Covert variations of a musician's loudness during collective improvisation capture other musicians' attention and impact their interactions

Armand Schwarz, Arthur Faraco, Coralie Vincent, Patrick Susini, Emmanuel Ponsot, Clément Canonne

Covert variations of a musician's loudness during collective improvisation capture other musicians' attention and impact their interactions

Armand Schwarz[1,2], Arthur Faraco[3], Coralie Vincent[1], Patrick Susini[1], Emmanuel Ponsot[1] and Clément Canonne[1]*

1- STMS UMR 9912 (CNRS/IRCAM/Sorbonne Université), Paris, France
2- Institut Pasteur (Université Paris Cité/INSERM/Institut de l'Audition/IHU Reconnect), Paris, France
3- Universidade de Sao Paulo, Sao Paulo, Brazil

* corresponding author: clementcanonne@hotmail.com

Funding. This work was supported by the INSEAD-Sorbonne Université Behavioural Lab and by the Collegium Musicae.

Acknowledgments. The authors thank all the wonderful improvisers who participated in this study.

Data availability statement. All data collected for the study and code used in the analyses are available at https://osf.io/bsgx8/.

Authors contributions. All authors designed the study. A.S. prepared the experimental material and programmed the software used in the experiment. C.C., A.S., A.F., and C.V. collected the data. C.C., E.P., and A.S. prepared the analysis plan. A.S. analyzed the data. A.F. ran the cross-recurrence quantification analyses. C.C. wrote the paper with edits and comments from E.P.

Ethical approval and consent. Ethical approval for this study was obtained at INSEAD/ Sorbonne University Center for Behavioural Science, Paris, France. All methods were carried out in accordance with their guidelines and regulations. All participants signed an informed consent.

Covert variations of a musician's loudness during collective improvisation capture other musicians' attention and impact their interactions

0. Abstract

While research on auditory attention in complex acoustical environment is a thriving field, experimental studies thus far have typically treated participants as passive listeners. The present study – which combined real-time covert loudness manipulations and online probe detection –investigates for the first time the effects of acoustic salience on auditory attention during live interactions, using musical improvisation as an experimental paradigm. We found that musicians were more likely to pay attention to a given co-performer when this performer was made sounding louder or softer; that such salient effect was not due to the local variations introduced by our manipulations but rather likely to be driven by the more long-term context; and that improvisers tended to be more strongly and more stably coupled when a musician was made more salient. Our results thus demonstrate that a meaningful change of the acoustical context not only captured attention but also impacted the ongoing musical interaction itself, highlighting the tight relationship between attentional selection and interaction in such social scenario, and opening novel perspectives to address whether similar processes are at play in human linguistic interactions.

1. Introduction

Research on auditory attention in complex acoustical environment is a thriving field. Starting with Cherry's seminal paper (1953) – which coined the expression "cocktail party effect" to describe our ability to focus on selected aspects of a given acoustic scene while blocking out non-relevant sonic streams – an impressive amount of psychophysical as well as neuroimag-

ing research has been conducted on both sides of the "cocktail party problem" (McDermott, 2009): how do we segregate concurrent sonic streams that are somehow mixed?; and how do we direct our attention to a source of interest while ignoring other sources?

Strikingly, the whole field was thus started by the description of an imaginary situation – the cocktail party – in which agents *participate* to the acoustic scene in which they are immerged in. They are not merely passive listeners, observing the scene from afar; rather, they direct their attention from one source to another as part of broader dialogic interests – namely, identifying the conversation(s) they are going to engage with, thus adding their own voice to the overall acoustic scene. However, this interactional aspect has all but disappeared from extent psychophysical studies addressing the cocktail party problem from the perspective of speech intelligibility with multiple concurrent talkers. Baring a few recent exceptions of a rather observational nature (e.g., Miles et al., 2023 or Ryan et al., 2023), it is fair to say that most studies approach speech-on-speech situations from an experimental perspective in which the participants are treated as external listeners of a scene rather than as agents *involved* in that scene (see Bidelman & Yoo, 2020 for a recent example). While there are obvious methodological reasons for this state of affair (studying participants' auditory attention in sound-proof booths using precisely calibrated sounds presented through headphones or spatialized speakers allows for maximum experimental control), disconnecting attention from interaction might also come at a cost. This is particularly clear when it comes to the problem of selective auditory attention – whose underlying mechanisms are arguably less studied than the mechanisms allowing for smooth sound segregation (McDermott, 2009). Selective auditory attention is indeed often seen as being strongly impacted by bottom-up factors – often described by the generic term "salience" to refer to acoustical events that draw auditory attention independently of volition, either because they possess acoustic characteristics that are known to catch auditory attention such as high levels of sound intensity or roughness (Arnal et al., 2019) or because

they stand out from the local sonic context (Dalton & Lavie, 2004). However, one might wonder whether stimulus-driven factors such as acoustic salience still play a significant role in shaping auditory attention over time in interactional contexts in which volitional attentional selection processes might seem to be predominant (Koch et al., 2011). Moreover, it could be the case that agents are typically less sensitive to local variations in the acoustic context in settings which are more cognitively demanding (as interactional settings typically are, comparatively to more passive settings), in line with already established effects of cognitive load on sound evaluation or auditory susceptibility (see Steffens et al., 2020; Van der Heiden et al., 2023). This would constitute yet another case of a finding on social attention obtained from studies in which participants merely observe an interaction between agents (for example by relying on pictures or videos) that does not straightforwardly generalize to participants' behaviors during live interactions (Dawson & Foulsham, 2022).

The present paper precisely seeks to investigate the impact of such acoustic salience on auditory attention during live interactions, by introducing a novel experimental paradigm involving musicians freely improvising together in a trio setting. Music provides a highly ecological setting – and yet still largely unexplored (but see Faraco et al., 2024, for a recent example) – to study selective auditory attention in interactive contexts, as music is, most often than not, a collective affair (Blacking, 1973), presenting simultaneously different instrumental or vocal parts. In performances, this means that musicians must navigate within a complex acoustic environment, focusing their attention on such or such aspect of the sonic tapestry collectively produced to fulfill various performance goals – which typically means focusing their auditory attention on such or such musician (or sub-group of musicians) (Keller 2001). Amongst the wide variety of collective musical practices, free improvisation is a particularly interesting case for our purpose, if only because joint improvisation plays such a central role in most of our social interactions (Noy et al., 2011). In collective free improvisation (CFI from now on),

musicians aim at spontaneously creating music without relying on pre-defined plans or pre-existing musical structures (Saint-Germier & Canonne 2020). CFI is thus in stark contrast with score-based, well-rehearsed performances, in which the distribution of musicians' auditory attention is likely to be mostly guided by the structural information contained within the score (e.g., where the melodic part lies) or the "ideal sound" and other performance goals that are slowly built through rehearsals. In CFI, on the contrary, how musicians distribute their auditory attention is an integral part of the interactional dynamics, and plays a crucial role in shaping how the performance will unfold (Clarke, 2005).

But while CFI might provide a highly relevant setting to assess whether acoustic salience play a significant role in driving musicians' auditory attention, a crucial problem is to find a way to track such auditory attention over the course of the performance. Recent advances in auditory attention decoding, based on the modeling of neural information collected through mobiles EEG (Straetmans et al., 2022) offer promising perspectives, but we are still far from being able to use such methods in a setting as sonically and interactionally complex as collective free improvisation. Relying on post-hoc verbalizations through interviews with musicians (Seddon, 2005) is of course also an option but such verbalizations tend to focus on conscious decisions made by the musicians (Canonne & Garnier, 2012), which might make the methodology ill-suited to study non-intentional shifts of musicians' auditory attention. In the following study, we thus relied on a behavioral approach, using a probe-based method which allowed us to assess, each time a probe is sent, whether or not a given musician in a trio was paying close auditory attention to one of the two other musicians. This probe-based method was then combined with real-time manipulations of the salience of the acoustic signals produced by the musicians. As they were playing together, and unknowingly to them, the musicians' individual signals were selectively made, for short periods of time of a few seconds, either a bit louder or a bit softer – a manipulation which presents the advantage of being ap-

plicable in the same way to every musical instrument, contrary to manipulations that would involve, e.g., real-time manipulation of the spectral information. This novel experimental paradigm thus made it possible to systematically study, for the first time, whether and how a change in loudness – which has been shown in previous studies to act as a salient cue (Dalton & Lavie, 2004) – could shape selective auditory attention in a context in which participants are able to freely interact with one another. On the one hand, contrasting the effects of both "louder" and "softer" manipulations on musicians' attention allowed us to assess whether this loudness manipulation was more a matter of intrinsic acoustic cues (which would predict that louder events are more salient than softer events) or of contextual cues (which would predict that both louder and softer events can act as salient cues when they stand out of given context, as was found in Dalton & Lavie, 2004). On the other hand, our repeated loudness manipulations combined with our probe-based methodology – with probes being sent to participants both *during* the manipulations and at various points *after* the manipulations – allowed us to assess whether such potential salience effect would be rather driven by the short-term deviation in loudness or more long-term contextual changes in musicians' typical loudness.

But beyond studying the impact of loudness changes on musicians' attention, our paradigm also allowed us to directly assess whether such events would impact the ongoing interaction. Attention and action have indeed been found to be closely related (Humphreys et al., 2010), to the point that a recent influential philosophical account of attention has explicitly defined attention in terms of action selection (Wu, 2014). More specifically, studies have shown that acoustic salience within musical stimuli was associated with spontaneous muscle activity from the participants exposed to such stimuli (Schultz et al., 2021), and that joint attention increased feelings of connectedness between participants (Wolf et al., 2015). All this makes plausible that musicians would be more likely to interact more strongly with their co-improvisers when they are more acoustically salient, and our experimental design made it

possible to test for this hypothesis. In other words, by assessing its impact on both attention and interaction, our study investigates how acoustic salience can act as a full-fledged communicational strategy in social settings.

2. Methods

2.1. Participants

Fifteen musicians participated in the experiment (mean age = 40.5 years, SD = 9.5, 9 male, 2 female, 1 non-binary and 3 that did not provide the information), divided into five trios. They were highly trained musicians (with a mean of 28.8 years of musical practice, SD = 8.6 years) and had a significant experience with collective free improvisation (with a mean of 19 years of practice, SD = 9.5), which they all practiced in a professional proficiency (invitation to important festivals; numerous live and studio recordings published on well-regarded labels; etc.). They were recruited from the Parisian Free Improvisation scene, that comprises a high number of musicians with a wide variety of different stylistic backgrounds (Roueff, 2006).

The overall instrumentation was saxophone (N = 4), guitar (N = 3), trumpet (N = 2), drums, piano, clarinet, contrabass clarinet, double bass and electronics (N = 1). One participant also used their voice during the improvisations. In each trio, the three musicians always played different instruments in order to minimize timbral – and thus source – confusion for the participants.

