The role of temporal coherence and temporal stability in the build-up of auditory grouping

Joseph Sollini[1], Kat Poole[2], Dominic Blauth-Muszkowski[2] and Jennifer Bizley[2]

1 Hearing Sciences, Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, England, United Kingdom

2 The Ear Institute, University College London, London, England, United Kingdom

Abstract

To form complex representations of sounds, i.e. auditory objects, the auditory system needs to make decisions about which information is part of one object and which is part of another. These decisions are usually taken over a short period of time at the beginning of the sound, known as build-up, by using the available grouping cues. Here we investigate the use of temporal coherence and temporal stability to influence subsequent grouping. We show that these two grouping cues behave independently from one another, except when put into conflict. In these situations, the cues available during the build-up period determine subsequent perception.

Introduction

It is often hard to fathom how the auditory system is able to turn a simple 2-dimensional signal, variations in pressure over time, into the complex scene of different auditory objects that we perceive. The auditory system uses 'grouping cues' to link together information arising from separate sources, or 'objects'[1-4]. For example, sound elements from common source will often fluctuate in power simultaneously[5]. These synchronous changes in amplitude across-frequency, known as temporal coherence, serve as a critical grouping cue [4]. Natural sound sources often have reliable sound statistics over time which also improve the segregation of auditory objects [6-8]. We shall call consistency in these statistics over time "stability". There is tolerance to small fluctuations in these statistics. For example, amplitude modulation rates in human speech fall over a restrictive range: 1-30Hz[9] and these statistics are used to aid the segregation of speech in mixtures[10]. We shall refer to the stability in temporal statistics as "temporal stability". An extreme example of this is temporal regularity where the temporal structure is repetitive over time such that the temporal statistics fall over a very narrow range[11].

Neither temporal coherence not temporal stability can be estimated instantaneously; in both cases a period of time is required to gather information, this process of accumulating knowledge about the sound is referred to as the "build-up" period. Build-up has been observed in a range of seemingly disparate auditory behaviours, such as: listening to speech-in-noise[12], speech in speech[6], auditory

stream formation[13], the discrimination of sound textures [14], comodulation masking release[15], the precedence effect[7] and transitions to new spectrotemporal sound statistics[16]. Generally speaking, the longer the build-up period the greater the behavioural benefit[7,14,17-19], though, this will asymptote and can decline due to inattention [20]. The optimal build-up period can be thought of as the shortest duration of build-up at which performance begins to asymptote, i.e. when additional build-up offers very little behavioural advantage. The optimal build-up period varies depending on the task, but appears to be between 0.4-2.5 seconds[7,12,14,17], though further modest gains can be made for more than a minute[13,21]. In the case of "regular" sounds, i.e. repeating spectrotemporal structure, this variation can be explained by the duration/period of the sound statistics themselves, i.e. the longer a repeating sequence the more time is required to learn that sequence[16].

Once an object has been built-up it can take several seconds for its influence to decay[13,22] unless object formation is reset by abrupt changes in an ongoing sound, e.g. a frequency, level or spatial deviation in one of the objects[21,23-26]. Overall, the existing literature suggests that build-up is a high-level evidence accumulation process[19] that influences subsequent perception. It can do this by either biasing perception towards, and/or enhancing perception of, the important features within an object, and operates on a range of grouping cues. This then serves a range of core auditory abilities, such as: segregation [6,13], detection [8,27] or discrimination[14]. It is not clear whether build-up is one general process or a number of processes (i.e. whether independent build-up mechanisms subserve different grouping cues). To better understand this, it is necessary to investigate interactions between different grouping cues in order to measure their influence upon one another. This is a particular important issue as current theories of sound segregation, i.e. the temporal coherence model [4], suggest that temporal coherence is the dominant grouping cue. To test this idea a direct comparison of temporal coherence and other grouping mechanisms is required.

