Effects of self-motion on auditory scene analysis

Hirohito M. Kondo[a,1,2], Daniel Pressnitzer[b,c,1], Iwaki Toshima[a], and Makio Kashino[a,d]

[a]NTT Communication Science Laboratories, NTT Corporation, Atsugi, Kanagawa 243-0198, Japan; [b]Laboratoire Psychologie de la Perception, Centre National de la Recherche Scientifique and Université Paris Descartes, Paris F 75006, France; [c]Département d'études cognitives, Ecole normale supérieure, Paris F 75005, France; and [d]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Kanagawa 226-8503, Japan

Edited by Dale Purves, Duke University Medical Center, Durham, NC, and approved March 12, 2012 (received for review August 13, 2011)

Auditory scene analysis requires the listener to parse the incoming flow of acoustic information into perceptual "streams," such as sentences from a single talker in the midst of background noise. Behavioral and neural data show that the formation of streams is not instantaneous; rather, streaming builds up over time and can be reset by sudden changes in the acoustics of the scene. Here, we investigated the effect of changes induced by voluntary head motion on streaming. We used a telepresence robot in a virtual reality setup to disentangle all potential consequences of head motion: changes in acoustic cues at the ears, changes in apparent source location, and changes in motor or attentional processes. The results showed that self-motion influenced streaming in at least two ways. Right after the onset of movement, self-motion always induced some resetting of perceptual organization to one stream, even when the acoustic scene itself had not changed. Then, after the motion, the prevalent organization was rapidly biased by the binaural cues discovered through motion. Auditory scene analysis thus appears to be a dynamic process that is affected by the active sensing of the environment.

bistable perception | cocktail party problem | sensory motor | hearing

An essential function of perceptual systems is to structure the incoming flow of sensory information into coherent scenes that are able to guide behavior. For hearing, this task is termed "auditory scene analysis" (1). It is also known as the "cocktail party problem" (2), or how to follow a conversation that is acoustically intermingled with competing conversations, loud music, glasses tinkling, and so on. Many studies have been devoted to understanding the acoustic cues and, more recently, the neural mechanisms underlying auditory scene analysis (reviewed in 3–5). However, in all these studies, a very odd "cocktail party" is considered: it is a cocktail party in which the listener is unable to move his/her head. Therefore, very little is known about the interaction between sensory processes of auditory scene analysis and motor processes. Audio-motor processes related to head movements, which are known to help sound localization (6–8), should nevertheless be an important component of auditory scene analysis, both initially as a listener actively explores a novel scene, and in the ongoing maintenance of perceptual organization, because attention-grabbing sounds are likely to induce rapid head turns.

Here, we studied sensory-motor interactions in auditory scene analysis by means of a streaming paradigm (9). Streaming refers to the ability to group sequences of sounds into perceptual objects that extend over several seconds, such as musical melodies or speech sentences. A classic streaming stimulus is a sequence of two tones, A and B, alternating at different frequencies in a repeating ABA-ABA- pattern, where "-" denotes a brief silent interval. For intermediate frequency separations of A and B, listeners usually report initially hearing a single stream (ABA-ABA-). However, after a few seconds, perception changes and two concurrent streams are formed (A-A-A- and -B---B---). Perception then becomes bistable (10). The initial switch of one stream to two streams is called the build-up of streaming (11). Interestingly, whenever something changes in the ABA- sequence, streaming is reset: for instance, a silent gap or a sudden change in location brings the listener back to a one-stream percept and stream formation starts all over again (11–14). It is as if, when faced with a change in acoustic evidence, the auditory system reevaluates the ongoing perceptual organization.

When moving our head in a stationary auditory scene, we should be able to realize that the scene itself has not changed. In such cases, does self-motion induce a reset of streaming? Several hypotheses can be considered. On the one hand, self-motion changes the cues at the ears of the listener, because head orientation determines the acoustic path between sources and ears. Therefore, some resetting attributable to acoustic changes could be expected (11, 14). Engaging and disengaging spatial attention also affects streaming, and also induces some resetting (15, 16). Because initiating head motion presumably involves attention, some resetting could again be expected here. On the other hand, self-motion is under the volitional control of the listener, who should be able to determine that the acoustic changes at the ears, correlated with head motion, are not a result of a change in the external world. In a classic study, Wallach (8) showed that head position was indeed combined with auditory cues for estimating sound location. In his experiments, the location of a sound source was changed while listeners moved their head. With the appropriate constraints, this led to the illusion of a static source at a location compatible with the dynamics of both head position and binaural cues. The importance of head motion for sound localization has since been confirmed for horizontal accuracy but especially for front-back disambiguation (17). It would therefore seem reasonable that head motion signals may also affect auditory scene analysis and suppress any resetting during self-motion.

Disentangling acoustic and motor cues is difficult because, for real head movements, they are fully correlated. Here, we circumvented the problem by using a virtual reality setup. Listeners heard through the ears of a telepresence robot, the "Telehead" system (18). The experimental setup is illustrated in Fig. 1*A* and Movie S1. Listeners were seated in a sound-insulating booth. Their head motion was tracked in real-time and sent to the Telehead robot, which could mirror the 3D motion with minimal latency and distortion (18). Bistable streaming sequences (10) were played over a single loudspeaker placed in front of the Telehead for experiments 1 and S1, and over two loudspeakers placed in different locations for experiments 2a and 2b (using narrow-band noises instead of tones to facilitate localization). Sound was recorded by microphones inserted in the Telehead's ear canal and transmitted in real-time to the listener via headphones.

Author contributions: H.M.K., D.P., I.T., and M.K. designed research; H.M.K., D.P., and I.T. performed research; H.M.K., D.P., and I.T. analyzed data; and H.M.K. and D.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]H.M.K. and D.P. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: kondo.hirohito@lab.ntt.co.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1112852109/-/DCSupplemental.