Trios were intentionally composed to minimize potential effects of familiarity on selective attention (i.e., avoiding that participants paid more attention to the musicians they were used to play with). Consequently, the majority of musicians within each trio were unacquainted or had not previously collaborated together. Participants assessed their prior familiarity with the

other members of their trio using a 7-point Likert scale. As expected, mean familiarity was low (M = 2.06; *SD* = 0.93).

All participants gave their informed written consent and were compensated at the standard rate for the employment of professional musicians in France.

2.2. Procedure

The experiment was held in a professional recording studio. Each musician of the trio was allocated to an individual booth, and was equipped with headphones (Beyerdynamics DT 770 pro, 80 ohms) in order to listen to each other. Musicians could thus not see each other, so that communication between them would only be based on acoustic information. Importantly, musicians' headphones were panned in such a way that they heard one improviser completely on the right side and one completely on the left side, while hearing themselves in the center. The panning (i.e., which musician is heard on the right side, which is heard on the left side) was made randomly.

Each trio performed between four and six improvisations so as to improvise for a total of approximatively 40 minutes (mean = 42 minutes, SD = 2.5 minutes). They had a 30-minute break after having performed the first two or three improvisations, depending on the duration of those first improvisations. The recording was made on *Protools* (version 2022.9) by a professional sound engineer.

While the musicians were playing, we introduced at various points real-time manipulations of the RMS level (i.e., average loudness) of a given musician during approximatively a 10-second window. There were two distinct patterns of variation. Taking the musician's actual loudness as the baseline, the *louder* pattern featured an +5 dB increase in the signal's amplitude for a duration comprised between 1.76s and 3.13s (randomly drawn from a gaussian distribution with mean=2.5s and SD=0.25), sustained the heightened level for a

duration comprised between 4.29s and 5.69s (randomly drawn from a gaussian distribution with mean=5s and SD=0.25), and then decreased for a duration comprised between 1.85s and 3.27s (randomly drawn from a gaussian distribution with mean=2.5s and SD=0.25) back to the baseline. Conversely, the *softer* pattern started with a -5 dB decrease from the baseline for approximatively 2.5s (drawn from the same distribution as for the louder pattern), maintained the reduced level for approximatively 5s, and finally increased for approximatively 2.5s back to the baseline.

As these manipulations were repeated numerous times over a given performance, the variations of durations described above were meant to make the manipulations more unpredictable, and thus less recognizable. Moreover, since the musicians did not see each other while playing, and that their interactions was entirely mediated by the sounds perceived in their headphones, they had no way to know that the level heard in their headphones was not the one at which the musicians were actually playing. The 5 dB increase/decrease was pre-tested by the authors of the present paper, in order to identify a threshold value that would be both perceivable but would not sound too "unnatural" (see Supplementary Material for the presentation of a perceptual follow-up study which confirmed, in a post-hoc fashion, the naturalness of our manipulations).

In order to implement these manipulations, six different 15-minute automation tracks were created for each trio, each of them defining the moments in which these RMS manipulations would occur for each musician. For each musician A in a given trio, 2 automation tracks were created: one applied to B's signal as received in A's headphones and one applied to C's signal as received in A's headphones. Note that musicians always heard themselves as they actually played, i.e., without any RMS manipulation. A *Max/MSP* patch was designed to synchronize in real time those automation tracks with the ongoing *Protools* recording session.

The following four constraints were used in the random generation of the automation tracks for a given trio to minimize the impact of our manipulations on the ongoing group interaction, while still collecting as many data points as possible. First, a given musician never heard both of their co-improvisers modified at the same time. Second, for a given musician, RMS manipulations were never applied to a same co-improviser twice in a row. Third, if a given musician's RMS was being modified for their two co-improvisers, then one co-improviser heard that musician through a louder pattern, and the other through a softer pattern. Fourth, and finally, when considering the 6 automation tracks of a given trio as a whole, RMS manipulations occurred continuously, with a 1-second buffer period between two successive manipulations.

To track the auditory attention of the participant in real time, participants had to perform a detection task while improvising. Target sounds were indeed added at different times to the music improvised by the musicians and the musicians were asked to press a MIDI pedal (ROLAND DP-10 Piano style Sustain Pedal) each time they noticed one of those target sounds – thus triggering a signal recorded as a separate MIDI track in the Protools session. Musicians were free to place their pedal wherever they felt it would be more convenient for them to press it while playing. At the beginning of each session, a 3-minute training period was organized so that musicians could familiarize themselves with the target sounds as well as with the pedal pressing.

Each target sound was associated to a given musician, meaning that the two other musicians heard the target sound on the channel associated to that musician. Musicians also never heard the target sounds associated with themselves. For example, every time a target sound was associated to musician A, the two other musicians B and C heard the target sound on the side attributed to musician A in their headphones panning (i.e., either the left or the right side) while A did not hear it (see Figure 1).

The target sound was composed of two bursts of noise of 60ms each, with 40ms of attack and 20ms of release and without any inter-onset interval, so as to make detection harder. The first burst was filtered between 400 and 1600 Hz and the second one between 320 Hz and 1280 Hz, with a strong decrease outside of the passing band (~ -300 dB/octave). Importantly, the level of the target sound, was always 1.5 dB louder than the level of the musician's track at the moment at which it was embedded. This was implemented through a Max/MSP patch which performed a real-time analysis of the RMS level of all the recorded tracks during the performance. The choice of this value for the emergence level of the target sound was made on the basis of observations from the experimenters during a pilot session in which several values were tested, trying to ensure sufficient audibility (which was also assessed in a post-hoc fashion, see section 2.3 below) while limiting salience, and thus avoiding both floor and ceiling effects regarding target detection performance.

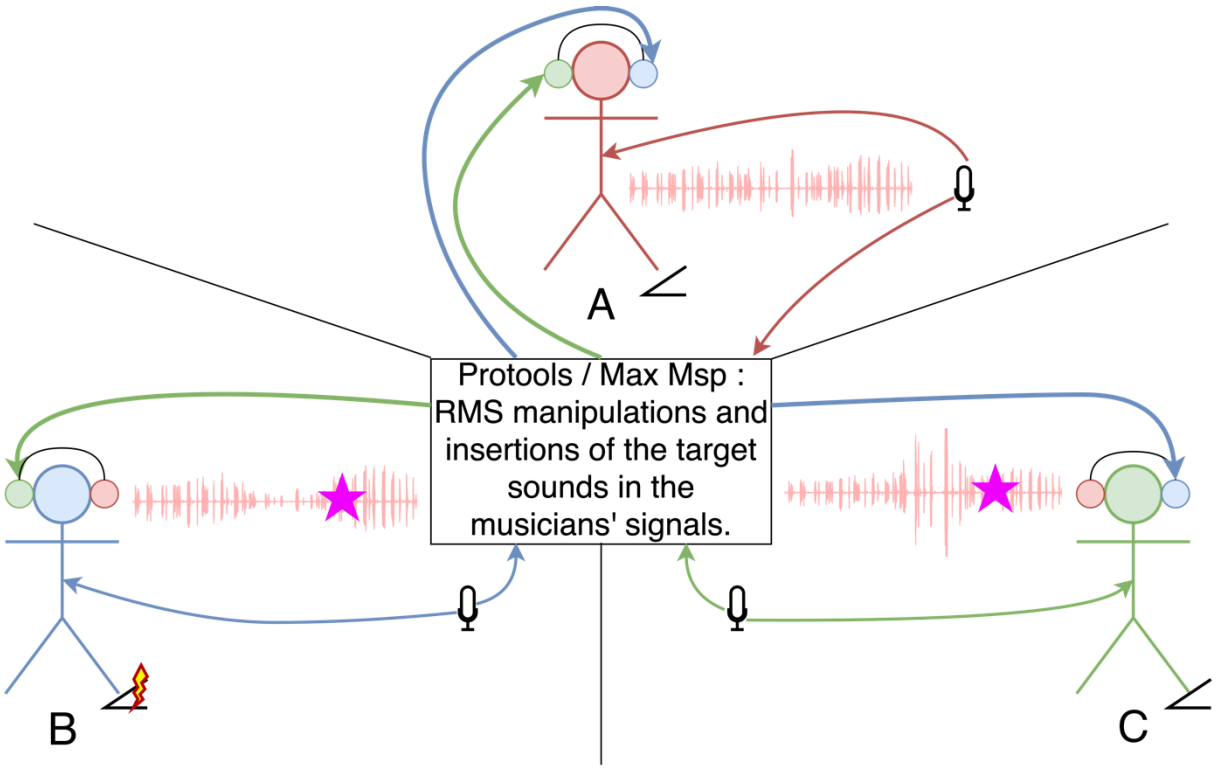

Figure 1. Experimental procedure. Musicians A, B and C are improvising together, without seeing each other and hearing each other through a panned mixed in their headphones. B hears A's signal through a "softer

pattern" real-time manipulation (see the smaller amplitude) and C hears A's signal through a "louder pattern" real-time manipulation (see the larger amplitude). A target sound (represented here by a purple star) is associated to musician A during these manipulations. B and C hear the target sound on the side on which they hear musician A (i.e., the left channel for musician B and the right channel for musician C). B detected the target sounds and pressed the pedal while C did not detect the target sound and did not press the pedal.

Target sounds appeared every 7 to 14 seconds (randomly drawn from a uniform distribution). Since musicians did not hear the target sounds associated to their own musical track, this means that, on average, a musician heard a target sound every 15.75 seconds. For each trio, three 15-minute sound files were created beforehand with all the target sounds associated to each of the three musicians. Each sound file was then superposed to the two different automation tracks associated with that musician (one for each of their co-improvisers within the trio) so that roughly 75% of the target sounds would be heard at moments where the level of the musician associated with the target sound was manipulated. The remaining target sounds were heard when no RMS manipulation was happening and were thus used as a baseline. On average, each participant heard 137 target sounds (SD = 20) over the duration of the whole experiment. This resulted in a total of 2170 target sounds to be detected by the musicians.

Importantly, in order to decrease potential learning effects, target sounds and RMS manipulations were not systematically associated and the distribution of the time intervals between two successive target sounds heard by a same participant was quite wide (see Supplementary Material for further analyses on potential learning effects). Furthermore, the first 30 seconds of each improvisation were without target sounds nor RMS modification, so that the musical interaction between the three musicians would begin in a way as natural as possible.

Two video examples of the improvisations produced during this experiment by two different trios, and edited in such a way as to visually illustrate our experimental procedure, can be seen in the Supplementary Material.

Ethical approval for this study was obtained at [Anonymized] (Protocol ID: 2023-16). All methods were carried out in accordance with their guidelines and regulations.

