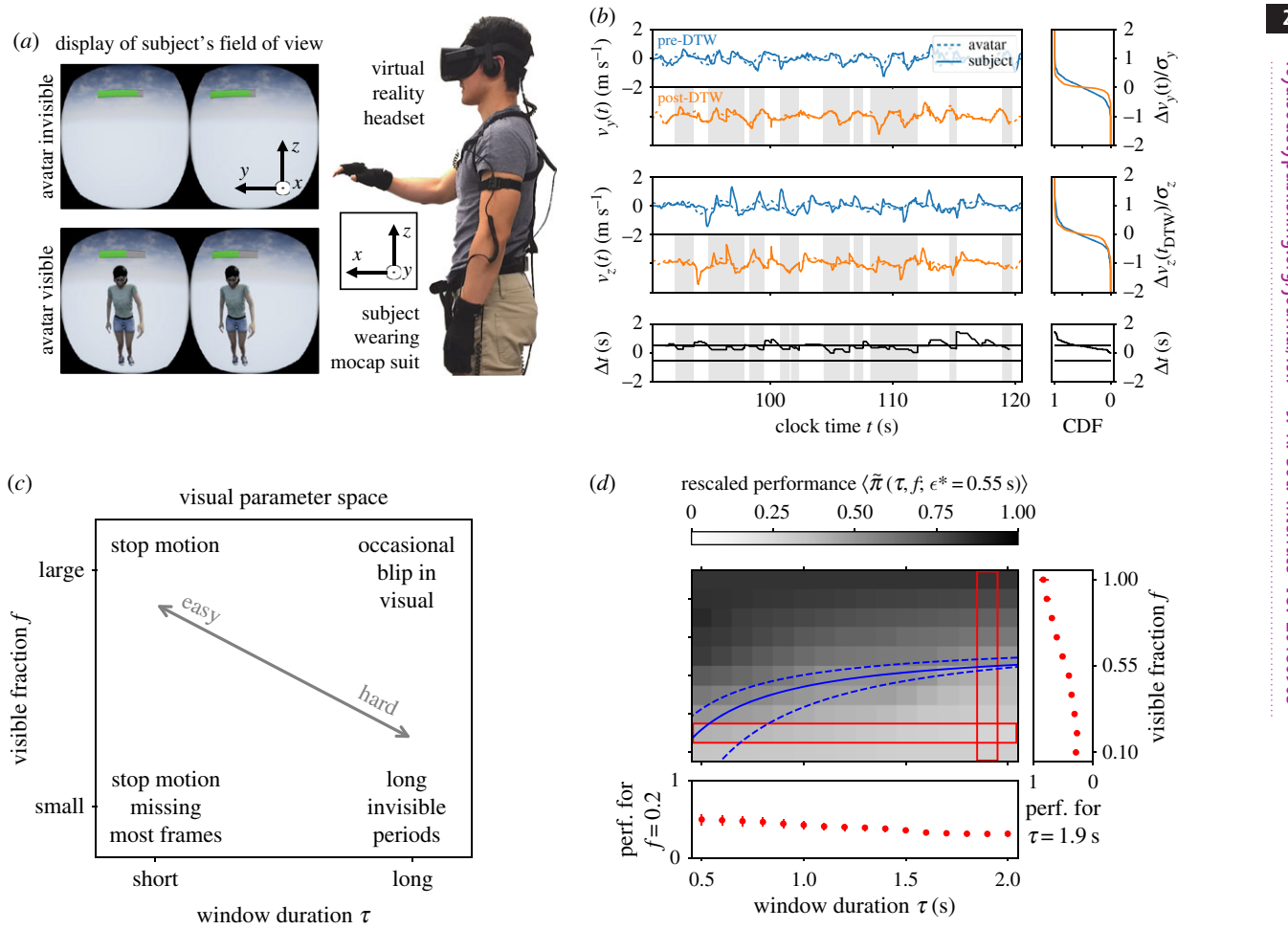

Figure 1. (*a*) Experimental set-up showing a subject wearing virtual reality goggles and the motion capture suit along with the subject's field of view when the avatar is invisible and visible. The green bar indicates current performance. (*b*) Dynamic time warping (DTW) aligns the measured velocities (blue) along the *y*- and *z*-axes. After DTW (orange), we identify runs of successful tracking (grey) and the fraction of the trial that these regions span is the estimated performance $\hat{\pi}$. DTW, as expected, reduces velocity error (normalized by standard deviation) as shown on right and returns time delays with distribution on the bottom right. (*c*) Parameter space diagram. The four corners represent the extremes of the parameter space. The visible fraction f determines the fraction of time, akin to the duty cycle, during which the avatar is visible. Where the visual representation of the avatar blinks on and off quickly, we call stop motion animation. (*d*) Rescaled performance landscape $\langle \tilde{\pi}(\tau, f) \rangle_{\epsilon^*=0.55\,\mathrm{s}}$ aggregated across all subjects in the Train + Audio condition ($M = 15$). One-dimensional cuts, outlined by red rectangles, are shown on the sides with predicted uncertainties. Error bars are one standard standard over the rescaled landscapes. Solid blue line traces the relation in equation (2.1) fit the level curve in performance given by $\langle \tilde{\pi} \rangle = 1/2$. Dashed blue lines indicate fits to rescaled landscapes one standard deviation above and below the shown mean landscape. (Online version in colour.)

pitch mapping based on speed. Given that these auditory cues provide complementary auditory information along with a visual of the avatar, we expect that performance at mirroring the avatar will be enhanced.