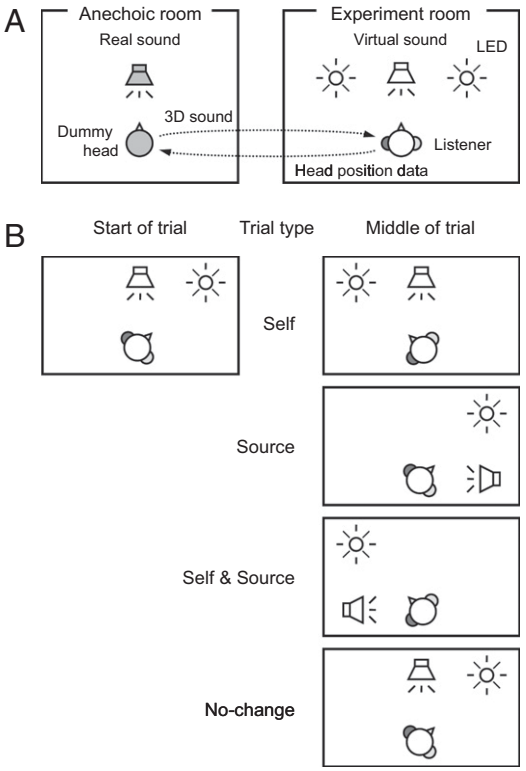


Fig. 1. Illustration of the experimental setup and trial types in experiment 1. (*A*) Auditory stimuli were presented to the Telehead robotic system. A loudspeaker was positioned in front of the robotic dummy head. Sounds were collected by microphones placed in the dummy head and transmitted in real-time to the listener via headphones. The head motion of the listener was tracked and could be mimicked with minimal latency by the robotic head. (*B*) Relative head and source positions at the start and end of each trial type. Details are provided in the main text.

Results

Experiment 1: Self-Motion Induces Resetting. Listeners always started a trial by fixating on a light-emitting diode (LED), which was lit, at random, to their right or to their left. They were instructed that the LED could change location during the trial, in which case they were to track this change by moving their head as fast as possible so as to maintain their gaze on the light. Our motivation for requesting rapid movement was twofold. First, we wanted to reduce experimental variability for head motion velocity. Second, we wished to contrast experimental conditions in which resetting occurred because of source motion to conditions in which the exact same acoustic cues were created by head motion. Because resetting by source motion occurs mostly for sudden changes in source location (14), we had to aim for rapid head movements.

Table 1. Structure of the different trial types

Trial type	ΔA	ΔS	ΔH
Self	1	0	1
Source	1	1	0
Self & Source	0	1	1
No-change	0	0	0
Control	1	0	1

For each trial type, the table indicates the presence (1) or absence (0) of changes in acoustic cues at the ears (ΔA), changes in apparent source location in allocentric coordinates (ΔS), and changes in head position during the trial (ΔH).

Four types of trials were included, as illustrated in Fig. 1*B* and Table 1. In the Self trials, after 10 s of sound presentation, the LED was turned off and another LED was lit on the contralateral side. The Telehead robot mimicked the head motion so that the Self trials simulated actual head motion. In the Source trials, the LED remained lit on the same side throughout the trial so that there was no head motion required from the listener. However, the Telehead robot initiated a motion previously recorded from the same listener. This motion, crucially, had the same acoustic cues at the ears as for the Self trials but without their motor, attentional, and volitional components. Such Source trials simulated the displacement of a sound source. In the Self & Source trials, listeners initiated a head motion to follow a change in the visual cue position but the robot did not move. Such trials have all the motor/attentional/volitional components of the Self trials but without any change in acoustic cues at the ears. They resulted in an apparent motion of the source in allocentric coordinates, which appeared to follow exactly the orientation of the head (as when one listens to music over headphones). In the No-change trials, the visual cue position was maintained throughout the trial and neither the listener nor the robot moved. Those No-change trials were used as a baseline. Finally, in a control experiment, the Telehead system was not used. Stimuli were delivered directly over a loudspeaker placed in front of the listener. The trial structure was otherwise identical to Self trials. These Control trials aimed at measuring the effect of self-motion without any potential artifact introduced by the Telehead system. Self, Source, Self & Source, and No-change trials were interleaved randomly within experimental blocks, whereas Control trials were run in separate blocks. In addition to tracking the visual cue, listeners were instructed to report continuously whether they heard one stream or two.

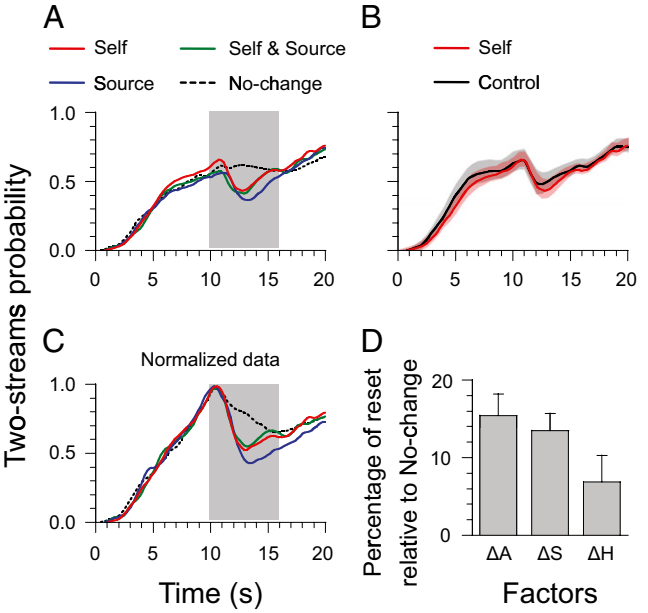


Fig. 2. Results for experiment 1. (*A*) Probability of two-stream responses averaged across listeners (N = 10) for the four trial types. (*B*) Same for Control trials, with the Self data replotted from *A*. Shaded areas indicate 95% confidence intervals. (*C*) Normalized data were computed by selecting the trials in which perception was two streams at the 10-s point. Resetting was evaluated over a 6-s time window (shaded area). (*D*) Estimated contributions to the resetting of (*i*) changes in acoustic cues at the ears, ΔA; (*ii*) apparent sound localization in allocentric coordinates, ΔS; or (*iii*) non-auditory factors related to head motion, ΔH. Those contributions were estimated for each listener by means of a linear additive model considering all trial types.