2.3. Estimating the target sounds' perceptual emergence

Since the music was entirely improvised, the difficulty to detect a target sound was likely to vary depending on the moment at which it appeared, regarding not only the acoustical characteristics of the track in which it was embedded, but also that of the other tracks. We assumed that the latter aspect was minimal in the present context, since the tracks were separated in the mix and therefore only subject to contralateral masking effects. However, regarding the former aspect, the masking effects related to the track associated with the target sound could have influenced detection. Indeed, even though target sounds were always inserted in a track using a constant level offset of +1.5 dB, their perceptual emergence is likely to have depended on more complex acoustical parameters of that specific track (e.g., average level, spectral content of the track, etc.). Thus, to ensure that the observed performance to detect these target sounds would mainly reflect the degree of attention to the track with which it was associated, and not simply their perceptual emergence at a pure psychoacoustical level, we estimated such perceptual emergence based on a well-adopted computational model of the human auditory system (King et al., 2019). This model allowed us to compute a perceptual emergence score for each one of the target sounds triggered during the experiment (see Supplementary Material for more details), and thus to assess whether our experimental manipulation had an effect on the detection of target sounds, over and beyond their perceptual emergence.

2.4. Acoustical features

As the musicians were recorded in separate studio booths, we had access to each musician's individual tracks. This allowed us to calculate audio descriptors of interest and explore potential relationships between these descriptors and our salience manipulations. Specifically, we computed two main audio descriptors often used to account for interaction between the musicians in collective free improvisation: Root Mean Square (RMS) and spectral centroid (Goupil et al., 2021). RMS is indicative of the loudness in each musician's signal, while the spectral centroid provides information on the signal's timbre (more specifically on its brightness). These two audio descriptors were computed using the python library *Librosa* (McFee et al., 2015) from each musician's individual WAV files, with a 100ms window size.

2.5. Variables

The target sounds presented to the performers were divided into three categories:

- "louder" target sounds: the target sounds associated to a musician heard through a louder pattern;
- "softer" target sounds: the target sounds associated to a musician heard through a softer pattern;
- baseline target sounds: the target sounds associated to a musician heard without any RMS manipulation.

This provided us with our main independent variable: RMS manipulations (baseline/softer/louder).

To account for the effects of RMS manipulations on attention, a target sound was considered to be detected by a given musician if the musician's pedal was pressed in the 7 seconds following the triggering of the target sound (7 seconds was the minimal temporal distance be-

tween two successive target sounds).[1] This provided us with our first dependent variable: Detection (yes/no). Note that the false alarms (i.e., pressing the pedal in the absence of a target sound) were rare (mean number of false alarms per participant = 5.133, SD = 3.631), amounting to less than 3% of the total number of target sounds participants had to detect. Since there was moreover no clear method for assigning them to a given experimental category (i.e., baseline, "louder pattern", or "softer pattern"), they were ultimately not taken into account in our analysis.

In order to assess whether the potential effects of RMS manipulations on attention were due to a change in the local acoustic context, we were also interested in investigating whether target sounds that would occur immediately after a RMS manipulation (and thus, immediately after a change back to the baseline level) would be better detected that target sounds that would occur farther away from a RMS manipulation. Baseline target sounds were thus further divided into two sub-categories: Baseline 1 – containing the target sounds occurring less than 10 seconds after the end of a RMS manipulation (i.e., during a time window of the same length as our patterns); and Baseline 2 – containing the target sounds occurring more than 10 seconds after the end of a RMS manipulation.

We were finally interested in assessing whether RMS manipulations would impact the musicians' interactions. Since all of our target sounds were associated with one of our three experimental conditions, we relied on the timings of such target sounds to estimate the interaction between a given pair of musicians (the musician who had to detect the target sound and the musician with who the target sound was associated). For each musician, we thus extracted the time series corresponding to our two audio descriptors (RMS and spectral centroid) over a 7-second window (the minimum temporal distance between two target sounds, so that there would never be any overlap in successive measurements) centred around

[1] See Supplementary Material for a replication of our main analysis using a much shorter time window but yielding similar results.

the target sound under consideration. Note that 14 target sounds appeared less than 3.5s before the end of a performance and were thus discarded from our analyses as they did not allow us to extract complete time series following the procedure described above. To assess the interaction between any two improvisers, we then used cross-recurrence quantification analysis (CRQA) – a method that provides insights into various correlational characteristics between two or more time series and that has been widely used in joint action studies (Wallot & Leonardi, 2018).

After identifying the relevant hyperparameters (see Supplementary Material for more details), we performed the CRQA using the PyRQA library in Python (Rawald, Sips & Marwan, 2017) on the extracted pairs of RMS time series and pairs of Spectral Centroid time series. Three metrics were used as variables to describe the interaction between the musicians, namely:

- *Determinism* (DET), which corresponds to the proportion of recurrent points that forms diagonal lines in the cross-recurrence plot. Diagonal lines represent periods where the two time-series follow similar trajectories for an extended period, meaning that there are patterns in one time-series that are also found in the other time-series. This provides an estimation of the degree to which the two musicians had similar behaviors.
- *Laminarity* (LAM), which measures the proportion of recurrent points that form vertical lines in the recurrence plot. Vertical lines indicate periods of time where the system remains in the same state or a similar state for an extended period, reflecting phases of low variability. This provides an estimation of the degree of stability of the interaction between the two musicians.
- *Divergence* (DIV), which is calculated as *1/Lmax, Lmax* being the longest diagonal line. The higher this metric, the more chaotic is the relation between the two time series. This provides an estimation of the degree of independence of the two musicians

As mentioned above, the CRQA was performed for both RMS and spectral centroid. For each moment of RMS manipulation, and for each one of our three metrics (Determinism, Laminarity, Divergence), we thus obtained two values (one for RMS and one for spectral centroid). Those two values were then averaged, resulting in a total of three Interaction variables (Average Determinism, Average Laminarity, Average Divergence) which, taken together, provides a rich picture of the interaction between the musician who had to detect the target sound and the musician with who the target sound was associated (see Figure 2).

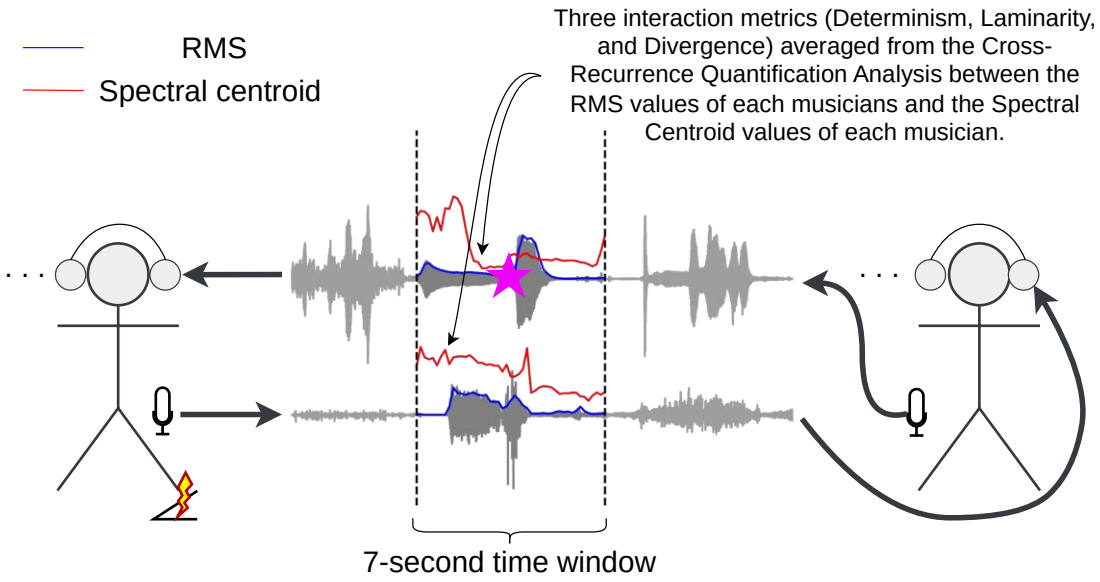

Figure 2. Computation of the three interaction variables.

2.6. Statistical analyses

As the duration of the improvisations were different each time, each participant ended up being presented with a different number of targets. To best account for this aspect and the fact that the overall detection performance greatly varied not only across individuals, but also within participants, depending on the musician (i.e., left or right) the target sounds were associated with, we relied on general linear mixed models for our statistical analyses, and

used the grouping between Detecting Participant (i.e., the participant that has to press the pedal) and Detected Participant (i.e., the participant with who the target sound is associated) as our random intercept.

To test the impact of our RMS manipulations on detection, we did a four-scale hierarchical regression by comparing nested models, starting with a null model (m0), then adding the emergence score as a predictor (m1), and finally adding the RMS category to which the target sound belongs (i.e., baseline, louder or softer) as a second predictor (m2). A last model was fitted which included the interaction term between our two predictors (m3). The models were fitted with the function *glmer* from the R package *lme4* (using a binomial family), and compared using a likelihood ratio test. "Baseline" was used as a base level whenever models included RMS manipulations as a predictor.

To test whether Baseline1-target sounds were better detected than Baseline2-target sounds, we did a two-scale hierarchical regression by comparing nested models starting with a model with only the emergence score as a predictor (m1), and then adding "Baseline type" (Baseline1 or Baseline2) as a second predictor. The models were fitted with the function *glmer* from the R package *lme4* (using a binomial family), and compared using a likelihood ratio test. "Baseline2" was used as a base level.

To test the impact of our RMS manipulation on the musicians' interactions, we did a series of two-scale hierarchical regressions by comparing nested models, with each one of our three Interaction variables as our dependent variables. We started with a null model (resp. m4, m6, and m8) and then added the RMS category (i.e., baseline, louder or softer) to which the interaction score under consideration was associated as our predictor (resp. m5, m7 and m9), using "Baseline" as the base level. The models were fitted with the function *lmer* from the R package lme4, and compared using a likelihood ratio test.

We also performed pairwise post-hoc tests using estimated marginal means, with Kenward-Roger degrees of freedom method and Benjamin-Hochberg method for p-values adjustment, using the *emmeans* package in R.

3. Results

3.1. RMS manipulations impact musicians' attention

The likelihood ratio test for model comparison between the null model (m0) and the model with perceptual emergence as a predictor (m1) was significant ($X2=103.93$, $p<0.001$). The likelihood ratio test for model comparison between m1 and the model with both perceptual emergence and RMS manipulations as predictors (m2) was also significant ($X2=8.343$, $p=0.015$), but adding the interaction term (m3) did not result in a better fit ($X2=1.964$, $p=0.375$). As expected, our model m2 showed a highly significant effect of perceptual emergence on detection (Estimate=2.496, SE=0.256, $z=9.739$ and $p<0.001$): target sounds which emerged more strongly from the musician's acoustic signal were better detected. But more importantly, the model also showed a significant effect of the "louder pattern" condition on detection (Estimate=0.343, SE=0.122, $z=2.813$, $p=0.005$) as well as a significant effect of the "softer pattern" condition on detection (Estimate=0.239, SE=0.119, $z=2.008$, $p=0.045$), as compared to the base level ("Baseline"). In other words, musicians were more likely to detect the target sounds when such sounds happened during a loudness manipulation, and such manipulations had an effect on our participants over and beyond the effects of the perceptual emergence of target sounds. Strikingly, a post-hoc pairwise test also showed that there was no significant difference in the participants' detection rate between the "louder pattern" and "softer pattern" conditions (Estimate=0.104, SE=0.119, $z=0.873$, $p=0.382$), suggesting that the effects of our manipulations on attention was not a matter of intrinsic acoustic features

(i.e., the louder, the more salient), but rather of dynamic contrast – in line with most current theories of salience (Kaya et al., 2020).