Beyond perceptual coupling, higher-level planning processes may play a role in learning how to mimic others' motions [18], an aspect that we study by training some of the participants in practice trials. In various studies of interpersonal coordination, there is evidence that anticipatory motor activation might help individuals respond to the motion of others [19,20]. One experiment showed that even imagining the motion of another subject prior to motion helped to synchronize behaviour [4]. Here, we explore how the provision of auditory cues might compare to the benefits of a training round where individuals have a chance to practice mirroring an avatar. We run two variations of the conditions where in one, participants do not have the opportunity to practice the task and in the other they do. This variation allows us to measure interplay between audio cues and practice with the task. In some experiments exploring the effect of audio cues, similar kinds of practice rounds precede measurement [13], whereas in others subjects only witness the task before immediately proceeding to performance [14]. We would expect that performance at the mirroring task would improve when subjects receive either training or audio cues in the absence of reliable visual cues, a prediction that would be consistent with other results in the literature [13,14,21].

In our experiments, subjects wore virtual reality goggles and a motion capture suit, stood face-to-face with an avatar that played a pre-recorded sequence of aperiodic motions generated by an experimenter and were instructed to mirror the motion of the avatar's hand as shown in figure 1*a*. Each experimental sequence consisted of 16 sequential 30 s trials with varying difficulty. To control the difficulty, we took windows of duration $1/2\,\mathrm{s} \leq \tau \leq 2\,\mathrm{s}$ and only showed the avatar for a contiguous visible fraction $0.1 \leq f \leq 1$ of the window (f is analogous to the duty cycle). In the first and last trials, the avatar was visible at all times, so $f = 1$. After a 16 trial sequence with a randomly chosen hand, the subjects repeated another sequence with the other hand for a different set of motions.

We assess how well subjects mirrored the motion of the avatar by comparing the two-dimensional velocity trajectory

of the subject $v_s(t)$ with the avatar's $v_a(t)$. We show an example for a single 30 s trial in figure 1*b*, where we measure the velocity of the subject's hand (solid blue line) in the mirror plane separating the subject from the avatar. The plane corresponds to the y and z axes as defined in figure 1*a*.[1] By inspecting the coarse features of the trajectories, we observe that the subject captures much of the lower frequency motions of the avatar but only after a varying temporal delay. To account for these delays, we use a standard algorithm for aligning two trajectories with local temporal modulation called dynamic time warping (DTW) [22–24]. We regularize the alignment problem so that solutions where the subject is more than 1/2 s ahead or 3/2 s behind are penalized to avoid pathologies that can arise from periodic motion (electronic supplementary material, section S3). We show an example of the time-warped velocity trajectories (orange lines) in figure 1*b*. After DTW, the velocity difference normalized by the typical size of the velocity fluctuations of the avatar, σ_a, is substantially narrower than the unaligned distribution. This narrowing indicates that accounting for temporal delays ϵ (black line in figure 1*b*) substantially improves feature matching between the curves. Since close mirroring corresponds to minimal delay, we use the distribution of delays found from aligning the curves as a measure of how well subjects mirrored the avatar; results are similar if we also consider the direction of the velocity vector (electronic supplementary material, section S4).

After alignment with DTW, we summarize mirroring performance with the estimated fraction of time that a subject is able to stay within a time threshold ϵ^* given the window duration τ and visible fraction f, $\hat{\pi}(\tau, f; \epsilon^*)$, which can only vary from 0 to 1. When the subject is consistently within a time delay of ϵ^* (as indicated by the shaded regions in figure 1*b*), the estimated performance measure $\hat{\pi} \approx 1$. With a short threshold ϵ^*, high performing subjects must mirror the avatar very closely with few deviations in both timing and velocity—we find that dissimilar trajectories lead to strong temporal variability with DTW. On the other hand with large ϵ^*, slower reaction times and bigger corrections will not affect the value of the performance. Thus, we vary ϵ^* to probe variation in how closely subjects mirror the avatar.

Given a particular value of the time threshold ϵ^*, we use Gaussian process regression to model a single subject's performance landscape using the 16 trials as training data to interpolate the unmeasured points [25,26]. These 16 data points represent a sparse sample of 160 discretized grid points. During an experiment, we chose these points by updating a Gaussian process model on previous trials and selecting points of maximum predicted uncertainty to explore quickly the performance landscape. After the experiments, we combined all subjects into another multi-subject Gaussian process that captures subject-specific variation and shared structure; this model agrees closely with the data. Checking with a leave-one-out cross validation procedure, we find that the multi-subject model works well as measured by the strong correlation of the prediction with the test point ($\rho = 0.95$) across all experiments (electronic supplementary material, sections S5 and S6).

2. Results

Looking across subjects, we find that performance varies with both visibility parameters. To show this trend, we combine performance landscapes across subjects after normalizing them to be centred about the same midpoint of performance (electronic supplementary material, section S6). We show an example for $\epsilon^* = 0.55$ s—a few times the fastest motor response time for humans [27]—in figure 1*c*. At $f = 1$, the avatar is always visible and subject performance is the highest. As avatar visibility is reduced by decreasing the fraction visible f, we observe poorer performance. We also tend to observe better performance at shorter window intervals τ. The variation with τ and f shows systematic trends in performance across subjects in this mirroring task.