Analyses of head motion confirmed that listeners followed the instructions on all trials, with an average duration from the visual cue to the end of the head motion of 1.4 ± 0.07 s and a corresponding motion speed of 133° ± 14° per second (means ± SEs). The proportions of two-stream reports are shown in Fig. 2*A*. All results display a typical build-up pattern for the first 10 s of sound presentation: the initial report was always one stream, and the proportion of two streams gradually increased over time before reaching a plateau for the No-change trials. A repeated-measures ANOVA showed that the two streams' probabilities at 10 s after stimulus onset did not depend on trial type [0.61 ± 0.05, 0.53 ± 0.05, 0.57 ± 0.05, 0.58 ± 0.05, and 0.60 ± 0.06 in the Self, Source, Self & Source, No-change, and Control trials, respectively; $F_{(4, 36)} = 2.03$, $\eta^2 = 0.03$, $P = 0.11$]. This is consistent with listeners being unable to guess the type of trials before presentation of the visual cue. In addition, results for Self trials did not differ from those for Control trials (Fig. 2*B*, overlapping confidence intervals). Thus, it is unlikely that the results of the main experiment were contaminated by artifacts from the Telehead system. From now on, we will only discuss the Self, Source, Self & Source, and No-change trials.

Fig. 2*A* shows that the resetting of stream segregation was observed for all trial types except the No-change trials. This was quantified through several analyses. First, as resetting was the effect of interest, we selected only those trials for which listeners reported a two-stream percept at 10 s after stimulus onset (Fig. 2*C*). This normalization procedure aimed at increasing the power of the analysis because it removed the trials for which, by definition, resetting could not be measured. By construction, the normalized two-stream probability at 10 s was equal to one. Note that this maximum has to be followed by a dip in the two-stream probability: because of the stochastic nature of bistable streaming, listeners will eventually switch back to one stream even when there is no resetting (e.g., No-change condition). Therefore, we always estimated the true amount of resetting attributable to the experimental manipulations relative to the No-change trials. These served as a baseline, capturing the natural dynamics of switching without resetting for each listener. To estimate the amount of resetting, *R*, in the Self, Source, and Self & Source trials, we selected a time window (Fig. 2*C*, shaded area), integrated the proportion of two-stream judgments across trials, and subtracted the baseline proportion of two-stream judgments obtained in the No-change trials (Eq. **S1**). The time window was chosen from 10 to 16 s, thus starting at the onset of motion and ending at the local minimum of the two-stream probability for the No-change trials (15.8 ± 0.27 s, which reflects the dynamics of random switching without resetting). An ANOVA gave the following pattern for *R*: Source (0.29 ± 0.04) > Self (0.22 ± 0.06) and Self & Source (0.20 ± 0.05) [$F_{(2, 18)} = 7.01$, $\eta^2 = 0.19$, $P < 0.01$].

We further estimated the time at which the amount of resetting reached a maximum after the build-up (i.e., for each listener, we computed the time of the minimum in the two-stream proportion after the 10-s point). There were no significant differences between trial types: 12.9 ± 0.44 s, 13.4 ± 0.42 s, and 13.2 ± 0.54 s in the Self, Source, and Self & Source trials, respectively [$F_{(2,18)} = 0.43$, $\eta^2 = 0.00$, not significant (ns)]. In all cases, the maximum resetting was at about 13 s, which was well within the 6-s analysis window chosen to estimate *R*.

There is a possibility that the amount of resetting was affected by the speed of head movements, which varied from trial to trial. We analyzed the speeds of head movements for the Self trials. To maximize sensitivity, we focused on trials in which the resetting occurred from 10 through 13 s. We split the Self trials into those in which perception was reset to one stream (66 ± 5.5%) and those in which perception was not reset (34 ± 5.5%). Head-movement speed did not differ significantly between the two subsets: 136° ± 18° per second and 144° ± 21° per second [$t_{(9)} = 0.44$, $\eta^2 = 0.02$, ns].

Finally, we estimated the amount of resetting attributable to (*i*) changes in acoustic cues, (*ii*) changes in apparent source location in allocentric coordinates, and (*iii*) changes in nonauditory processes related to head motion. Usually, those factors cannot be distinguished: a change in apparent source location is accompanied by changes in acoustic cues at the ears; conversely, head motion will usually cause changes in acoustic cues. However, because of our design, we could enter these three factors in a linear model. The amount of resetting, *R*, was modeled as being the sum of resetting caused purely by each of the three factors, which intervened in some but not all trial types (Table 1). A system with three equations and three unknowns could then be solved for each listener (Eq. **S2**). The results are presented in Fig. 2*D*. Changes in cues at the ears and in allocentric source location obtained positive weights, indicating that those factors were able to cause some resetting. This is consistent with previous results (11, 13, 14). The factor related to head motion (when the accompanying acoustic and source location changes were factored out) obtained much smaller weights. An ANOVA confirmed that the contribution of acoustic cues, ΔA, and apparent location, ΔS, was greater than that of head motion factor, ΔH [$F_{(2, 18)} = 7.01$, $\eta^2 = 0.19$, $P < 0.01$]. Further *t* tests showed that the effects of ΔA and ΔS were significantly different from zero [$t_{(9)} = 5.15$ and 5.73, $\eta^2 = 0.75$ and 0.78, $P < 0.001$], whereas the effect of ΔH was not [$t_{(9)} = 1.96$, $\eta^2 = 0.30$, $P = 0.08$].

Experiment 2a: Self Motion with More than One Source. In experiment 1, all conditions produced some resetting relative to No-change. This could be because any sudden change, attributable to either head or source motion, triggered a reevaluation of perceptual organization. However, another explanation could be that all the changes we tested produced dynamic localization cues that favored the one-stream interpretation. As mentioned previously, dynamic localization cues participate in source localization (8, 17). Furthermore, because our stimuli were presented from a single loudspeaker, dynamic localization cues were always the same for the A and B noises. This would strongly favor the one-stream interpretation (19). The "resetting" we observed may thus simply reflect an increased likelihood of a one-stream interpretation.

We tested this hypothesis in experiment 2. Again, we used the Telehead system and included Self trials similar to those of experiment 1. Importantly, we also included trials in which the A and B noises were played from different loudspeakers positioned in distinct spatial locations (Fig. 3*A*). In Front-front trials, the two loudspeakers faced the robotic head but were located at a distance of 1 m and 2 m from the dummy head, respectively. In Front-back trials, one loudspeaker was located 1 m in front of the robotic head and the other loudspeaker was located 1 m behind the robotic head. All trials started with the listener's head facing the midline. This was intended to maintain ambiguity as to the presence of one or two sources at the onset of the trial, because distance and front-back cues are relatively poor for static sounds (17). Then, as in experiment 1, the LED location changed to one side during sound presentation and subjects were instructed to track the change with a rapid head movement. As soon as head movement was initiated, because of the spatial configuration of the loudspeakers, dynamic localization cues for the A and B noises were fully correlated (Self), partially correlated (Front-front), or anticorrelated (Front-back). The latter two trial types should favor a two-stream interpretation. As baselines, we also included three types of No-change trials, with spatial configurations matching those for head motion trials.