Note however that the effect size of the manipulations remained small (Cohen's d=0.097 for louder patterns and 0.114 for softer patterns). This might be due to the subtleness of our manipulations, which was precisely designed to be as undisruptive as possible.

3.2. The impact of RMS manipulations on musicians' attention is not driven by short-term contrasts

While we established that both louder and softer patterns had an effect on our participants' auditory attention, the reason why they did so remains unclear. In line with the many studies highlighting the link between surprise and attention (see Hortsmann, 2015, for a review), prevalent accounts of acoustic salience generally highlight the role of contrasts on both short-term and long-term contexts (Huang & Elhilali, 2017; Kaya et al., 2020), but it is not clear that they necessarily have the same weight, particularly in an interactional setting.

If the salience of our RMS manipulations was mainly driven by contrastive variations within the local sonic context, this would imply that the target sounds that occur immediately *after* a change back to the baseline (Baseline 1) should also be better detected than the target sounds that occur farther away from the end of the manipulation (Baseline 2), during periods of overall RMS stability. However, the likelihood ratio test for model comparison between m1 and our model with Baseline Type as an additional predictor was not significant ($X^2=0.130$, $p=0.718$), suggesting that whether a target sound immediately followed or not a loudness change did not make a difference in the participants' detection score.

It thus seems plausible that the effect of our manipulations on the participants' attention was driven by the more long-term context, although the precise temporal scale of such long-term context remains unknown. The change back to the baseline could be less salient than the de-

parture from the baseline simply in virtue of being the second loudness change within a 10-second interval, thus making it less surprising. Alternatively, it could be less salient because it simply signals a return to a "normal" state of affair. To further substantiate this latter interpretation, we ran two additional analyses. First, we computed the amount of time a musician was hearing a given co-improviser under each one of our 3 conditions (baseline, louder pattern, softer pattern). We found that, on average, a participant heard a given co-improviser in the baseline condition 78% of the time, 10.8% of the time in the louder pattern condition, and 11.2% of the time in the softer pattern. Second, we computed, for each participant, the mean RMS level they received from their two co-improvisers' across our 3 conditions, confirming that, on average, co-improvisers were indeed heard significantly louder in the "louder pattern" condition, and significantly softer in the "softer pattern" condition, as compared to the baseline (see Supplementary Material for more details). In other words, participants were exposed on a long-term basis to the typical sound of their co-improvisers as heard in the baseline condition, so that in particular, the RMS range perceived during the baseline condition was much more likely to occur than the RMS range perceived during the experimental manipulations. This makes it highly plausible that the improvisers formed internal models of the standard sound of their co-improvisers based on how they typically sounded during the baseline condition, and that the experimental manipulations disrupted the broader expectations associated with these models. The ensuing prediction error is likely to have resulted in a heightened attention to what the manipulated musician was playing (Den Ouden et al., 2012), and thus in a better detection score when a target sound occurred during that same period.

3.3. RMS manipulations impact musicians' interactions in a targeted manner

The likelihood ratio tests for model comparisons between the null models (m4, m6 and m8) and the models with RMS manipulation as a predictor (m5, m7 and m9) were significant (X2

= 13.747, p = 0.001; X2 = 23.326, p < 0.001; and X2 = 7.569, p = 0.022, respectively). Model m5 showed a significant effect of the "louder pattern" condition on the average determinism (Estimate=0.03, SE=0.014, t=2.172, p=0.03) as well as a significant effect of the "softer pattern" condition (Estimate=0.05, SE=0.013, t=3.70, p< 0.001), as compared to the base level ("Baseline"). Model m7 also showed significant effects of both the louder pattern (Estimate=0.049, SE=0.014, t=3.382, p<0.001) and softer pattern (Estimate=0.067, SE=0.014, t=4.725, p< 0.001) on average laminarity. Finally, m9 showed significant effects of both the louder pattern (Estimate=-0.03, SE= 0.012, t=-2.440, p= 0.015) and softer pattern (Estimate=-0.029, SE=0.012, t=-2.366, p= 0.018) on average divergence.

In other words, our results suggest that the detecting musician and the detected musician had more similar musical behaviors (i.e., similar recurring patterns were more likely to be found for the two musicians), had a more stable interaction (i.e., the two musicians were more likely to reach a state that persisted over time with little variation), and were less independent from one another (i.e., the two musicians' behaviors were more likely to be predictive of one another) during louder and softer patterns as compared to the baseline.

Once again, the effect sizes were small (Cohen's d=0.058 for louder patterns and 0.143 for softer patterns on average determinism; Cohen's d=0.124 for louder patterns and 0.213 for softer patterns on average laminarity; Cohen's d=0.106 for louder patterns and 0.106 for softer patterns on average divergence). The highly complex and ever-changing nature of the music produced in a CFI context means that results obtained through algorithmic tools primarily designed to extract acoustical features from much simpler musical signals (e.g., pop music) are bound to be very noisy. What is remarkable here is not so much the effect sizes themselves but rather that a significant effect on all three of our interaction metrics was found for both patterns, despite the noisy nature of the acoustical data, suggesting a clear and consistent effect of salient cues on the musicians' interactional dynamics.

4. Discussion

Our study – which combined real-time covert loudness manipulations and online probe detection – made it possible to systematically investigate for the first time the effects of acoustic salience on auditory attention during live and complex interactions, using musical improvisation as an experimental paradigm. We found, first, that musicians were more likely to pay attention to a given musician when this musician had been made sounding louder or softer; second, that such salience cue appeared to be primarily driven by the long-term acoustic context, as no significant salience-like effect was observed when loudness returned back (increasing or decreasing) to the baseline value after our manipulation; and third, that salient cues also had an effect on the ongoing musical interaction – with improvisers tending to interact more strongly and in a more stable way with a musician that had been made sounding louder or softer. Taken together, these results shed new light on the role of salience in social-attentional processes and extend previous experimental observations to a full-fledge, complex interactional setting.

Attention can take many forms and objects, and attentional processes in musical contexts are no exception: musicians' attention can be alternatively mostly focused on themselves or on the other performers, depending on one's own musical part – i.e., how demanding or virtuosic it is – or on coordination requirements – i.e., how the sonic and temporal unfolding of your own part is dependent of that of the other performers (Keller, 2001; Faraco et al., 2024; Abalde et al., 2024); it can also shift from deep absorption to something more akin to mind-wandering (Høffding, 2019). But – as hinted by the observed inter-participant variability in detection score (see Figure S3 in Supplementary Material) – improvisers' attentional profiles can also be widely different from one another, for reasons that might have to do with the in-

strument they play and the musical functions that are traditionally associated with such instrument (e.g., in a jazz context, whether one own's instrument is usually part of the frontline or of the rhythm session, see Monson, 1996), their musical background, and their own representations of the practice of improvisation itself. Our paradigm opens new avenues to address these questions empirically.

In contexts as indeterminate and open-ended as collective improvisations, it can become crucial for agents to come up with strategies that can capture their co-agents' attention as a way to signal a given intention (e.g., the intention to end the performance, see Goupil et al., 2021). Previous observational works on musical collective improvisation (Canonne & Garnier, 2012) have suggested that modifying one's own production in such a way as to make it more salient within the group precisely is one of these strategies. Our results provide considerable additional ground to this suggestion by demonstrating that the interactional impact of our loudness manipulations goes over and beyond simple behavioral adaptations such as the Lombard effect (which was indeed also present in our study, see Supplementary Material): on the one hand, improvisers' interactions were also impacted when someone in the group was made to be perceived as playing *softer* (whereas the Lombard effect is only observed when the ambient noise *increase*); and on the other hand, salient events also impacted the very dynamics of the interaction between improvisers, creating points of stability within the musical flux. In other words, we should not think of acoustic salience as only a set property of a given auditory stream; it is also something that can be actively manipulated by the agents themselves for communicative purposes. Acoustic salience thus acts as a kind of ostensive gesture which guide not only attention but also the various inferential processes that are bound to emerge in most forms of social communication (Frith, 2008).

The salience effect of our loudness manipulations can be further interpreted in the light of studies on attentional selection. Most studies on salience have considered the salience of local

events in the auditory signal in a stimulus-driven, purely bottom-up way, and thus have developed models that mostly approach events' salience with respect to the emergence of their instantaneous acoustical statistics in a short-term context (see Kaya & Elhilali, 2014) and more recently to their semantics (Kothinti & Elhilali, 2023). But it might also be relevant to frame acoustic salience within a more general predictive-coding/Bayesian theory (Friston et al., 2012), and approach it not through the event/context relationship as this is typically done but rather through the short-term/long-term statistical relationship. In that respect, it is interesting to mention three factors that have been found to lead to an enhancement of attentional selection beyond perceptual emergence (see Theeuwes et al., 2019, for a review): (i) history-driven selection – which relates to statistical learning, with stimuli that belong to high-probability categories leading to less attentional capture compared to those associated to low-probability categories; (ii) value-driven selection – where a reward associated to a stimulus can enhance its attentional capture; and (iii) goal-driven selection – where a change in stimulus feature that is useful for a particular task can enhance attentional capture. Beyond the role of the statistical structure of the stimuli in explaining the effect of loudness manipulations on attentional capture observed here, one might wonder whether a change in a specific musician's overall loudness carries an intrinsic musical value in this context, e.g., by providing opportunities for the group to renegotiate the interactional dynamics. Yet, since disentangling those various mechanisms was not the initial purpose of our research, new studies with specific experimental designs would be needed to further explore the different factors shaping attentional selection in collective musical improvisation, beyond the too simple bottom-up/top-down dichotomy (Awh et al., 2012). It would also be interesting to study whether other acoustic dimensions would play a comparable role in such a musical context, e.g., whether experimentally-induced variations in timbre or average pitch would lead to similar effects on the other musician's attention and interaction. We could hypothesize that as soon as these

changes are meaningful musically-speaking, that they would induce comparable effects. This remains to be tested empirically.

The preeminence of the long-term context over the short-term context observed in our study might also precisely be due to its interactional nature. The sonic information perceived by the participants was not treated as mere acoustical cues but also as full-fledged behavioral cues, ultimately providing them with a sense of their (musical) personalities. This certainly led the improvisers to pay particular attention to the acoustic cues that were highly informative about their co-performers. In that perspective, a sonic event that strongly stands out from the local acoustic context might be less relevant than a less spectacular sonic event that does not match with the expected sonic and musical personality of a given co-improviser – as typically inferred over a longer time scale. The complex relations between short-term and long-term contexts, as well as between acoustic cues and interactional cues, should provide an exciting topic for further experimental explorations.