We characterize the typical form of the transition between high and low performance by inspecting the level contours of the aggregated performance landscape in detail in figure 1*d*. A simple parametrization for the level contours is the nonlinear, inverse relation

$$f = a - \frac{b}{\tau}, \tag{2.1}$$

where a and b are constants. This form captures the fact that for large τ performance must become a linear function of f—because performance becomes an average between long visible and invisible windows—and captures the intuition that as τ decreases subjects do better because the rapid, intermittent views simulate stop motion animation. Fitting to the level contour $\tilde{\pi} = 1/2$ on the aggregated landscape, we find that nearly all the landscapes we consider are well captured by equation (2.1) and better than by a linear relation between f and τ. The results of the best-fit parameters are shown on the landscape in figure 1 (blue line, electronic supplementary material, section S7). Thus, the shape of the transition region shows that faster windows typically increase the range of f where good performance is accessible across subjects.

To determine if audio cues can affect performance, we introduce the Audio experimental condition where subjects hear a tone whose frequency increases with the speed of the avatar's hand (Material and methods). Though the tone does not provide directional information, it can be used to deduce when the avatar is making long sweeping motions or changing directions. We compare Audio with the Train condition, where subjects first undergo a 5 min practice version of the experiment. Finally, we combine these two changes in the Train + Audio condition in which subjects are reminded to use and coached on how to use the audio signals. This schema gives four different experimental conditions with N subjects and M unique subject and hand combinations: Visual Only ($N = 10$, $M = 17$), Audio ($N = 10$, $M = 10$), Train ($N = 7$, $M = 13$) and Train + Audio ($N = 8$, $M = 15$).[2]

The presence of audio and training enhances average performance taken over the predicted performance landscapes across subjects $\langle \pi \rangle (\epsilon^*)$. Since this measure depends on ϵ^*, we lower ϵ^* to assess how well subjects track the motion of the avatar at shorter time scales. We expect to find that the points converge at large and very small ϵ^* corresponding to the regimes of generous time delay where subjects do equally well and the limits of human reaction time where subjects perform equally poorly, respectively. Across nearly the entire range shown, where ϵ^* varies from 0.1 s to 1 s, we find large improvement from the Visual Only to all other conditions as shown in figure 2. When comparing the other conditions with each other in the intermediate regime (circled regions), we observe significant differences in the mean performance of up to approximately 25% with the highest performance consistently in the Train + Audio condition. Interestingly,

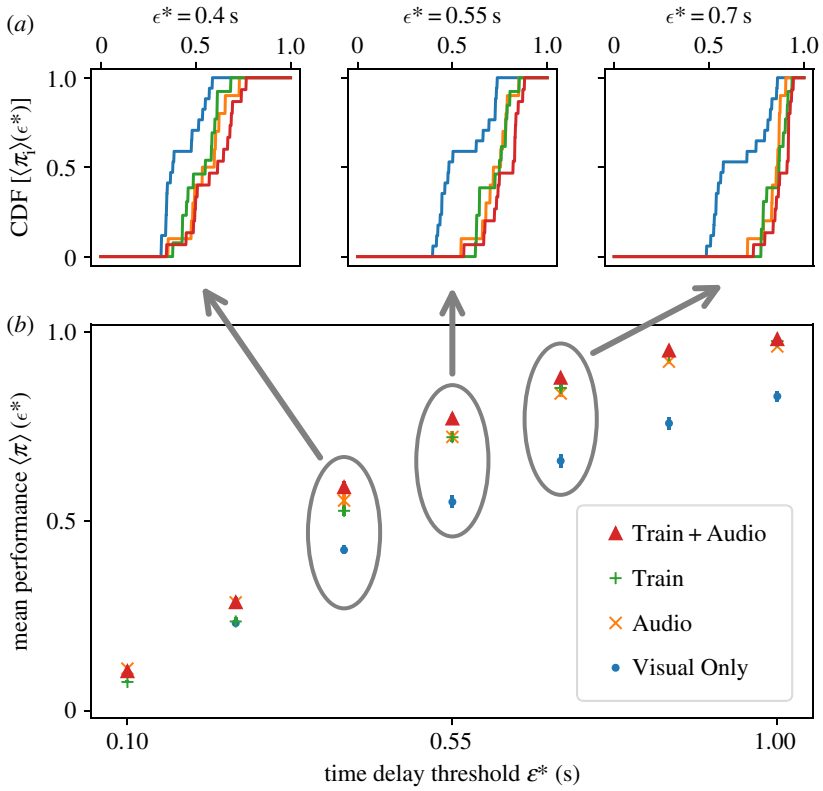

Figure 2. Mean of the predicted performance landscape as a function of the time error threshold ϵ^*. (*b*) The mean over performance landscapes $\langle \pi \rangle_{\epsilon^*} = (1/M) \sum_i (1/145) \sum_{(\tau,f)} \pi_i(\tau, f; \epsilon^*)$. At the physical limit of reaction times of small ϵ^*, performance converges. For large ϵ^*, performance again converges except for untrained individuals who do poorly in general. For 0.1 s $\leq \epsilon^* \leq$ 0.4 s, mean performance in Audio becomes better than Train, evidence that audio enhances performance at faster time scales. Error bars are two standard deviations of the mean by subject normalized by the square root of the number of subjects and samples $\sqrt{16N}$ and are small, reflecting precise estimates. (*a*) Variation in performance across subjects by cumulative distribution functions (CDF) of the average performance per subject i, $\langle \pi_i \rangle(\epsilon^*) = (1/145) \sum_{(\tau,f)} \pi_i(\tau, f; \epsilon^*)$. (Online version in colour.)