Here, we investigate how temporal stability and temporal coherence interact with one another during build-up. Given the reasonably long-lasting effects of build-up (up to 1 second) these interactions could have important consequences on subsequent perception. Grouping processes have been shown to effect auditory stream segregation[13,28] and also sound detection[29-31]. We chose to measure grouping by employing a simple tone in background sound detection task. Increases in tone detection threshold represented an impaired ability to represent both sounds separately, whereas decreases in detection threshold signal an enhanced ability to segregate the tone from competing background sounds. We sought to promote the use of either temporal coherence and/or temporal stability as a grouping cue by selectively including or excluding these cues during the build-up period. By varying the background sound statistics during build-up, while providing both cues later in the background (when the tone was presented) we were able to bias the use of different cues and control subsequent grouping. Using this approach, we were able to show that temporal stability and coherence independently influence grouping and that manipulating temporal coherence during the build-up did not exert a greater effect than that of temporal stability. Our results demonstrate the importance of the build-up period on organising an auditory scene and thus the subsequent perception of auditory objects.

Materials and methods

Participants

19 listeners participated in the experiment. All were aged between 19 and 25 and none reported long-term or current problems with their hearing. Experimental procedures were carried out in accordance with the protocols approved by the research ethics committee of University College London and written informed consent was provided by each participant. Participants were given a short training session (usually lasting 15-20 minutes) to familiarise themselves with the experimental setup and the task itself before starting.

Stimulus and apparatus

Apparatus

Participants were seated in a double-walled sound attenuating booth (IAC, Winchester, UK), wearing headphones (Sennheiser HD 600), in front of a computer monitor and mouse. Sound was delivered through the headphones via a preamp connected to a signal generator (Tucker Davis Technology, RP2) which was in turn controlled by custom written software (MATLAB and TDT's RPvdsEX). Participants registered responses to sounds via a custom designed graphical user interface (MATLAB) through mouse clicks at the end of each trial.

Sound stimuli were generated de novo for each trial using MATLAB and loaded onto the signal generator at the start of each trial. All sounds were generated and presented with a ~48kHz sampling rate.

Background stimuli

To create the backgrounds modulated pure tones (800ms duration) were created with a random phase at the necessary frequencies for the narrowband (1kHz, 65dB SPL) and broadband conditions (1kHz as before but also at: 0.25, 0.5, 2 and 4kHz, termed 'flanker bands' where each was calibrated to be 65dB SPL).

The background was composed of two periods of time; the build-up (the first 500ms) and the masker (the next 300ms). The masker was always temporally stable (i.e. had a regular rhythm) with tone pips presented at 20 Hz, but could be narrowband or broadband depending on the bandwidth of the preceding precursor. The precursor could be narrow or broadband and could be temporally regular across all frequency bands, jittered within the signal band, the flanker bands, or both (see Fig. 2). Individual "pip" envelopes were generated using ½ a cycle of a 40 Hz cosine which were then positioned at regular intervals at 20Hz such that for temporally regular conditions, 25ms tone pips were presented at intervals of 50ms separated by 25ms intervals of silence. For temporally jittered conditions each pip onset within the pre-cursor was adjusted by a uniformly sampled random amount of time between the bounds ±25ms.

Signal

Participants were required to detect a short (25ms duration, 12ms $\cos^2$ ramp) pure tone signal (1 kHz) that was embedded in the background. Tone signals were presented in the trough between peaks 13 and 14 of the background, meaning they were centred, temporally, in the middle of the masker (e.g. Fig. 1A). Tones were varied in sound level according to a staircase rule (see procedure).

Testing protocol

Staircase procedure

An adaptive staircase procedure was used to vary signal level between trials. A 3up 1down design was used with 3 sequential rules for varying the change in sound level: rule 1 = 3 reversals with a 6 dB step size, rule 2 = 3 reversals with a 2 dB step size and rule 3 = 4 reversals with a 1dB step size. The tone signal level started each block at 65 dB SPL.

General procedure

Tone detection threshold for each of the 7 backgrounds were measured in blocks and 3 repeats of each threshold were measured. The intention was for the order of the blocks within a repeat to be randomised on each repeat cycle. Owing to an experimenter error this was only performed for 8 of the 19 participants. However, statistical comparison of the mean thresholds (across 3 repeats for each condition) for the two groups (i.e. randomised vs same order participants) demonstrated there was no significant effect of the order within a repeat cycle between the two groups (paired t-test, uncorrected for multiple comparisons to add confidence, $p>0.05$ for all conditions). Within each block a 3-interval forced choice design was used to probe tone signal detection. On a given trial 3 identical backgrounds, presented consecutively, were presented with a 1 second interval between each. Within one of these intervals a tone signal was added and the participant was asked to identify the interval containing the tone using a mouse click on a response screen.