It is also important to acknowledge that the effects of our loudness manipulations (on both attention and interaction) were quite small. Our main motivation for selecting such a subtle loudness variation as our experimental manipulation was to make the manipulation as unnoticeable as possible, so as to not artificially disrupt the interactions between the musicians. As such, it remains an open question whether our small effect size is due to a slightly too subtle manipulation, or whether more overt acoustic manipulations would have yield similarly small effects, because, in this kind of context, musicians' attention is merely driven by top-down factors (associated to specific interactional goals or idiosyncratic musical intentions) rather than salient cues. Further studies could shed light on this issue by manipulating top-down attention at the same time (e.g., by providing musicians with dynamical attentional scripts that they must follow as they improvise) to assess the relative importance of both factors. On the other hand, our diotic-based target detection paradigm could be directly transferred to further

address such questions in lab-based experiments with isolated participants, extending previous psychophysical paradigms to assess auditory attention (Huang & Elhilali, 2017). Participants would not be interacting with the music played in their headphones but rather performing an additional unrelated cognitively demanding task, while we would measure how their performance for detecting a target appearing in one ear or the other of a diotic auditory scene varies with both bottom-up and top-down factors. Results from such in-lab experiments could provide insightful benchmarks for interpreting the size of the effects observed in online experiments, where the active task of the listener is arguably the most demanding.

Another interesting issue is the extent to which our results could generalize to other forms of sonic interactions between humans. In particular, it is an open question whether loudness manipulations could play a similar role in impacting interactions in multi-speaker conversations in natural languages. CFI is mainly a hedonistic practice, devoid of any practical concerns, making it possible for the improvisers to simply follow whatever catch their attention, and to let the music emerge from the ongoing interactions. Moreover, the lack of a clear musical semantics also means that salience is bound to play an even greater role in making the communication between the musicians possible, as it is often the case in semantically undetermined contexts (Kecskes, 2013). In contrast, it might well be the case that the role of auditory salience is lessen in the regulation of verbal interactions in which participants have more practical concerns (which might strongly constrain how the interaction will unfold) and can all rely on clear semantic resources. Here, we believe that our experimental procedure could be, in a principled-way, easily and fruitfully transposed to the study of complex multi-speaker conversations, allowing us to contrast effectively the role of auditory salience in both verbal interactions and musical interactions. Further studies addressing the respective impact of short-term feature emergence vs. changes in long-term stimulus statistics will thus be key for

a better understanding of how meaningful dynamics in speech prosody are integrated and consequently shape conversational interactions.

Finally, although our study suggests a link between attention and interaction, its exact nature remains unclear. Two concurrent explanations could account for this connection. On the one hand, attention might be seen as a condition for interaction: paying more attention to a fellow agent might make it more likely to engage in an interaction with them. But on the other hand, it is also possible that attention (to someone) emerges from the interaction itself – considered as an autonomous system made of mutual couplings (De Jaegher et al., 2010): what is felt or noticed first is the interaction, which then increases the attention towards the other parts of the interactive network. This hints at a variety of communicational strategies that are mediated either through interaction or through attention, and which could be more sharply teased apart in further studies. Moreover, while our study shed light on the dyadic musical interactions between the manipulated musician and the musician that listened to the manipulated musician, further experimental protocols have to be set up in order to fully understand how these induced changes in salience on a single musician shape the interactional dynamics at the broader group level. In particular, one could assume that if musicians rely on a constant amount of interactional resources (Keller, 2001), the strengthening of the interaction with the salient musician would necessarily come at the cost of a weaker interaction with the non-salient musicians. Addressing such questions of course requires further conceptual elaboration as well as the development of novel experimental protocols that can also provide information about the overall amount of attention deployed by each musician, but doing so might produce key data to further develop theoretical models of musical interaction.

Systematically studying attention in the wild – and moving from individual lab booths to more complex interactional settings – is arguably one of the most pressing challenges for social cognition research. Music – and in particular CFI – might provide a relevant experimental

paradigm for such research program, allowing us to balance experimental control with ecological validity in a satisfying way (D'Ausilio et al., 2015). Our study illustrates once again the value of musical settings for investigating social cognition, by providing a first attempt at tracking auditory attention in real-time during complex interactions. While music is arguably a much more harmonious environment than the standard noisy cocktail party situation, there is no doubt that it has still a lot to reveal on the inner workings of auditory attentional processes in highly social contexts.

References

Abalde, S. F., Rigby, A., Keller, P. E., & Novembre, G. (2024). A framework for joint music making: behavioral findings, neural processes, and computational models. *Neuroscience & Biobehavioral Reviews*, 105816.

Arnal, L. H., Kleinschmidt, A., Spinelli, L., Giraud, A. L., & Mégevand, P. (2019). The rough sound of salience enhances aversion through neural synchronisation. *Nature communications*, *10*(1), 3671.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, 16(8), 437-443.

Bidelman, G. M., & Yoo, J. (2020). Musicians show improved speech segregation in competitive, multi-talker cocktail party scenarios. *Frontiers in Psychology*, *11*, 1927.

Blacking, J. (1973). *How musical is man?*. University of Washington Press.

Canonne, C., Garnier, N. (2012). Cognition and Segmentation in Collective Free Improvisation: An Exploratory Study. *Proceedings of the 12th International Conference on Music Perception and Cognition 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*, Thessaloniki, Greece.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*(5), 975-979.

Clarke, E. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press.

Dalton, P., & Lavie, N. (2004). Auditory attentional capture: Effects of singleton distractor sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 30(1), 180. https://doi.org/10.1037/0096-1523.30.1.180

D'Ausilio, A., Novembre, G., Fadiga, L., & Keller, P. E. (2015). What can music tell us about social interaction?. *Trends in cognitive sciences*, *19*(3), 111-114.

Dawson, J., & Foulsham, T. (2022). Your turn to speak? Audiovisual social attention in the lab and in the wild. *Visual Cognition*, *30*(1-2), 116-134.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition?. *Trends in cognitive sciences*, *14*(10), 441-447.

Den Ouden, H. E., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, *3*, 548.

Faraco, A., Schwarz, A., Vincent, C., Susini, P., Ponsot, E., & Canonne, C. (2024). Listening Behaviors and Musical Coordination in Collective Free Improvisation. *Music & Science*, *7*, 20592043241257023.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, 3, 151

Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1499), 2033-2039.

Goupil, L., Wolf, T., Saint- Germier, P., Aucouturier, J. J., & Canonne, C. (2021a). Emergent shared intentions support coordination during collective musical improvisations. *Cognitive Science*, *45*(1), e12932.

Høffding, S. (2019). *A phenomenology of musical absorption*. Springer.

Horstmann, G. (2015). The surprise–attention link: A review. *Annals of the New York Academy of Sciences*, *1339*(1), 106-115.

Huang, N., & Elhilali, M. (2017). Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, *141*(3), 2163-2176.

Humphreys, G. W., Yoon, E. Y., Kumar, S., Lestou, V., Kitadono, K., Roberts, K. L., & Riddoch, M. J. (2010). The interaction of attention and action: From seeing action to acting on perception. *British Journal of Psychology*, *101*(2), 185-206.

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in human neuroscience*, 8, 327.

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160101.

Kaya, E. M., Huang, N., & Elhilali, M. (2020). Pitch, timbre and intensity interdependently modulate neural responses to salient sounds. *Neuroscience*, *440*, 1-14.

Kecskes, I. (2013). Cross-cultural and intercultural pragmatics. Oxford University Press.

Keller, P. E. (2001). Attentional resource allocation in musical ensemble performance. *Psychology of Music*, *29*(1), 20-38.

King, A., Varnet, L., & Lorenzi, C. (2019). Accounting for masking of frequency modulation by amplitude modulation with the modulation filter-bank concept. *The Journal of the Acoustical Society of America*, 145(4), 2277-2293.

Koch, I., Lawo, V., Fels, J., & Vorländer, M. (2011). Switching in the cocktail party: exploring intentional control of auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1140.

Kothinti, S. R., & Elhilali, M. (2023). Are acoustics enough? Semantic effects on auditory salience in natural scenes. *Frontiers in Psychology*, *14*, 1276237.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, *19*(22), R1024-R1027.

McFee, B., Raffel, D., Dawen L., Ellis, D., McVicar, M., Battenberg, E., Nieto, O. (2015). "librosa: Audio and music signal analysis in python." *Proceedings of the 14th python in science conference*, pp. 18-25.

Miles, K., Weisser, A., Kallen, R. W., Varlet, M., Richardson, M. J., & Buchholz, J. M. (2023). Behavioral dynamics of conversation, (mis) communication and coordination in noisy environments. *Scientific Reports*, *13*(1), 20271.

Monson, I. (1996). *Saying something: Jazz improvisation and interaction*. University of Chicago Press.

Noy, L., Dekel, E., & Alon, U. (2011). The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proceedings of the National Academy of Sciences*, *108*(52), 20947-20952.

Rawald, T., Sips, M., Marwan, N. (2017). PyRQA. Conducting Recurrence Quantification Analysis on Very Long Time Series Efficiently. *Computers and Geosciences*, 104, pp. 101-108. doi: https://doi.org/10.1016/j.cageo.2016.11.016.

Roueff, O. (2006). L'invention d'une «scène» musicale, ou le travail du réseau: La programmation d'un club de musiques improvisées entre radicalisation et consécration (1991-2001). *Sociologie de l'Art*, (1), 43-76.

Ryan, F., Jiang, H., Shukla, A., Rehg, J. M., & Ithapu, V. K. (2023). Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14663-14674).

Saint-Germier, P., Canonne, C. (2020). Coordinating free improvisation: An integrative framework for the study of collective improvisation. *Musicae Scientiae,* 26(3), 455–475, https://doi.org/10.1177/1029864920976182.

Schultz, B. G., Brown, R. M., & Kotz, S. A. (2021). Dynamic acoustic salience evokes motor responses. *Cortex*, *134*, 320-332.

Seddon, F. A. (2005). Modes of communication during jazz improvisation. *British Journal of Music Education*, *22*(1), 47-61.

Steffens, J., Müller, F., Schulz, M., & Gibson, S. (2020). The effect of inattention and cognitive load on unpleasantness judgments of environmental sounds. *Applied Acoustics*, *164*, 107278.

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2022). Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of neural engineering*, *18*(6), 066054.

Theeuwes, J. (2019). Goal-driven, stimulus-driven, and history-driven selection. *Current opinion in psychology*, 29, 97-101.

Van der Heiden, R. M., Kenemans, J. L., Donker, S. F., & Janssen, C. P. (2022). The effect of cognitive load on auditory susceptibility during automated driving. *Human factors*, *64*(7), 1195-1209.