at $\epsilon^* \approx 1/2$ s the order of the mean performance of the Train and Audio conditions reverses suggesting that audio cues enhance mean performance more at shorter time scales. This result is consistent with studies showing that human reaction times to audio cues are faster than reactions to visual cues [27], as if effects of training were visually mediated or if training engaged higher-level anticipatory motor responses acting at slower time scales [19,20]. Collectively, these data demonstrate that for time scales spanning up to four times the human motor response time, from 200 to 800 ms, there is notable variation in performance depending on whether or not audio cues and training are provided.

To gain insight into what distinguishes good performers across conditions, we investigate the dynamics of how individuals mirror the avatar. We inspect runs of successful mirroring that are indicated by the shaded regions in figure 1*b*. Each of these runs has a duration t. When subjects are able to mirror the avatar closely, they show two kinds of dynamics: either long runs of close mirroring or a dearth of immediate failures (electronic supplementary material, section S8). We map where these behaviours appear in the parameter space given the condition of high performance $\hat{\pi} \geq 1/2$ (figure 3). We plot in blue where at least one high-performance trial appears on the performance landscape for the Train + Audio condition. We plot in red where at least one high-performance trial appears in the Audio or Train conditions. Where blue and red overlap, we colour the grid grey. For this comparison, we ignore the Visual Only condition where average performance is clearly poor. At $\epsilon^* = 0.7$ s, we inspect the region in the bottom right corner where visual gaps are the largest. To identify this region, we draw a line with the form of equation (2.1) for which it is significantly more probable to find a high-performance trial from Train + Audio ($p = 0.18 \pm 0.07$) than from Train or Audio ($p = 0.07 \pm 0.04$) as given by probabilities estimated with the Jeffreys prior and 90% confidence intervals. As we decrease ϵ^* to 1/2 s, this region is still significantly dominated by Train + Audio high-performance trials. This effect is no longer significant once $\epsilon^* = 0.4$ s, where high-performance trials are rare across all conditions. This asymmetry in the distribution across parameter space spanned by high-performance trials shows that for a limited range of ϵ^* some subjects in the Train + Audio condition are able to maintain stable mirroring under more difficult scenarios than subjects from the other experimental conditions.

3. Discussion

How might the few high performers in Train + Audio that do well across the extended parameter range be doing better? One explanation is that they reflect natural variation in the population and are affected by neither training nor audio cues [3]. Although this is a possibility we could only rule out completely by testing the same subjects *de novo* under multiple conditions, the significant increase in typical performance over many subjects implies that our experimental conditions are changing behaviour. Thus, these dynamic signatures provide evidence that the highest performing individuals are better at learning to use information from the audio cues to mirror the avatar, enabling them to perform well in regions of parameter space

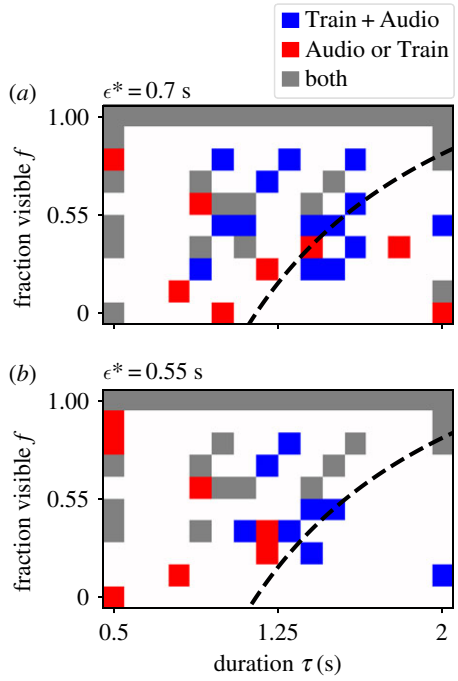

Figure 3. Combination of audio cues and training expands the region of parameter space accessible to stable mirroring in the regime of $\sim$1/2 s. Red indicates where at least one sticky or robust trial with high performance ($\hat{\pi} \geq 1/2$) appears in the Audio or Train conditions. Blue indicates the same for the Train + Audio condition. Grey indicates where red and blue overlap. Black dashed lines denote areas below which there are significantly more trials in the Train + Audio condition than the other two conditions using the Jeffreys prior with p-value < 0.1 and parametrized by the functional form given in equation (2.1). (Online version in colour.)

inaccessible to low performers. Similar examples of high performers have been identified in a number of experiments involving individuals mirroring the behaviour of another subject between a leader and a follower and in joint coordination without a designated leader [21,28,29]. The enhanced stability of mirroring runs that we observe is consistent with studies showing that short training sessions reduce error rates and temporal variability in motor tasks [16]. We find that the observed dynamics have signature distributions of temporal variation that have not been explored in other experiments though studies have shown that dancers show stronger coherence when following new motions versus non-dancers over a range of time scales [28]. Our experiments also reveal that high performance is facilitated by audio cues, consistent with prior work showing that auditory information enhanced entrainment [30] as well as a performance at joint coordination between individuals conducting complementary actions [15,31]. In the context of these previous studies, our results suggest that people can be trained to use audio cues to perform coordination tasks in regimes where visual cues are sparse.