Results are shown in Fig. 3*B*. A repeated-measures ANOVA showed that the two-stream probability at 10 s after stimulus onset did not differ between the Self, Front-front, Front-back, and No-change trials: 0.66 ± 0.07, 0.65 ± 0.09, 0.70 ± 0.05, and 0.67 ± 0.06, respectively. This shows that, as we hypothesized, the static cues for distance and front-back location did not affect

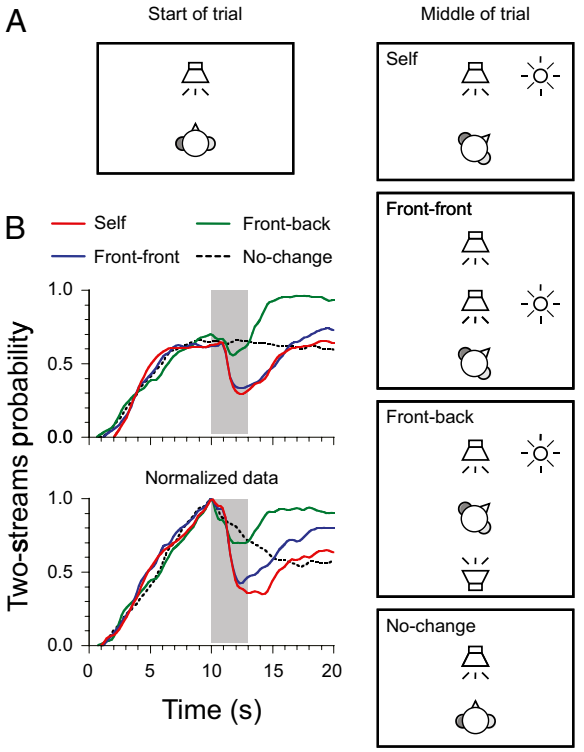


Fig. 3. Setup and results for experiment 2a. (*A*) Trial types differed according to the spatial location of sound sources. Because of the spatial arrangement of loudspeakers, for Self trials, dynamic localization cues during head motion were fully correlated (identical) for A and B noises, partially correlated for Front-front trials, and anticorrelated for Front-back trials. There were also three types of No-change trials, with spatial configurations corresponding to Self, Front-front, and Front-back trials. (*B*) Probability of two-stream response averaged across listeners (N = 8) for the four trial types (*Upper*) and the corresponding normalized data (*Lower*).

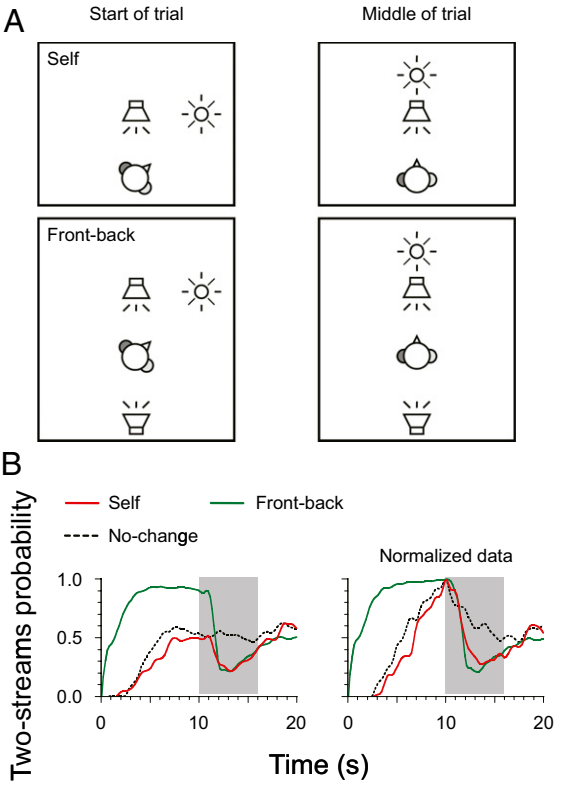


Fig. 4. Setup and results for experiment 2b. (*A*) Self and Front-back trials were as in experiment 2a, except that listeners started the trials by facing toward the side and subsequently moved toward the midline. The dynamic localization cues were as in experiment 2a, but the static binaural cues after head motion were now similar between trial types. (*B*) Probability of two-stream response averaged across listeners (N = 4) for the three trial types (*Left*) and the corresponding normalized data (*Right*).

the build-up of streaming. Using the normalized data (computed relative to each No-change baseline with the correct spatial configuration), we performed an ANOVA on the amount of resetting R for each condition in a window from 10 to 13 s. The main effect contrasting trial types was significant: Self (0.21 ± 0.06), Front-front (0.18 ± 0.07), and Front-back (0.08 ± 0.03) [$F_{(2,\ 14)} = 4.57$, $\eta^2 = 0.16$, $P = 0.03$]. There was a significant difference between Self and Front-back trials, with the latter exhibiting less resetting. However, crucially, statistical testing of R against zero showed that some resetting to one stream was observed in all conditions, even when dynamic localization cues favored a two-stream interpretation: Self [$t_{(7)} = 3.29$, $\eta^2 = 0.61$, $P = 0.01$], Front-front [$t_{(7)} = 2.64$, $\eta^2 = 0.50$, $P = 0.03$], and Front-back [$t_{(7)} = 2.61$, $\eta^2 = 0.49$, $P = 0.04$].

Another notable feature of the results is that the Front-back trials displayed a much higher probability for two streams after head motion compared with other trial types. We quantified the probability of two streams after the resetting period by averaging the raw data within a window from 13 to 20 s. A repeated-measures ANOVA confirmed that the probability for two streams was greater during the Front-back trials (0.92 ± 0.03) than during the Self trials (0.56 ± 0.09), Front-front trials (0.60 ± 0.07), and No-change trials (0.63 ± 0.06) [$F_{(3,\ 21)} = 17.67$, $\eta^2 = 0.61$, $P < 0.001$].