Wallot, S. & Leonardi, G. (2018). Analyzing Multivariate Dynamics Using Cross-Recurrence Quantification Analysis (CRQA), Diagonal-Cross-Recurrence Profiles (DCRP) and Multidimensional Recurrence Quantification Analysis (MdRQA) – A Tutorial in R. *Frontiers in Psychology*, 9(2232). DOI: doi: 10.3389/fpsyg.2018.02232.

Wolf, W., Launay, J., & Dunbar, R. I. (2016). Joint attention, shared goals, and social bonding. *British Journal of Psychology*, *107*(2), 322-337.

Wu, W. (2014). *Attention*. Routledge.

Supplementary information

1. Video examples

The two video examples provided show musicians from two different trios improvising during our experimental protocol. The videos are displayed in such a way as to illustrate our experimental procedure. Each line is meant to represent the subjective perspective of the musician displayed in the center. The musician shown on the left was the musician heard on the left channel of their headphones, and the musician shown on the right was the musician heard on the right channel of their headphones. RMS manipulations applied to these musicians (i.e., louder and softer patterns) are shown through changes in brightness (the video of a musician gets darker when this musician is heard through a softer pattern, and the video of a musician gets brighter when this musician is heard through a louder pattern). Finally, sudden flashes in the video happen at times a target sound associated with that musician was presented.
Those videos were filmed and edited by Elsa Laurent.

2. Additional information on the participants

We also tested our participants for hearing losses. Audiometric thresholds were measured between 0.25 and 8 kHz in all participants, using an Echoscan device. Importantly, we ensured that participants had clinically-normal thresholds (<15 dB HL) in the frequency range (from 0.25 to 2 kHz) of the probe used for the detection task. Most of them also had a normal hearing at high frequencies, except for two participants who had a moderate hearing loss: one at 4 kHz and one at 8 kHz. For both of them, the loss was predominant in the left ear (~ 40 dB HL) in regard of the right ear (25 dB HL). Both participants were aware of this. As their results were similar to the others, and because the target sounds were lower in frequency and never presented at threshold, we decided to keep them in the analysis.

3. Assessing the naturalness of the RMS manipulations.

This follow-up experiment was designed to estimate, in a post-hoc fashion, the degree of naturality of the RMS manipulations introduced during our recording session. In particular, we were wondering whether external listeners would be more likely to perceive the overall musical output as less natural when one of the musicians was subjected to a RMS manipulation, as compared to the same output but without the RMS manipulation.

3.1. Participants

27 participants took part in study 2 (mean age = 26 years, SD = 5, 14 women and 13 men). They all had practiced music for at least 5 years (mean number of years = 12.5 years, SD = 5.5). All participants gave their informed written consent and were paid at a standard rate.

3.2. Stimuli

18 audio excerpts were chosen randomly at moments where one musician's RMS level was modified (either with a high-level pattern or a low-level pattern) for another musician. Those excerpts were directly taken from the audio tracks heard by the improvisers in their headphones. Each excerpt was 15 seconds long, corresponding to the 10-second period of the RMS manipulations, with the addition of two 2.5-second buffer zones at the start and at the end of the excerpt. For each excerpt, we then created two alternative versions: one containing

the opposite pattern from the one originally applied (i.e., a louder pattern instead of a softer pattern); and one with no RMS manipulation whatsoever. This resulted in a total of 54 excerpts. Finally, 50-ms fade-ins/fade-outs were imposed at the beginning and the end of each excerpt to avoid clipping.

3.3. Procedure

The following cover story was provided to the participants: "We are testing an artificial intelligence designed to help sound engineers in mixing various instruments playing live together. For each rough recording given to the AI, the AI offers various mixes. In this study, you will have to compare, for several different excerpts, two of the mixes that were created by the AI. We ask you to tell us which version you think sounds the most natural". Our design thus followed a two alternative forced choice (2-AFC) procedure. For each trial, two versions of the same musical excerpt (e.g., the same music but once with a high-level pattern applied to one of the musicians of the trio, and the other time with a low-level pattern applied to the same musician) were successively presented to the participants (with a 2-second gap in between), and participants had to select the version that felt the most "natural" for them.

The participants heard the stimuli through headphones (Beyerdynamics DT 770 pro, 80 ohms) and answered via a custom-made interface designed with Max/MSP.

Over the course of the experiment, participants had to compare each one of the three versions of each musical excerpt to the other two (i.e., louder pattern vs softer pattern; louder pattern vs no-manipulation; softer pattern vs no-manipulation). This resulted in a total of 54 trials, which were divided in three successive blocks of 18 trials. The order of the stimuli was pseudo-randomized so that, over all trials and participants, each stimulus was presented an equal number of times as the first excerpt and as the second excerpt. Participants were invited to take a break in between blocks.

3.4. Statistics

A two-scale hierarchical regression was conducted by comparing nested models, starting with a null model and adding our experimental condition (i.e., whether the stimulus belong to the no-pattern, softer pattern or louder pattern category). The effect of our experimental condition on whether a stimulus was chosen to sound more natural than the comparison stimulus was tested through a mixed binomial regression model with Condition as predictor and participants' choice (natural: yes/no) to be predicted. "No-pattern" was used as our model base level. In the random structure we included the random slope for each listener. The models were fitted with the function *glmer* from the R package *lme4* and compared using a likelihood ratio test.

3.5. Results

Figure S1 shows the results of this follow-up experiment. The likelihood ratio test for model comparison was not significant ($X2=3.730$, $p=0.155$), suggesting that whether the RMS of a musician had been modified or not did not make a clear difference in third-party listeners' naturalness judgments. In particular, it does not seem that listeners were more likely to find the excerpt's sound more natural when we presented them with the musicians' signals as actually played by them – without any RMS manipulations. This shows that our RMS manipulations were indeed subtle enough to not be clearly noticeable by external listeners, suggesting that they could have passed as real musical intentions (i.e., playing a bit louder or a bit softer) during the improvisers' interactions.

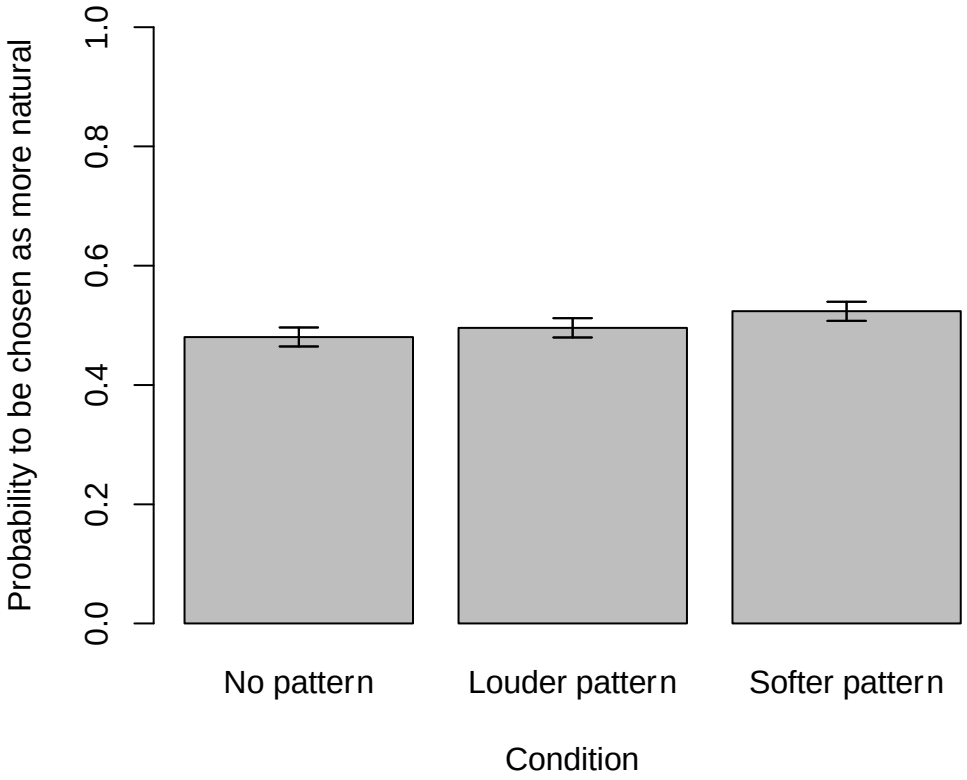

Figure S1. Bars represent the probability of a snippet from a certain Condition (No-pattern, Louder pattern, or Softer pattern) to be chosen as more natural than the comparison snippet. There were 36 datapoints for each participant and condition (36 * 3 = 108 datapoints per participant), which were averaged per participant (N = 27) and Condition before they were used in the plot. Error bars show standard error.

4. Analyses suggest that participants did not merely learn to associate RMS manipulations and target sound occurrences.

In order to decrease potential learning effects, target sounds and RMS manipulations were not systematically associated: on the one hand, roughly 31% of the target sounds occurred in the "baseline" condition, during which no RMS manipulation was applied; on the other hand, roughly 6.5% of the RMS manipulations were not associated to any target sound. Similarly, the distribution of the time intervals between two successive target sounds heard by a same participant was quite wide (see Figure S2 below), making it unlikely that the participants would press their pedals based on their learning a typical interval between two consecutive target sounds.

Post-hoc analyses of our results were performed to ensure that the effects of RMS manipulations on target sound detection was not the mere result of a learning effect. As our participants were asked to perform several improvisations during the experiment, we were able to compare detection scores for their first and last improvisations. Our reasoning was that if that there was any RMS-driven learning association issue, the effect of loudness patterns on target detection would be larger in the last improvisation (i.e., after learning) compared to the first improvisation. To do so, a hierarchical regression was conducted by comparing nested

models starting with a model with only the emergence score as a predictor (m1), second, adding "Order" (First improvisation or Last improvisation) as a predictor (m2), and third, the interaction term between Order and RMS manipulations (baseline, louder pattern, softer pattern) as a predictor (m3). The likelihood ratio test for model comparison between m1 and m2 was significant ($X2=13.763$, $p<0.001$). Our model m2 showed a highly significant effect of Order on detection (Estimate=0.059, SE=0.160, $z=3.692$ and $p<0.001$): target sounds were indeed better detected in the last improvisation than in the first performance. However, the likelihood ratio test for model comparison between m2 and m3 was not significant ($X2=8.553$, $p=0.073$), showing that participants were not significantly better to detect the target sounds associated with higher or louder patterns in the last improvisation than in the first improvisation. Hence, these analyses support the view that the effect of patterns on detection could not be simply explained by a RMS-driven learning association.