Our results show that even simple low-dimensional, scalar representations of three-dimensional motion using pitch can enhance the ability to mirror. Similarly, one experiment showed that sonification of a cyclic target motion along with sonification of the subject's own motion could have a beneficial effect on learning greater than natural sounds of motion [14]. Other experiments likewise support the observation that auditory cues can enhance the learning of motor tasks [15]. Furthermore, pitch information has been found to substitute for missing visual information in motor perception, consistent with our finding of enhanced performance [32]. Interestingly, subjects in another experiment responding freely to music showed no association between pitch and speed of hand in one experiment, but our results show that a pitch-to-speed mapping can help when subjects are instructed to use such a mapping. Indeed, humans are adept at recognizing abstracted motion in various representations through both visual and auditory modalities [12,33] (see [12] for an overview). Flexibility in how human motion can be encoded suggests that our approach may be just one way of generating helpful auditory cues for mirroring tasks [12,34]. More broadly, frequency coding of motion is a commonly used approach for representing kinematic and motion values in human experiments [11]. In contrast with experiments using multidimensional encoding of motion [12,15], we give a simple representation only mapping the speed of the avatar's hand to the frequency of a pure tone such that faster speed corresponds to a higher frequency.

We find that overall performances in the Audio and Train conditions are quite similar despite small but significant differences in the time scales at which performance is improved. One explanation for this difference is that responses to visual stimuli are slower than to auditory stimuli as measured by reaction times [27], assuming that performance in the Train condition is visually mediated. Indeed, many subjects in the Visual Only condition had a difficult time responding to changes in the direction of the avatar's hand, showing considerable latency that lessened with the provision of auditory cues. Another possible explanation for the difference between Audio and Train is that training engages higher-level cognitive processes that act at slower time scales. Subjects in the Train condition had verbal reinforcement and a brief conversation to talk through the task with the experimenter, perhaps engaging higher-level cognitive functions for motor planning and social context [19,20,35]. Furthermore, familiarity with a motor task can improve performance [36–38]. We note that the presence of longer time scales over which subjects are able to mirror the avatar well in the Train condition compared to the Visual Only condition suggests that subjects are engaging cognitive processes with commensurate time scales going beyond visual reaction times. When we additionally include audio in the Train + Audio condition, stable trajectories appear over many seconds and many changes in direction, posing an interesting question: in which aspects is time-consuming training substitutable with intuitive perceptual cues?

Our study is just one example of how virtual reality technology combined with a set of statistical learning tools can advance the study of human behaviour [39]. An analogous toolkit has been used in cognitive neuroscience where control over the sensory apparatus in model systems has led to significant advances in the understanding of cognitive mechanisms [40–42]. To adapt this approach for human subjects, we used learning techniques to cover quickly a large parameter range across four different conditions over an order of magnitude in visual duration. Similar expansive experiments for mapping multiple conditions and parameters could be used to explore the efficacy of machine–human interfaces [43,44], determine parameters for athletic performance and diagnose motor or cognitive conditions with characteristic dynamics [45]. In the context of this study, this combination of techniques has been used to illustrate how visual perception can be augmented with audio signals to enhance coordination. Such developments could prove useful for medical teams synchronizing

different tasks, enhancing the fluidity of human–robot interactions, or even learning to improve one's tango.

4. Material and methods

All subjects were informed about the purpose and goal of the study at the beginning of the experiment and gave consent. After a preliminary survey about the experience in sports or performing arts and questions about any conditions that would exclude them from the study (including vision, hearing and arm motion problems and history of poor experience with virtual reality headsets), they were shown how to use the motion capture suit and virtual reality headset comfortably. The subject was familiarized with the mirror game outside of the virtual reality environment through two quick practice rounds (one hand at a time) with the researcher. Subjects were then instructed to 'mirror [simultaneously] the motion, or velocity, of the avatar' where the word 'simultaneously' was included in the training conditions because it was unclear if all subjects understood what was implied by mirroring in the untrained conditions. When audio cues were used, they were also told, 'Try to use the sound to predict the motion of the avatar's hand.' Immediately previous to the start of the mirroring task, they were reminded visually by a floating script to 'Mirror the hand.' Periodically throughout the trial, the comfort of subjects in the virtual environment was assessed verbally. At the end of the experiment, all subjects filled out a post-experimental survey to assess the comfort of the suit and virtual headset, importance of fatigue, clarity of instructions and to check if they had been following instructions.

A sequence of trials for a single hand consisted of 16 different 30 s trials where the first and last trials were always a fully visible condition. During the experiments, the task was paused every 2–3 min to assess the subject for any poor reactions to the virtual environment and to ask explicitly about fatigue. If the subject expressed any sign of fatigue, a rest of time of at least 15 s was taken.