Experiment 2b: Dynamic or Static Localization Cues. Two factors may have contributed to the large bias toward two streams at the end of the Front-back trials of experiment 2a: dynamic localization cues during motion and static binaural cues subsequent to the motion. Indeed, after motion, there were differences in binaural cues between A and B noises for Front-back trials but not for Self trials (Fig. 3). In experiment 2b, we aimed at minimizing those static binaural cues by introducing a single change compared with experiment 2a. Here, listeners started the trial by looking toward the side and ended the trial by looking toward the midline (Fig. 4*A*). This is simply the opposite motion sequence compared with experiment 2a. As a result, dynamic localization cues were the same as for experiment 2a, but there were no differences in static binaural cues between Self and Front-back trials after motion.

Results are shown in Fig. 4*B*. A repeated-measures ANOVA showed that the two-stream probability at 10 s after stimulus onset was greater for the Front-back trials (0.91 ± 0.05) than for Self trials (0.50 ± 0.10) and No-change trials (0.54 ± 0.12) [$F_{(2,\ 6)} = 11.56$, $\eta^2 = 0.63$, $P < 0.01$]. This is consistent with the presence of static binaural cues only for Front-back trials before motion (in this experiment, No-change trials had the spatial configuration of Self trials; *Materials and Methods*). Right after the motion, we estimated the amount of resetting, R, derived from the normalized data in a 6-s time window. R was always significantly different from zero: Self [$t_{(3)} = 4.88$, $\eta^2 = 0.89$, $P = 0.02$] and Front-back [$t_{(3)} = 4.19$, $\eta^2 = 0.85$, $P = 0.03$]. Moreover, the amount of resetting did not differ whether dynamic localization cues favored one or two streams: Self (0.40 ± 0.08) and Front-back (0.41 ± 0.10) [$t_{(3)} = 0.09$, $\eta^2 = 0.00$, ns].

Finally, we quantified the probability of two streams after the resetting period by averaging the raw data in a window from 16 to 20 s. There were no differences between trial types: Self (0.53 ± 0.10), Front-back (0.48 ± 0.11), and No-change (0.57 ± 0.06)

[$F_{(2, 6)} = 0.58$, $\eta^2 = 0.03$, ns]. These values are similar to those of the Self trials of experiment 2a.

Discussion

Whereas most previous studies involving head motion measured the accuracy of localization, we focused on auditory scene analysis in a streaming paradigm in this study. Importantly, streaming is only partially determined by spatial cues (20, 21). Our results demonstrate that self-motion affects streaming in at least two important ways: a partial resetting of perceptual organization is observed just after motion, and streaming then benefits from the spatial cues revealed by motion. We now discuss these two effects in turn.

The data first confirmed that rapid source motion, which is accompanied by a sudden change of acoustic cues at the ears, causes resetting to the one-stream interpretation (14). Perhaps more surprisingly, resetting also occurred when the changes at the ears were attributable to voluntary self-motion. This held true even when the acoustic scene itself did not change and when the dynamic localization cues associated with head motion strongly favored a two-stream interpretation. A smaller resetting was found in one experimental condition (experiment 2a, Front-back), but this was likely attributable to high bias toward two streams after self-motion in this case. When the bias was controlled for, in experiment 2b, resetting was just as large when dynamic cues favored two streams. Several hypotheses can be considered to account for the resetting.

A first possibility is that the presence of a visual cue induced a lapse in attention of the listeners. When attention is diverted from the streaming sequence by an auditory (12, 16) or visual (22) task, less build-up is observed. We performed a supplemental experiment, experiment S1 (*SI Experiment S1* and Fig. S1) to test directly for the effect of the visual cue. We found its effect to be negligible. Attention could also have been diverted by motor processes associated with head motion. However, the linear model of Fig. 2*D* strongly suggests that motion, per se, did not contribute to resetting.

A second possibility is that head motion induced some perceptual changes in the auditory scene, which may have triggered a reevaluation of perceptual organization. A compression of auditory space has been reported just before the initiation of rapid head movements (23). The velocity of head movements in our study was, on average, smaller than in the study by Leung et al. (23); nevertheless, it may have caused some compression. However, in our data, we found no influence of the speed of head motion on the amount of resetting in the Self trials. If perceptual changes induced by rapid head motion had been a significant contributor to resetting, we should have observed more resetting with increasing speed.

A third possibility considers that head position and binaural cues must be pooled for the computation of allocentric source location (6, 7). In the context of scene analysis, the head position signals could be used to account for the changes in cues at the ears, and thus suppress any resetting potentially caused by such changes (e.g., changes in binaural cues, changes in monaural intensity). If this mechanism were imperfect or noisy, at least for rapid head movements, the system could not be sure that the acoustic scene had not changed during motion and a resetting of perceptual organization is a reasonable outcome. We suggest that our behavioral data can be best explained by this hypothesis.

Although this remains speculative, we note that the broadly distributed network currently thought to be involved in scene analysis seems consistent with such an interpretation. Neural correlates of streaming have been found at many stages of the auditory pathways: in primary and secondary auditory cortex (3); in higher cortical regions, such as the intraparietal sulcus (24); and also subcortically in the auditory thalamus (25), inferior colliculus (26), and even before binaural convergence in the cochlear nucleus (27). At least parts of such a network may not be fully modulated by head position signals during self-motion (6, 7, 28). This could explain why resetting was experienced whenever acoustic cues changed at the ears.

After the resetting period, self-motion affected the probability of segregating the scene into two streams. For experiments 1 and 2a, the main cue to segregation at the beginning of trials was the frequency difference between A and B noises. It produced bistable alternations between one and two streams, even though A and B appeared to share the same location (6, 20). In experiment 1, the coherent dynamic cues during motion confirmed to the listener that A and B came from the same location. After resetting, the average probability of two streams returned to that of the No-change trials. The situation was different for experiment 2a: For the Front-back trials in particular, what may have seemed to be a single location for A and B at the onset of the trial (because of front-back confusions) was revealed to be two locations during and after the head motion. In this case, a large bias toward segregation was induced by head motion. The bias was mostly attributable to the binaural cues revealed after the motion rather than to the dynamic cues during the motion. When dynamic cues were preserved but static binaural cues were minimized (experiment 2b), the segregation bias all but disappeared. The relative importance of static vs. dynamic cues observed here may be specific to the streaming paradigm, and does not necessarily reflect their respective contributions to sound localization (29, 30). In any case, for scene analysis, it seems that a major benefit of self-motion is to probe different sets of binaural cues, which, after a brief period of resetting, can be integrated into the estimate of perceptual organization. This is likely to provide a distinct advantage for the active sensing of a labile acoustic environment.