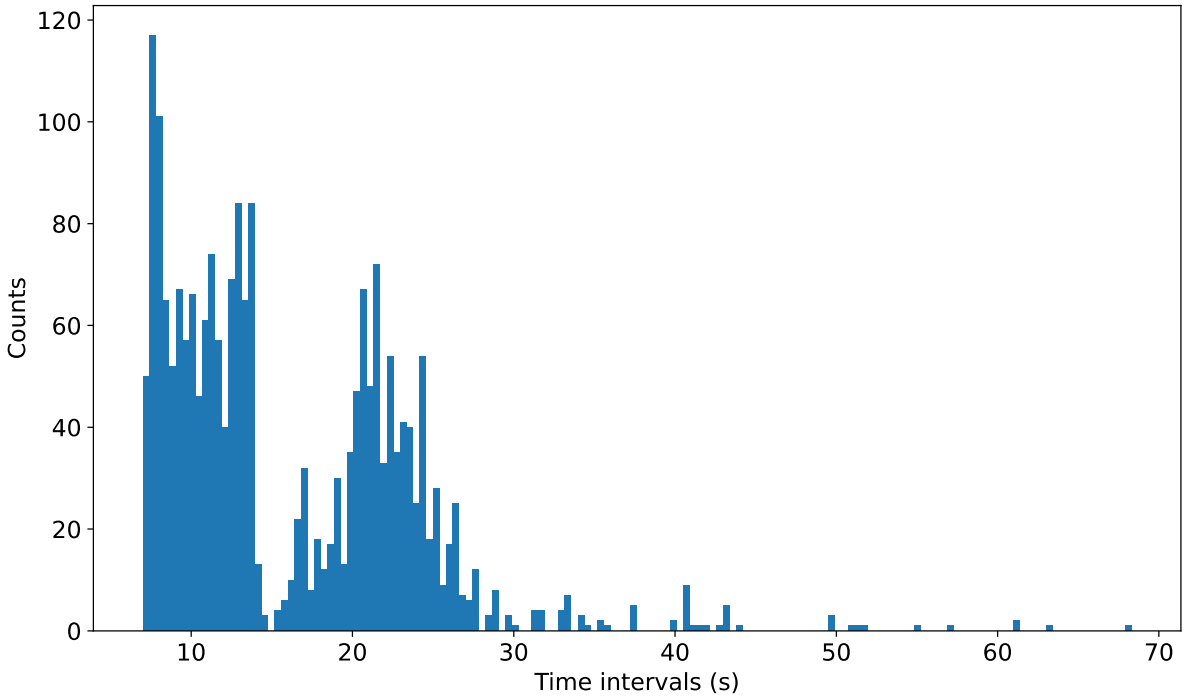

Figure S2. Distribution of the time intervals between two consecutive target sounds heard by a same participant

5. Computing the perceptual emergence of target sounds

Although target sounds were embedded in real time in the musical mix at a constant signal-to-noise ratio, this did not account for masking effects and therefore did not guarantee equal "audibility" of all target sounds. To account for such effects, we *a posteriori* estimated the "perceptual emergence" associated to each embedded target using the auditory modulation filterbank model (MFB), a computational model[2] of the human auditory system that mimics the main peripheral and central processing stages and can therefore account for basic perceptual masking effects and target-in-noise psychometric characteristics.

We used the MFB implementation of King et al. (2019) using the code provided by the *Auditory Modelling Toolbox* (Majdak et al., 2022; AMToolbox v. 1.5.0). A detailed description of

[2] The model is built from a cumulative body of works since the 90's (Dau et al., 1997) and is now used to account for a large range of human perceptual capacities ranging from complex targets detection in noise to speech recognition in noise (e.g., Relaño-Iborra et al., 2016).

the MFB model and its different processing stages can be found in Ponsot et al. (2021). The main parameters of the MFB model were chosen to best reflect auditory processing of normal-hearing listeners (frequency range of analysis was limited to that of the target, i.e., 100-2000Hz; N=10 cochlear filters with ERB=1; Q-value of modulation filters = 1; modulation phase cut-off at 5 Hz). Importantly, since we the overall sound pressure level of the mix received by the musicians in their headphones remained unknown (each musician could indeed freely adjust the overall level of the mix that they received), we used a linear version of the model with respect to sound level and entered a value of 1 for the compression parameter (i.e., no compression), meaning that the model did not attempt to account for level-dependence of masking and other more complex non-linear effects with respect to sound level.

We used the 3-D internal representation {frequency channels x modulation channels x time} of sounds provided at the output of the MFB model. We began by computing the internal model representation of the target sound alone (i.e., without any additional musical signal) to create a 'clean template'. Next, we extracted from each individual track 1-second temporal segments corresponding to the temporal windows of the overall mix at which target sounds were inserted, and computed their model representations, which yielded 'noisy templates'. The similarity between the clean and each noisy template (matrix were transformed into 1-D vectors) was then evaluated using Pearson correlation. This provided, for each extract, a scalar value of "perceptual emergence" of the target sound, which is assumed to reflect the amount of perceptual evidence that the target sound was present in the acoustical signal from the viewpoint of a human auditory observer. By controlling for the degree of "perceptual emergence" of each target sound on detection performance in the regression models, we aimed to more precisely assess the degree of attention to the track with which it was associated, so as to optimally probe the attentional dynamics of the musicians during the performance.

6. Identifying the hyperparameters for the cross-recurrence quantification analysis

CRQA has been described as "a class of multivariate and generalized correlational analyses, that are suited for joint action data because they make very few assumptions and are particularly robust in case of non-linearities, non-stationary dynamics and time-series with extreme outliers" (Wallot & Leonardi, 2018, p. 2). As such, CRQA has been widely used in joint action studies, such as ones involving synchronized arousal between performers in collective rituals (Konvalinka et al., 2011), conversations between medical doctors and patients (Angus et al, 2012), communicative interaction between children (Lira-Palma et al, 2018), and interaction between parents and children (Dale & Spivey, 2006; Reddy et al., 2013).

CRQA involves reconstructing two or multiple time series in a phase space, identifying recurrent states within the system. In other words, the analysis detects moments when the time series return to similar states, focusing on patterns of recurrence that can reveal the deterministic or stable nature of the dynamics between the two time-series. In order to perform the analysis, we followed the procedures described in Wallot & Leonardi (2018) and Wallot & Monster (2018). In order to do so, we needed to estimate the hyperparameters for the CRQA, namely the *embedding dimension*, the *time delay* and the *radius.* The time delay defines the number of time steps between consecutive points in the phase space reconstruction, while the embedding dimension refers to how many consecutive points of the time series are considered together when reconstructing the phase space. As for the radius, it sets the threshold for considering two points in the phase space as "close" or recurrent, determining the "sensitivity" of the recurrence. To achieve this, we first normalized our time series data (RMS and spectral centroid) by rescaling them. This step was important for two reasons: (1) as we used the average of the found optimal radii as a fixed parameter (see

below), both time series needed be on the same scale for the radius to be meaningful; and (2) since our interaction measurements are based on the average of the CRQA metrics, both time series needed to be standardized to the same scale for accurate comparison. Second, we looped over all pairs of segments (comprised of the detecting participant and the detected participant) for both RMS and spectral centroid and found the optimal time delay by using the average mutual information function, and the embedding dimension by using the false nearest-neighbors algorithm (both functions are taken from Wallot & Monster (2018) and were adapted from MATLAB to Python). Third, we used the average time delay and maximum embedding dimension to calculate the optimal radius based on an iterative process that finds the radius in which the percentage of recurrence (%REC) is between 1-10% (Wallot & Leonardi, 2018). Fourth, we averaged all hyperparameters thus identified.

7. Participants' detection scores and response times

Figure S3 shows, for each participant and each experimental condition, the number of target sounds they had to detect as well of the number of target sounds they actually detected. Figure S4 shows the distribution of the participants' response time for the detection of target sounds, revealing a fair amount of inter-participant variability, which is probably due to the large variety of instruments involved (e.g., it might have been easier to press a pedal for the saxophone players than for the drummers, who already have to use their feet to play their instrument). That being said, most pedal presses occurred quite shortly after the target sound occurrence (M = 2.224; SD = 0.541), suggesting that, when a target sound was detected, participants were rather fast at pressing their pedal, despite the very demanding setting they were placed in.

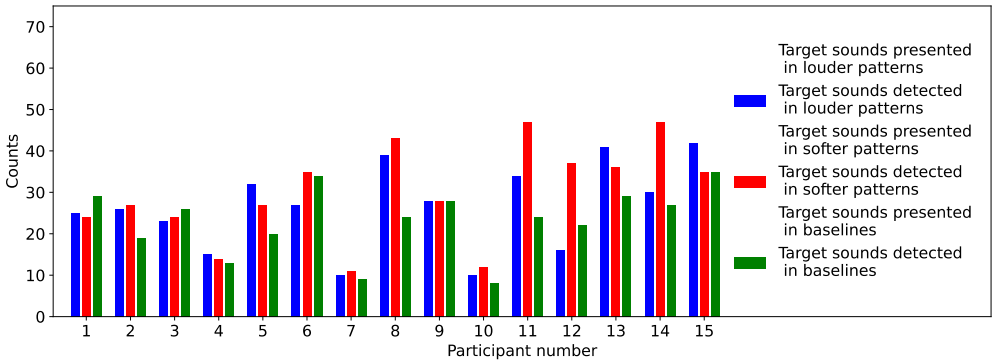

Figure S3. Target sounds presented and detected for each participant in each condition