Data analysis

Data analysis was performed in MATLAB. For each participant and testing block the threshold was calculated as the mean signal sound level across the final 4 reversals and averaged across the three repeats of each condition. Statistical analysis was performed in SPSS (IBM SPSS Statistics 27).

Results

Temporal stability and temporal coherence both independently influence grouping

To test the interactions between temporal stability temporal coherence we employed a simple tone in background stimulus design. In brief, participants (n=19) were presented the same background 3 times in consecutive intervals (0.8 second duration, 1 second inter-stimulus interval), in one of these intervals a tone was also presented, and the participant had to identify this interval (i.e. 3-IFC). From trial-to-trial the sound level of a target tone was varied with a 3 up, 1 down adaptive staircase in blocks to yield a threshold for each block (the mean of the last 4 reversals) and then for each condition (the mean of all 3 blocks/repeats for that condition).

4 backgrounds were compared in a 2x2 design (Fig. 1A), with two manipulations being tested. The effect of temporal coherence (in the masker) was tested by altering the bandwidth, using a narrowband (1kHz tone pips) and a broadband configuration (5 tone pip complex between 0.25 and 4kHz with 1 octave spacing). In addition, the effect of temporal stability was tested by varying only the stimulus statistics in the build-up of the background (i.e. first 500ms of the background) and using the same stimulus statistics in the masking portion of the background, across comparisons (500-800ms). The statistics in the masking period were always regular but the temporal structure in each frequency channel during the build-up could either be regular (i.e. temporal stability, Fig. 1A) or jittered (temporal instability, Fig1A). This yielded two narrowband conditions to measure the effect of temporal stability in the absence of temporal coherence (within channel): one with temporal stability ($TS^{+SIG}TC^{0}$, Fig. 1A) and one with a temporally jittered build-up ($TS^{-SIG}TC^{0}$) and two broadband conditions to allow the assessment of varying temporal stability on the use of temporal coherence: one with temporal stability and a temporally coherent build-up ($TS^{+}TC^{+}$) and one without ($TS^{-}TC^{-}$, though the masker did have temporal coherence).

Varying either bandwidth or temporal stability had a significant effect on subsequent tone detection thresholds (Fig 1B). Adding temporal coherence in the masker produced significant unmasking of the tone target, irrespective of temporal structure in the build-up (μ = 3.72 dB). Altering the temporal structure, by transitioning from a jittered to a temporally regular modulation rate, whether broadband or narrowband (Fig. 2A), produced a larger (μ = 9.39 dB) reduction in tone detection thresholds. These findings were confirmed with a two-way RM ANOVA with main effects of adding bandwidth and stability, and no significant interaction (stability: $F = 28.45$, $p = 0.000045$, coherence: $F = 5.171$, $p = 0.035$, stability*coherence. $F = 0.151$, $p = 0.702$).While when combined the effect of adding bandwidth was significant, post-hoc tests revealed this unmasking fell short of significance in the temporally stable (μ = 46.55dB vs 42.5dB, $TS^{+SIG}TC^{0}$ vs $TS^{+}TS^{+}$ respectively, paired t-test, sidak correction, $p = 0.241$) and unstable conditions individually (μ = 55.61dB vs 52.21dB, $TS^{-SIG}TC^{0}$ vs $TS^{-}TC^{-}$ respectively, paired t-test, sidak correction, $p = 0.374$). Post-hoc tests revealed the effect of stability was highly significant for both the narrowband conditions (μ = 55.61dB vs 46.55dB, $TS^{-SIG}TC^{0}$ vs $TS^{+SIG}TC^{0}$ respectively, paired t-test, sidak correction, $p = 0.000096$) and the broadband conditions (μ = 52.21dB vs 42.5dB, $TS^{-}TC^{-}$ vs $TS^{+}TC^{+}$ respectively, paired t-test, sidak correction, $p = 0.003$). Improvements in threshold produced by unmasking due to bandwidth (Fig. 1C) and stability (Fig. 1D) remained consistent across the other manipulation. These results suggest that varying temporal stability did not disrupt the across-channel benefit of temporal coherence as unmasking caused by adding bandwidth was similar irrespective of the temporal stability in the build-up (Fig. 1C). In addition, the effect of varying temporal stability was also similar regardless of the bandwidth (Fig. 1D). This was supported by the fact there was no interaction between the two effects. In addition,

stability produced larger reductions in threshold than adding coherent (μ = 9.39 vs 3.72dB, respectively) potentially suggesting its greater importance. To explore this further we turned to conditions in which we systematically altered the temporal structure of the signal band and flanking bands.