Materials and Methods

Listeners. Ten listeners were recruited for experiment 1 (5 male and 5 female; mean age = 25.3 y, range: 19–30 y). Eight listeners were recruited for experiment 2a (6 male and 2 female; mean age = 31.9 y, range 24–44 y). Four listeners were recruited for experiment 2b (3 male and 1 female; mean age = 32.8 y, range 27–40 y). A given listener only participated in a single experiment (experiment 1, 2a, or 2b). All gave written informed consent, which was approved by the Ethics Committee of Nippon Telegraph and Telephone (NTT) Communication Science Laboratories.

Apparatus. Listeners were seated in the center of a double-walled, soundproof room and wore headphones (HDA 200; Sennheiser). The Telehead system was used to present the sounds (18). For experiment 1 and control, acoustic stimuli were delivered through a loudspeaker (MG10SD0908; Vifa) located 1 m in front of the Telehead dummy head in an anechoic chamber. For experiments 2a and 2b, three identical loudspeakers were used, located at 1 m, 2 m in front of the dummy head, and 1 m behind the dummy head. Sound was recorded by small microphones (ECM77B; Sony) placed 2 mm inside the entrance of the dummy head's outer ears and transmitted in real-time to the headphones. Two LEDs were used as visual cues to direct head movements, one on the right and one on the left of the listener, at eye level and at a 2-m distance (visual angle with respect to midline = 60°). The room was darkened for the duration of the experiment. In the Control trials, the apparatus was identical, except that sounds were delivered directly by a loudspeaker placed inside the darkened room and the Telehead system was not used.

The Telehead dummy head was made by molding a human head using impression material. The surface was covered by soft polyurethane resin with a thickness of 1 cm. The head positions of listeners were measured by a 3D head tracker (FASTRAK; Polhemus) placed on the top of the headphones. The position data were obtained at a 120-Hz sampling rate and used to synchronize yaw, pitch, and roll motions (maximum range: 180°, 80°, and 60°, respectively) of the listener's head with those of the dummy head. The pitch and roll motions of the dummy head were controlled by two 400-W ac servomotors, whereas the yaw motion was regulated by a direct-drive brushless dc servomotor. To reduce mechanical noises, forces for the pitch and roll motions were transmitted via driving rods and belts. The low feedback gain of the servomotor system contributed to the reduction of

mechanical noises. The delay of mechanical responses was 20 ms, and line noise was not more than 24 dB sound pressure level (SPL). The experimental setup led to accurate sound localization (18). We interviewed the listeners after practice trials and confirmed that sounds were externalized (not heard inside their head).

Stimuli and Task Procedures. Auditory stimuli were made from 50 repetitions of a triplet of narrow-band pink noises (roll off = 3 dB per octave), arranged in an ABA- pattern, where A and B represent different noise bands and - represents a silent interval. The A and B bands were centered on 1 kHz with a six-semitone frequency difference between them and a four-semitone bandwidth. This yielded cutoff frequencies of [749–944] Hz for the A band and [1,060–1,335] Hz for the B band. The noise bands were generated in the frequency domain and equated in root mean square (rms) amplitude. The duration of each noise was 62.5 ms, which included rising and falling cosine ramps of 10 ms. Onset asynchrony between successive bands was 100 ms. A background of pink noise was also included to mask any residual line noise of the Telehead system. The pink noise was generated in the frequency domain with cutoff frequencies of [0.1–5] kHz, with a level of −30 dB RMS relative to the A and B bands. The sound pressure level was measured by using International Electrotechnical Commission (IEC) couplers with microphones and a measuring amplifier (Brüel & Kjær). The presentation level of the stimuli was set at 65 dB SPL.

Experiment 1 and control. Listeners were tested individually. We first explained the concept of auditory streaming by means of a visual illustration of the stimuli. Listeners were then instructed to judge whether they perceived one stream (ABA-ABA-...) with a galloping rhythm or two streams (A-A-... and -B---B---) with an isochronous rhythm for each stream. They reported their percept by means of a computer keyboard, without looking down at the keyboard. Listeners were also instructed to move their head to track an LED light and maintain it at the center of gaze. Before the beginning of each trial, they oriented toward the midline and the position of their head was calibrated. Then, a blinking LED was presented either to their left or right side in a random and counterbalanced fashion across trials. Listeners had to orient their head to the light. The LED stopped blinking, and listeners were instructed to maintain it at the center of gaze. In half of the trials, the LED was turned off and the contralateral LED was lit in the middle of the trial (10 s from onset). If that happened, listeners were asked to move their head as quickly as possible to follow the LED. At least 12 practice trials were run before data collection began. The head movement of listeners in the final practice trial was recorded to generate the Telehead motion in the Source trials. The main experiment consisted of six blocks of 24 trials. The order of the trial types was randomized, and 36 repeats were collected for each trial type. The control experiment was run with three blocks of 12 Control trials that were interleaved with the experimental blocks. The entire experimental session lasted approximately 2 h.

Experiment 2a. The procedure was the same as for experiment 1, except that the blinking LED, signaling the beginning of a trial, was located at the midline. In Self, Front-front, and Front-back trials, the midline LED was turned off at 10 s and another LED was lit to the right or left of the midline in a random and counterbalanced fashion. There were three types of No-change trials, with spatial configurations of A and B noises corresponding to the Self, Front-front, and Front-back cases.

Experiment 2b. The procedure was the same as for experiment 2a, except that the blinking LED, signaling the beginning of a trial, was located at the right or left of the midline, in a random and counterbalanced fashion. In Self and Front-back trials, the LED on the side was turned off at 10 s and another LED was lit at the midline. No-change trials had the same configuration as Self trials, but the LED remained lit on the same side throughout the trial.

ACKNOWLEDGMENTS. We thank Brian C. J. Moore and an anonymous reviewer for their insightful comments, which led to the design of experiments 2a, 2b, and S1. The work was partially supported by an NTT grant to DP and by the Agence Nationale de la Recherche.

1. Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
2. Cherry EC (1953) Some experiments on the recognition of speech, with one and two ears. *J Acoust Soc Am* 25:975–979.
3. Micheyl C, et al. (2007) The role of auditory cortex in the formation of auditory streams. *Hear Res* 229(1-2):116–131.
4. Shamma SA, Micheyl C (2010) Behind the scenes of auditory perception. *Curr Opin Neurobiol* 20:361–366.
5. Snyder JS, Alain C (2007) Toward a neurophysiological theory of auditory stream segregation. *Psychol Bull* 133:780–799.
6. Goossens HHLM, van Opstal AJ (1999) Influence of head position on the spatial representation of acoustic targets. *J Neurophysiol* 81:2720–2736.
7. Vliegen J, Van Grootel TJ, Van Opstal AJ (2004) Dynamic sound localization during rapid eye-head gaze shifts. *J Neurosci* 24:9291–9302.
8. Wallach H (1940) The role of head movements and vestibular and visual cues in sound localization. *J Exp Psychol* 27:339–368.
9. van Noorden LPAS (1975) Temporal coherence in the perception of tone sequences. PhD thesis (Eindhoven University of Technology, Eindhoven, The Netherlands).
10. Pressnitzer D, Hupé JM (2006) Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr Biol* 16:1351–1357.
11. Anstis S, Saida S (1985) Adaptation to auditory streaming of frequency-modulated tones. *J Exp Psychol Hum Percept Perform* 11:257–271.
12. Cusack R, Deeks J, Aikman G, Carlyon RP (2004) Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J Exp Psychol Hum Percept Perform* 30:643–656.
13. Roberts B, Glasberg BR, Moore BCJ (2008) Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. *J Exp Psychol Hum Percept Perform* 34:992–1006.
14. Rogers WL, Bregman AS (1998) Cumulation of the tendency to segregate auditory streams: Resetting by changes in location and loudness. *Percept Psychophys* 60: 1216–1227.
15. Best V, Ozmeral EJ, Kopco N, Shinn-Cunningham BG (2008) Object continuity enhances selective auditory attention. *Proc Natl Acad Sci USA* 105:13174–13178.
16. Thompson SK, Carlyon RP, Cusack R (2011) An objective measurement of the build-up of auditory streaming and of its modulation by attention. *J Exp Psychol Hum Percept Perform* 37:1253–1262.
17. Perrett S, Noble W (1997) The contribution of head motion cues to localization of low-pass noise. *Percept Psychophys* 59:1018–1026.
18. Toshima I, Aoki S, Hirahara T (2008) Sound localization using an auditory telepresence robot: TeleHead II. *Presence-Teleoperators and Virtual Environments* 17:392–404.
19. Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* 34(3):114–123.
20. Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. *Acta Acustica United with Acustica* 88:320–332.
21. Best V, Gallun FJ, Carlile S, Shinn-Cunningham BG (2007) Binaural interference and auditory grouping. *J Acoust Soc Am* 121:1070–1076.
22. Carlyon RP, Plack CJ, Fantini DA, Cusack R (2003) Cross-modal and non-sensory influences on auditory streaming. *Perception* 32:1393–1402.
23. Leung J, Alais D, Carlile S (2008) Compression of auditory space during rapid head turns. *Proc Natl Acad Sci USA* 105:6492–6497.
24. Cusack R (2005) The intraparietal sulcus and perceptual organization. *J Cogn Neurosci* 17:641–651.
25. Kondo HM, Kashino M (2009) Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J Neurosci* 29:12695–12701.
26. Schadwinkel S, Gutschalk A (2011) Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus. *J Neurophysiol* 105:1977–1983.
27. Pressnitzer D, Sayles M, Micheyl C, Winter IM (2008) Perceptual organization of sound begins in the auditory periphery. *Curr Biol* 18:1124–1128.
28. Altmann CF, Wilczek E, Kaiser J (2009) Processing of auditory location changes after horizontal head rotation. *J Neurosci* 29:13074–13078.
29. Middlebrooks JC, Green DM (1991) Sound localization by human listeners. *Annu Rev Psychol* 42:135–159.
30. Griffiths TD, et al. (1996) Evidence for a sound movement area in the human cerebral cortex. *Nature* 383:425–427.

Supporting Information

Kondo et al. 10.1073/pnas.1112852109

SI Materials and Methods

Perceptual Data Analyses. *Experiment 1 and control.* We recorded the time-series data of perceptual states. After performing an initial ANOVA on the raw data, normalized data were computed to estimate the amount of resetting for each listener and trial type. Only those trials in which listeners reported a two-stream percept at 10 s were retained (61% ± 5%, 53% ± 5%, 57% ± 5%, and 58% ± 5% for the Self, Source, Self & Source, and No-change trials, respectively). The results were then averaged between 10 and 16 s, the range at which a local minimum of the two-stream probability for the No-change trials was observed (15.8 ± 0.27 s). We used the No-change condition as a baseline. The amount of resetting, R, was computed for each trial type, TT, and listener, L, as follows:

$$R_{TT,L} = \int_{t=10}^{t=16} \left[P_{TT,L}(2\,stream) - P_{\text{No-change},L}(2\,stream)\right]\, dt. \qquad \text{[S1]}$$

We also built a linear model to estimate the contribution of (*i*) changes in acoustic cues at the ears (ΔA), (*ii*) changes in the source apparent location in allocentric coordinates (ΔS), and (*iii*) changes in head position (ΔH). The amount of resetting, R, was modeled as follows:

$$R = K_A\Delta\mathrm{A} + K_S\Delta\mathrm{S} + K_H\Delta\mathrm{H}. \qquad \text{[S2]}$$

For each listener, three measures of R were available (Self, Source, and Self & Source trials). For each measure, the values of ΔA, ΔS, and ΔH were set to either zero or one depending on whether the trial type included changes in the corresponding factor (Table 1). The system of three equations and three unknowns was then solved for each listener. ANOVAs were performed on both the R values and the K values. Tukey honestly significant difference tests were used for post hoc comparisons (α-level = 0.05).

Experiment 2a. The same method was used to estimate the amount of resetting, except that each condition (Self, Front-front, and Front-back) was compared with the appropriate *No-change* baseline (with the same spatial configuration). A 3-s time window was selected to estimate R because the crossing point between the Front-back and No-change time-series data was 13.0 ± 0.29 s.

Experiment 2b. The same method was used to estimate the amount of resetting. A 6-s time window was selected to compare R between experiments 1 and 2b. The crossing point between the Front-back and No-change time-series data was 16.9 ± 0.80 s.