The avatar's hand motion were recordings of the entire body of the experimenter while he was moving his arm naturally. Correspondingly, there were small displacements of the shoulders, upper torso, hips and legs with manual corrections to obvious distortion using MotionBuilder. Arm movements were relatively slow ($<2\ \mathrm{m\ s^{-1}}$) and smooth with frequent pauses and changes in direction ($\sim$1 Hz) with a conscious attempt to avoid repetitive motions or meaningful shapes to which subjects might entrain. In the electronic supplementary material, we provide some more detail about the motion and a motion file is available online on the GitHub repository.

For the Train and Train + Audio conditions, subjects were told that the first 5 min of the experiment would consist of a practice round with a single break in the middle. During the break, subjects were asked if they had any questions about their performance. When audio cues were used, the experimenter emphasized the instruction to use the audio cue and asked the subjects to explain how they were using the audio cues. If they made incorrect inferences about how the audio corresponded to the motion—for example, one subject thought the volume of the audio changed with the location of the avatar's hand—the experimenter explained to them how they were incorrect. To all subjects, the experimenter explained that the audio cue had pitch proportional to the speed of the avatar's hand and became higher in pitch when the avatar was moving faster and lower when the avatar was slowing down or changing directions. The tone's instantaneous frequency $F(t)$ increases with the speed of the avatar's hand $|v_a(t)|$ as

$$F(t) = c_1 e^{|v_a(t)|} + c_0, \tag{4.1}$$

where c_0 and c_1 were chosen to keep the frequency between 100 and 340 Hz, where the bounds were chosen to maximize the easily audible range while minimizing discomfort. This exponential dependence of the instantaneous frequency on the speed of motion ensures that changes in the velocity of the avatar's hand are clearly distinguishable by frequency, a scaling distinct from Weber's law [46].

We collected data from 35 participants, but one subject was excluded from the analysis because of professed disinterest in the experiment and cursory completion of the post-experimental survey that included answering an inapplicable question without any mention or question to the experimentalist. Subjects ranged in ages from 18 to 42 with varying levels of experience in physical activities requiring coordination with others. Experimental protocol was approved by the institution's IRB and the HRPO at the DoD.

For aligning the velocity trajectories, we use DTW with a cost function for the trajectory comparing times with indices i and j,

$$g(\mathrm{i},\mathrm{j}) = \begin{cases} 0, & \left|t_\mathrm{i} - t_\mathrm{j} + \frac{1}{2}\right| < 1 \\ \left|t_\mathrm{i} - t_\mathrm{j} + \frac{1}{2}\right|^6, & \left|t_\mathrm{i} - t_\mathrm{j} + \frac{1}{2}\right| \geq 1. \end{cases} \tag{4.2}$$

To control the strength of this regularization, we set the coefficient of g to be $\lambda = 10^{-3}$ in the minimized objective function (electronic supplementary material, section S3). We first use FastDTW which can calculate the time warp in nearly linear time instead of quadratic time [23]. If the found trajectory ventures outside of the bounding interval $\Delta t \in [-1/2\ \mathrm{s}, 3/2\ \mathrm{s}]$, we then solve the problem using our own (slower) implementation including the regularization specified in equation (4.2). We find that about 60% of the untrained trials were regularized whereas only 35% of the trained trials were. We might expect this difference because untrained individuals typically do not replicate the trajectory of the avatar as well and the algorithm is more prone to misaligning stretches of motion.

To measure mirroring error, we measure the fraction of time that the subject is within some time delay ϵ^* measured from alignment with DTW:

$$\hat{\pi}(\tau, f; \epsilon^*) = \frac{1}{\tilde{T} + 2}\left(1 + \sum_{\tilde{t}}^{\tilde{T}} \Theta\left[\epsilon^* - |\epsilon\ (\tilde{t})|\right]\right), \tag{4.3}$$

which is regularized by the Laplace counting estimator. The indicator function, given by the Heaviside theta function $\Theta\ (x \geq 0) = 1$ and $\Theta\ (x < 0) = 0$, counts when the subject is within or beyond temporal error threshold. We use the warped time $\tilde{t}$ and normalize by the length of the warped trajectory $\tilde{T}$.

The distributions of durations of mirroring runs are given by three classes: an exponential, a 'sticky' gamma-like function with a dearth of the shortest decay times, and a heavy-tailed 'robust' distribution. Although the exponential decay is a signature of a memoryless process, the remaining two distributions suggest that the dynamics of how subjects are tracking the motion of the avatar are generated from a history-dependent process.

The 'sticky' distribution is described by the complementary cumulative distribution function (CDF) of decay times, otherwise known as the survival function, as a function of a single rate constant K

$$1 - \mathrm{CDF}\ (t') = e^{-Kt'} \sum_{n=0}^{N} \frac{K^n t'^n}{n!}. \tag{4.4}$$

In the limit of $N \to \infty$, we recover the gamma distribution. We find that the measured values of N as calculated with maximum likelihood are concentrated at smaller values. Over 50% of the observed values are smaller than or equal to 5 when $\epsilon^* = 1/2$ s, suggesting that enhanced dynamical stability corresponding to the 'sticky' distribution is slight. The 'robust' distribution describes the first passage time for simple

diffusion,

$$1 - \text{CDF}(t') = 1 - \sqrt{\frac{\alpha}{\pi}} \int_{1/30}^{t'=t\alpha^*/\alpha} t^{-3/2} e^{-\alpha/t}\, dt. \qquad (4.5)$$

Here, the lower limit is important and is given by our interpolation of the velocity trajectories at 30 Hz (see electronic supplementary material to find more information about methods).