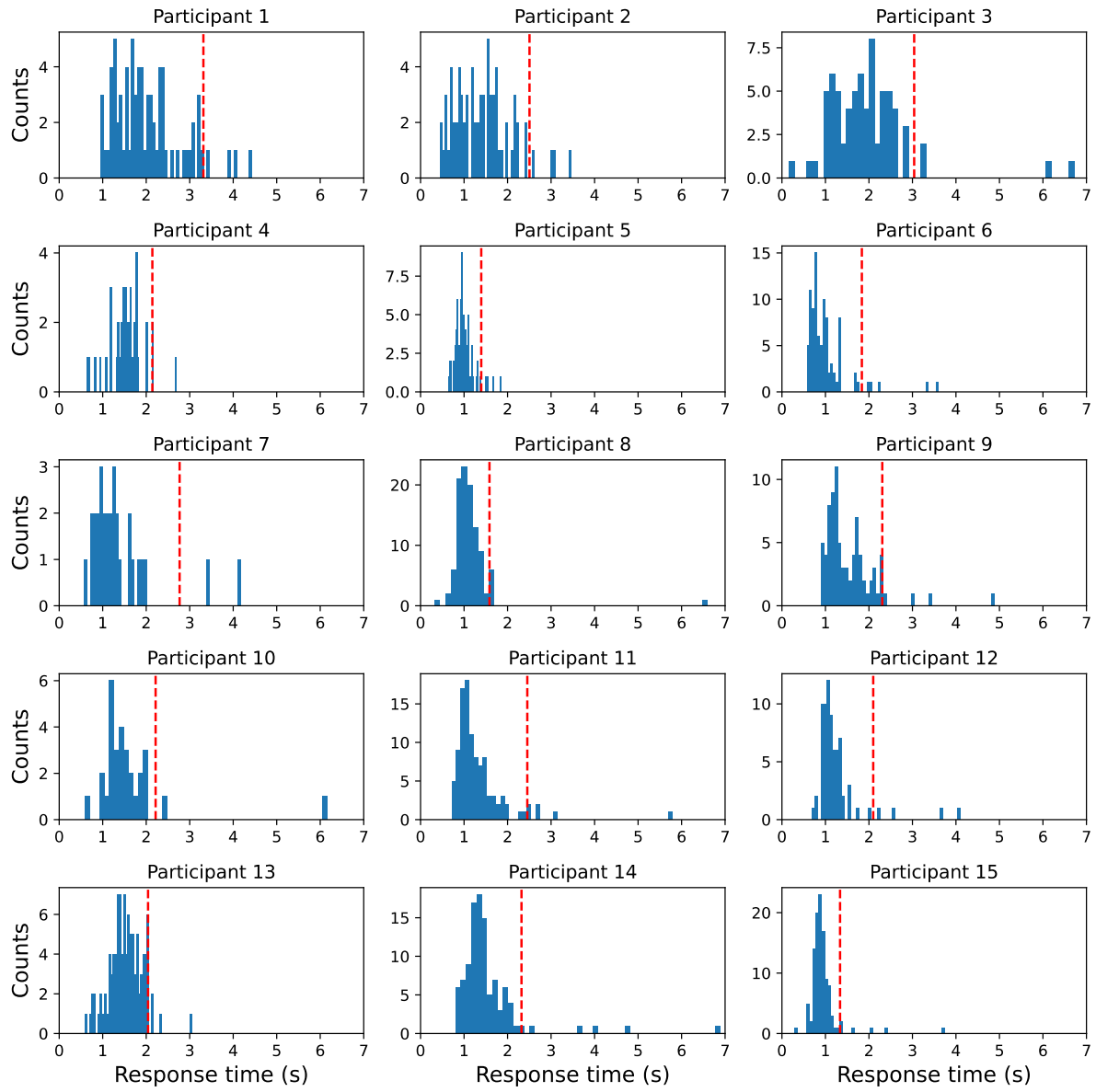

Figure S4. Occurrences of pedal presses as a function of response time for individual participants. The vertical dashed red line shows the cutoff for the 95% percentiles of these distributions.

8. Replicating our analyses using a much shorter time window for detection yields similar results

To ensure that the pattern of results reported in our main analysis did not depend on the time window we chose to compute the participants' detection score (a target sound was considered to be detected by a participant if the participant pressed their pedal in the 7 seconds following the target sound), we ran the same analyses using a much shorter time window, different for each participant, corresponding to the 95% percentile shown in Figure S4 (which was below 3 seconds for every participant). Strikingly, changing the size of the time window in such a way did not modify our pattern of results. The model still revealed a significant effect of the "louder pattern" condition (Estimate = 0.343, SE = 0.119, z = 2.872, p = 0.004) and a significant effect of the "softer pattern" condition (Estimate=0.258, SE=0.116, z=2.216, p=0.027).

9. Computing the RMS level received by each participant from their two co-improvisers across the three experimental conditions

For each pair of detecting participant/detected participant, we segmented the audio track of the detected participant in 10-s windows. Each window was ascribed to one of our 3 conditions (baseline, louder pattern, and softer pattern) depending on the manipulation (or absence of manipulation) applied to the detected participant during that window. The RMS level was then averaged over each window. On average, the mean RMS received by a given performer from a co-performer in the baseline condition was -31.493 dB; the mean RMS received by a given performer from a co-performer in the louder pattern condition was -23.097 dB; and the mean RMS received by a given performer from a co-performer in the softer pattern was -37.502 dB (see Figure S5).

We then conducted a statistical analysis. As the RMS data did not follow a normal distribution, we did a ranked transformation of our RMS data using the R library bestNormalize. This allowed us to run a two-scale hierarchical regression by comparing nested models, starting with a null model and then adding our experimental condition (i.e., whether the window belong to the baseline, softer pattern or louder pattern category) as a predictor. The likelihood ratio test for model comparison was significant ($X^2=537.22$, $p<0.001$). Our model with the experimental condition as a predictor showed that the RMS level received from musicians exposed to a louder pattern was significantly higher as compared to the baseline (Estimate = 0.688, SE=0.034, t= 20.087, $p<0.001$), and that the RMS level received from musicians exposed to a softer pattern was significantly lower as compared to the baseline (Estimate = -0.321, SE=0.033, t=-9.713, $p<0.001$).

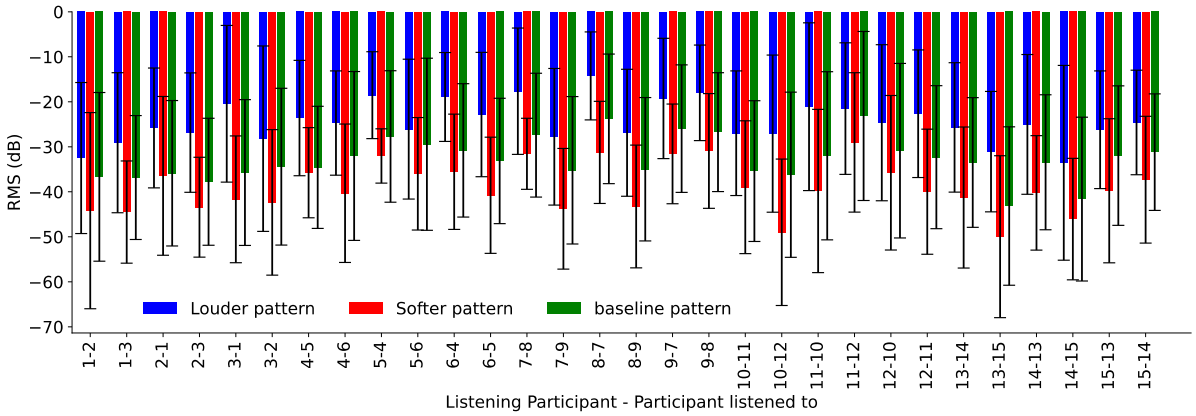

Figure S5. Mean RMS level received by each participant from their two co-improvisers across the three experimental conditions. Error bars show the standard deviation.

10. Improvisers were susceptible to a Lombard-like effect

We were interested to explore whether our participants would tend to play louder when they heard one of their co-improvisers under a louder pattern (and thus, heard this co-improviser playing louder than usual). This would be in line with the many studies on the so-called Lombard effect in the speech literature (Kunc et al., 2022), which showed that humans tend to speak louder as the ambient noise increase – a phenomenon that has also be identified in choral singing (Tonkinson, 1994).

First, we segmented the recorded audio track of each participant in 10-s windows. Each window was ascribed to one of our 3 conditions (baseline, louder pattern, and softer pattern) depending on the manipulation (or absence of manipulation) applied to their co-improvisers during that window. Second, we eliminated the windows in which the mean RMS was < -60 dB (meaning that the musician was simply not playing at that time). Averaging RMS values over each type of window revealed that musicians played +1.58 dB louder (SE = 0.43) when exposed to a louder pattern (as compared to the baseline) and +0.6 dB (SE = 0.41) louder when exposed to a softer pattern (as compared to the baseline).

We then conducted a statistical analysis. As the RMS data did not follow a normal distribution, we did a ranked transformation of our RMS data using the R library bestNormalize. This allowed us to run a two-scale hierarchical regression by comparing nested models, starting with a null model and adding our experimental condition (i.e., whether the window belong to the baseline, softer pattern or louder pattern category). The likelihood ratio test for model comparison was significant ($X^2=8.768$, $p=0.012$). Our model with the experimental condition as a predictor showed that musicians indeed tended to play significantly louder when exposed to a louder pattern (Estimate = 0.126, SE=0.023, $t=2.954$, $p=0.003$). Conversely, no significant effect was found for the softer pattern (Estimate = 0.048, SE=0.042, $t=1.141$, $p=0.254$).

Additional References

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualizing conversation structure across time: insights into effective doctor-patient consultations. *PLoS One*, 7:e38014. DOI: doi: 10.1371/journal.pone.0038014

Dale, R. &Spivey, M. J. (2006). Unraveling the dyad: using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Lang. Learn*. (56). DOI: doi: 10.1111/j.1467-9922.2006.00372.x.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of Am*erica, *102*(5), 2892-2905.

King, A., Varnet, L., & Lorenzi, C. (2019). Accounting for masking of frequency modulation by amplitude modulation with the modulation filter-bank concept. *The Journal of the Acoustical Society of America*, *145*(4), 2277-2293.

Kunc, H. P., Morrison, K., & Schmidt, R. (2022). A meta-analysis on the evolution of the Lombard effect reveals that amplitude adjustments are a widespread vertebrate mechanism. *Proceedings of the National Academy of Sciences*, *119*(30), e2117809119.

Konvalinka, I., Gálatas, D., Bulbulia, J., Schjodt, U., Jegindo, E., Wallot, S., Van Orden, G., & Roepstorff, A. (2011). Synchronized arousal between performers and related spectators in a fire-walking ritual. *Proceedings of the National Academy of Sciences*, 108(20). DOI: http://www.pnas.org/cgi/doi/10.1073/pnas.1016955108.

Lira-Palma, D., González-Rosales, K., Castillo, R. D., Spencer, R., & Fresno, A. (2018). Categorical Cross- Recurrence Quantification Analysis Applied to Communicative Interaction during Ainsworth's Strange Situation. *Complexity*, *2018*(1), 4547029.

Ponsot, E., Varnet, L., Wallaert, N., Daoud, E., Shamma, S. A., Lorenzi, C., & Neri, P. (2021). Mechanisms of spectrotemporal modulation detection for normal-and hearing-impaired listeners. *Trends in hearing*, *25*, 2331216520978029.

Reddy, V., Markova, G., & Wallot, S. (2013). Anticipatory adjustments to being picked up in infancy. *PLoS One*, 8:e65289. DOI: doi: 10.1371/journal.pone.0065289.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., & Dau, T. (2016). Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain. *The Journal of the Acoustical Society of America*, *140*(4), 2670-2679.
Tonkinson, S. (1994). The Lombard effect in choral singing. *Journal of Voice*, *8*(1), 24-29.
Wallot, S. & Leonardi, G. (2018). Analyzing Multivariate Dynamics Using Cross-Recurrence Quantification Analysis (CRQA), Diagonal-Cross-Recurrence Profiles (DCRP) and Multidimensional Recurrence Quantification Analysis (MdRQA) – A Tutorial in R. *Frontiers in Psychology*, 9(2232). DOI: doi: 10.3389/fpsyg.2018.02232.
Wallot, S., Monster, D. (2018). Calculation of Average Mutual Information (AMI) and False Nearest-Neighbors (FNN) for the Estimation of Embedding Parameters of Multidimensional Time Series in MATLAB. Frontiers in Psychology, 9. DOI: https://doi.org/10.3389/fpsyg.2018.01679.