Head-Motion Analyses. The head position was recorded for all trials in which the Telehead system was used (all except Control). For trials with head movements, we measured the maximum head angle (H_{max}) and corresponding time of occurrence (t_{max}) between 10 s (the time of the visual cue to move the head) and the end of the trial. The onset angle of head movement (H_{onset}) and its corresponding time were defined as 10% of H_{max}. The head movement speed (S_{hm}) was estimated as follows:

$$S_{hm} = (H_{max} - H_{onset})/(t_{max} - t_{onset}). \qquad \text{[S3]}$$

For experiment 1, H_{max} was 80° ± 4° on average. t_{onset} was 600 ± 20 ms, and the duration of head movement ($t_{max} - t_{onset}$) was 760 ± 60 ms. Thus, it took less than 1.5 s for listeners to complete their head motion after the visual cue. The corresponding S_{hm} was 133° ± 14°.

Experiment S1

Experiment S1 was designed to control for the effect of the visual cue used to initiate head movement. In experiment 1, all trials that included head movements (Self and Self & Source) also included a change in the location of the visual cue, because the LED indicating target gaze direction changed sides in the middle of the trial. In contrast, the location of the visual cue did not change for trials without head movement (Source and No-change baseline). Thus, there is a possibility that a change in visual cue influenced the amount of resetting differently for head motion trials and no-motion trials. For instance, the change in visual cue could have temporarily distracted listeners and contributed to the resetting observed in trials with head movement. Alternatively, listeners could have been surprised by the sudden onset of source motion in the Source trials, and this may add some resetting to those trials. We tested for those possibilities in experiment S1.

Materials and Methods. Five listeners were recruited (three male and two female; mean age = 32.2 y, range: 23–44 y) who had not participated in the main experiments. All were right-handed with normal hearing and normal vision.

The apparatus and stimuli were the same as those in the main experiments. Listeners were instructed to listen to stimulus sounds of a repeating ABA-ABA- pattern and report by a button press whether they perceived one stream (ABA-ABA-. . .) or two streams (A-A-. . . and -B---B---. . .).

Experiment S1 included Self, Source, and No-change trials, identical to the main experiment 1. It also included an additional Cued-source condition (Fig. S1*A*). For Cued-source, the visual cue blinked in the middle of the trial (9.5–10.5 s) but did not change position. Accordingly, no head motion was required from listeners, but a distracting visual cue was presented (alternately, the cue may have warned listeners of the impending change in source location).

A least 12 practice trials were run before data collection began. The order of the trial types was randomized, and 36 repeats were collected for each trial type: Self, Source, Cued-source, and No-change.

Results. The proportions of two streams are shown in Fig. S1*B* for each trial type. A repeated-measures ANOVA showed that the two-stream probability at 10 s after stimulus onset did not differ between trial types: means ± SEs were 0.50 ± 0.07, 0.48 ± 0.07, 0.48 ± 0.07, and 0.49 ± 0.05 in the Self, Source, Cued-source and No-change trials, respectively. To compute the amount of resetting for each trial type, we used normalized data by selecting only trials for which listeners reported a two-stream percept at 10 s after stimulus onset (see main text). When the No-change trials were used as a baseline, the amount of resetting, R, was 0.16 ± 0.06, 0.17 ± 0.08, and 0.14 ± 0.06 in the Self, Source, and Cued-source trials, respectively. An ANOVA revealed that there was no significant main effect [$F_{(2,\ 8)} = 0.18$, $\eta^2 = 0.00$, ns]. In particular, the lack of difference between Source and Cued-source trials suggests that the amount of resetting was not influenced by the change in visual cue in the middle of the trial. Thus, it is unlikely that the results of the main experiments were influenced by either lapses of attention or surprise caused by the visual cue.

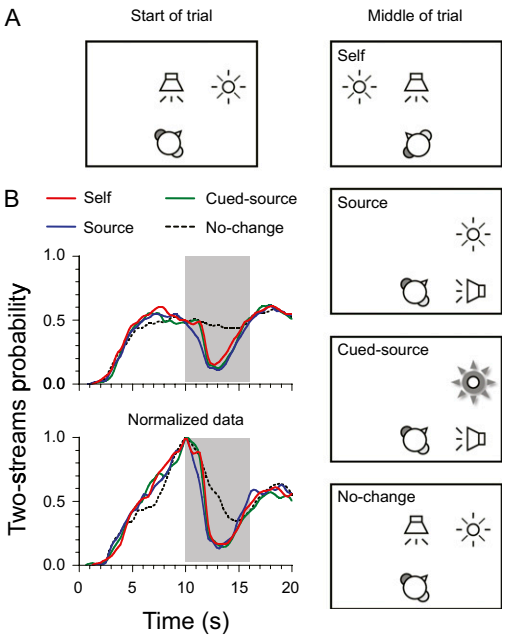
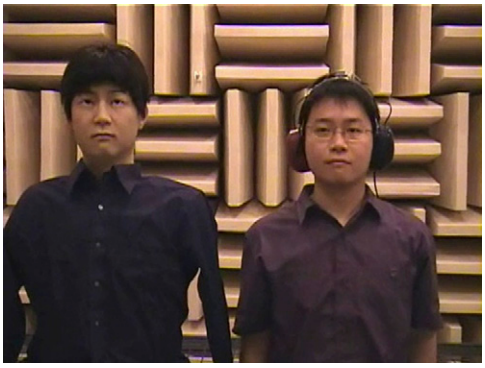


Fig. S1. Illustration of trial types and results of experiment S1. (*A*) All trials started with the listener fixating on an LED on one side, chosen at random. The Self, Source, and No-change trials were identical to those in the main experiment (main text and Fig. 1). The Cued-source trials were identical to the Source trials, except that the LED blinked in the middle of the trial, from 9.5 to 10.5 s. (*B*) Resetting of auditory streaming. The analysis is as in Fig. 2: raw data (*Upper*) and normalized data computed by selecting the trials in which listeners reported two streams at the 10-s point (*Lower*). Resetting was evaluated over the 6-s time window indicated by the shaded area.



Movie S1. The Telehead dummy head was located in an anechoic chamber. The head position of listeners was measured by a 3D head tracker placed on the top of the headphones. Their head motion was tracked in real-time and sent to the Telehead robotic system (in the actual experiment, the listener was located in a different room, see Fig. 1*A* for details).

Movie S1