Data accessibility. Project code and data are available at https://doi.org/10.17605/OSF.IO/58J6T. The anonymized datasets generated during the current study are provided.

Authors' contributions. E.D.L. and I.C. designed the study. E.D.L. and E.E. ran the experiments. E.D.L. and I.C. wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. The current work is supported by the ARO through award 69189-NS-II. E.D.L. was supported by an NSF graduate fellowship under grant no. DGE-1650441. I.C. was supported in part by a Feinberg Fellowship and a Braginsky Grant.

Acknowledgements. We thank Guy Hoffman, Giles Hooker, Joe Guinness, Lena Bartell, Uri Alon and his group, and Tamar Flash and her group for helpful and inspiring discussion; Jim Jing for helping us experiment with the Vicon system; Caeli MacLennan, Saerom Choi and Vincent Chen for contributing to our experimental apparatus; and all the volunteers for our experiments.

Endnotes

[1]We do not consider the *x*-axis which points from the subject to the avatar. This axis is particularly problematic for the motion capture system that we used and we found that timing errors could be significant (electronic supplementary material, section S3).

[2]Some pairs of subjects and hands were not considered because of errors in the code.

References

1. Dyer JR, Ioannou CC, Morrell LJ, Croft DP, Couzin ID, Waters DA, Krause J. 2008 Consensus decision making in human crowds. *Anim. Behav.* **75**, 461–470. (doi:10.1016/j.anbehav.2007.05.010)
2. Konvalinka I, Vuust P, Roepstorff A, Frith CD. 2010 Follow you, follow me: continuous mutual prediction and adaptation in joint tapping. *Q. J. Exp. Psychol.* (*Colchester*) **63**, 2220–2230. (doi:10.1080/17470218.2010.497843)
3. Pecenka N, Keller PE. 2011 The role of temporal prediction abilities in interpersonal sensorimotor synchronization. *Exp. Brain Res.* **211**, 505–515. (doi:10.1007/s00221-011-2616-0)
4. Vesper C, Knoblich G, Sebanz N. 2014 Our actions in my mind: motor imagery of joint action. *Neuropsychologia* **55**, 115–121. (doi:10.1016/j.neuropsychologia.2013.05.024)
5. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006 The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765. (doi:10.1037/0033-295X.113.4.700)
6. Brunton BW, Botvinick MM, Brody CD. 2013 Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95–98. (doi:10.1126/science.1233912)
7. Wolpert DM, Ghahramani Z, Jordan MI. 1995 An internal model for sensorimotor integration. *Science* **269**, 1880–1882. (doi:10.1126/science.7569931)
8. Goodbody SJ, Wolpert DM. 1999 The effect of visuomotor displacements on arm movement paths. *J. Stat. Phys.* **127**, 213–223. (doi:10.1007/s002210050791)
9. Scheidt RA, Conditt MA, Secco EL, Mussa-Ivaldi FA. 2005 Interaction of visual and proprioceptive feedback during adaptation of human reaching movements. *J. Neurophysiol.* **93**, 3200–3213. (doi:10.1152/jn.00947.2004)
10. Young W, Rodger M, Craig CM. 2013 Perceiving and reenacting spatiotemporal characteristics of walking sounds. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 464–476. (doi:10.1037/a0029402)
11. Dubus G, Bresin R. 2013 A systematic review of mapping strategies for the sonification of physical quantities. *PLoS ONE* **8**, e82491. (doi:10.1371/journal.pone.0082491)
12. Vinken PM, Kröger D, Fehse U, Schmitz G, Brock H, Effenberg AO. 2013 Auditory coding of human movement kinematics. *Multisens. Res.* **26**, 533–552. (doi:10.1163/22134808-00002435)
13. Dyer J, Stapleton P, Rodger M. 2017 Transposing musical skill: sonification of movement as concurrent augmented feedback enhances learning in a bimanual task. *Psychol. Res.* **81**, 850–862. (doi:10.1007/s00426-016-0775-0)
14. Effenberg AO, Fehse U, Schmitz G, Krueger B, Mechling H. 2016 Movement sonification: effects on motor learning beyond rhythmic adjustments. *Front. Neurosci.* **10**, 67. (doi:10.3389/fnins.2016.00219)
15. Hwang T-H *et al.* 2018 Effect- and performance-based auditory feedback on interpersonal coordination. *Front. Psychol.* **9**, 404. (doi:10.3389/fpsyg.2018.00404)
16. Scheurich R, Zamm A, Palmer C. 2018 Tapping into rate flexibility: musical training facilitates synchronization around spontaneous production rates. *Front. Psychol.* **9**, 1. (doi:10.3389/fpsyg.2018.00458)
17. Schmidt RC, Richardson MJ. 2008 Dynamics of interpersonal coordination. In *Coordination: neural, behavioral and social dynamics*, pp. 281–308. Berlin, Germany: Springer.
18. Vesper C, van der Wel RP, Knoblich G, Sebanz N. 2013 Are you ready to jump? predictive mechanisms in interpersonal coordination. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 48–61.(doi:10.1037/a0028066)
19. Kilner JM, Vargas C, Duval S, Blakemore SJ, Sirigu A. 2004 Motor activation prior to observation of a predicted movement. *Nat. Neurosci.* **7**, 1299–1301. (doi:10.1038/nn1355)
20. Kourtis D, Sebanz N, Knoblich G. 2010 Favouritism in the motor system: social interaction modulates action simulation. *Biol. Lett.* **6**, 758–761. (doi:10.1098/rsbl.2010.0478)
21. Noy L, Dekel E, Alon U. 2011 The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proc. Natl Acad. Sci. USA* **108**, 20 947–20 952. (doi:10.1073/pnas.1108155108)
22. Itakura F. 1975 Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust.* **23**, 67–72. (doi:10.1109/TASSP.1975.1162641)
23. Salvador S, Chan P. 2007 Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **11**, 561–580. (doi:10.3233/IDA-2007-11508)
24. Müller M. 2007 *Information retrieval for music and motion*. Berlin, Germany: Springer.
25. Bishop CM. 2006 *Pattern recognition and machine learning*. Singapore: Springer Verlag.
26. Rasmussen CE, Williams CKI. 2006 *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
27. Shelton J, Kumar GP. 2010 Comparison between auditory and visual simple reaction times. *Neurosci. Med.* **1**, 30–32. (doi:10.4236/nm.2010.11004)
28. Washburn A, DeMarco M, de Vries S, Ariyabuddhiphongs K, Schmidt RC, Richardson MJ, Riley MA. 2014 Dancers entrain more effectively than non-dancers to another actor's movements. *Front. Hum. Neurosci.* **8**, 1. (doi:10.3389/fnhum.2014.00800)
29. Caramiaux B, Bevilacqua F, Palmer C, Wanderley M. 2017 Individuality in piano performance depends on skill learning. In *Proc. 4th Int. Conf. on Movement Computing, London, UK, 28–30 June 2017*, article 14. New York, NY: ACM. (doi:10.1145/3077981.3078046)
30. Richardson MJ, Marsh KL, Isenhower RW, Goodman JR, Schmidt RC. 2007 Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Hum. Mov. Sci.* **26**, 867–891. (doi:10.1016/j.humov.2007.07.002)

31. Knoblich G, Jordan JS. 2003 action coordination in groups and individuals: learning anticipatory control. *J. Exp. Psychol. Learn. Mem. Cogn.* **29**, 1006–1016. (doi:10.1037/0278-7393.29.5.1006)
32. Effenberg AO, Schmitz G. 2018 Acceleration and deceleration at constant speed: systematic modulation of motion perception by kinematic sonification. *Ann. N. Y. Acad. Sci.* **1425**, 52–69. (doi:10.1111/nyas.2018.1425.issue-1)
33. Johansson G. 1973 Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 201–211. (doi:10.3758/BF03212378)
34. Bidet-Caulet A, Voisin J, Bertrand O, Fonlupt P. 2005 Listening to a walking human activates the temporal biological motion area. *NeuroImage* **28**, 132–139. (doi:10.1016/j.neuroimage.2005.06.018)
35. Wolpert DM, Doya K, Kawato M. 2003 A unifying computational framework for motor control and social interaction. *Phil. Trans. R. Soc. Lond. B* **358**, 593–602. (doi:10.1098/rstb.2002.1238)
36. Jeannerod M. 2001 Neural simulation of action: a unifying mechanism for motor cognition. *NeuroImage* **14**, S103–S109. (doi:10.1006/nimg.2001.0832)
37. Ramnani N, Miall RC. 2004 A system in the human brain for predicting the actions of others. *Nat. Neurosci.* **7**, 85–90. (doi:10.1038/nn1168)
38. Calvo-Merino B, Grèzes J, Glaser DE, Passingham RE, Haggard P. 2006 Seeing or doing? Influence of visual and motor familiarity in action observation. *Curr. Biol.* **16**, 1905–1910. (doi:10.1016/j.cub.2006.07.065)
39. Minderer M, Harvey CD, Donato F, Moser EI. 2016 Virtual reality explored. *Nature* **533**, 324–325. (doi:10.1038/nature17899)
40. Aronov D, Tank DW. 2014 Engagement of neural circuits underlying 2D spatial navigation in a rodent virtual reality system. *Neuron* **84**, 442–456. (doi:10.1016/j.neuron.2014.08.042)
41. Morcos AS, Harvey CD. 2016 History-dependent variability in population dynamics during evidence accumulation in cortex. *Nat. Neurosci.* **19**, 1672–1681. (doi:10.1038/nn.4403)
42. Stowers JR *et al.* 2017 Virtual reality for freely moving animals. *Nat. Methods* **14**, 995–1002. (doi:10.1038/nmeth.4399)
43. Wickens CD. 2002 Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* **3**, 159–177. (doi:10.1080/14639220210123806)
44. Iqbal T, Gonzales MJ, Riek LD. 2015 Joint action perception to enable fluent human–robot teamwork. In *Proc. 24th IEEE Int. Symp. on Robot and Human Interactive Communication, Kobe, Japan, 31 August–4 September 2015*, pp. 400–406. (doi:10.1109/ROMAN.2015.7333671)
45. Wu D, José JV, Nurnberger JI, Torres EB. 2018 A biomarker characterizing neurodevelopment with applications in autism. *Sci. Rep.* **8**, 614. (doi:10.1038/s41598-017-18902-w)
46. Wier CC, Jesteadt W, Green DM. 1977 Frequency discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.* **61**, 178–184. (doi:10.1121/1.381251)