Build-up critically influences the subsequent use of both temporal stability and coherence cues

In our previous conditions it was difficult to quantify the relative contribution of temporal stability, coherence and bandwidth to subsequent perception. For example, providing stability and coherence in the build-up ($TS^{+}TS^{+}$) could create enhancement in grouping of the background and, hence, enhance tone detection. Alternatively, disrupting stability and coherence during the build-up ($TS^{-}TC^{-}$) could interfere with grouping and tone detection. When comparing these conditions, and observing a difference, it is unclear whether we are measuring either one or both. To allow quantification of the relative enhancement/interference on subsequent performance we introduced a "neutral" condition where the build-up period had been removed but where the masker contained temporal coherence (Fig 2A, $TS^{0}TC^{0}$). This allowed us to determine the relative direction of change in performance that was elicited by the inclusion of each respective precursor.

We also sought to quantify the contribution of each cue toward subsequent object formation. The previous conditions (Fig. 1) suggest that temporal stability and coherence are independent processes, as the size of one was not significantly varied when changing the parameters of the other (and confirmed by the non-significant interaction term in the ANOVA). However, in the narrowband condition the only grouping cue available was temporal stability (in the $TS^{+SIG}TC^{0}$ condition) and when this transitioned from jittered to regular ($TS^{+SIG}TC^{0}$ vs $TS^{-SIG}TC^{0}$) this cue would suggest a new object had begun. In addition, in the broadband condition both cues were consistent, i.e. in $TS^{+}TC^{+}$ both are consistent with a continuing object from build-up to masker and $TS^{-}TC^{-}$ both are consistent with two separate objects (one in the build-up and another during the masker period). Therefore, for all of these conditions there was no conflict between grouping cues as both cues either promoted one or two objects.

If the dominant cue for grouping is temporal coherence we would predict that where temporal coherence and another cue conflict the temporal coherence cue would dominate and thus define grouping. To test this idea we generated two additional conditions, each designed to promote different grouping configurations. The first promoted the use of temporal stability, by continuous regularity in the signal channel, while disrupting temporal coherence ($TS^{+SIG}TC^{-}$, Fig. 2A), in effect promoting the signal channel as one object during the masker period. The second sought to block both temporal coherence between flankers and the signal band but also temporal stability within the signal band ($TS^{-SIG}TC^{-}$). These 5 conditions combined gave the opportunity to not just test the relative size and direction of masking, i.e. by comparing the thresholds of 4 of the broadband conditions condition ($TS^{+}TC^{+}$, $TS^{-}TC^{-}$, $TS^{+SIG}TC^{-}$ and $TS^{-SIG}TC^{-}$) relative to the neutral broadband control condition ($TS^{0}TC^{0}$, Fig. 2A).

A significant effect of varying the masker properties during build-up was found across the 5 conditions (RM-ANOVA of all five conditions, n = 19, F = 11.729, p = 0.0000002). Tone detection thresholds were lowest when the build-up contained identical sound statistics to the masker period ($TS^{+}TC^{+}$, μ = 42.5dB SPL) suggesting an enhancement relative to the neutral condition ($TS^{0}TC^{0}$, μ = 45.12dB SPL, 2.62dB difference), however, post-hoc test revealed this was not significant (paired t-test, sidak correction, p = 0.897). By contrast, conditions with temporal instability in the build-up produced interference, relative to the neutral condition (Fig. 2B: μ = 52.21, 54.5, 48.89dB SPL or 7.1, 9.38 and 3.77dB increases, $TS^{-}TC^{-}$, $TS^{-SIG}TC^{-}$ and $TS^{+SIG}TC^{-}$ respectively). Post-hoc tests revealed interference was significant when all frequency channels were jittered ($TS^{-}TC^{-}$) or the signal band was jittered ($TS^{-SIG}TC^{-}$) but not when the signal band had temporal stability ($TS^{+SIG}TC^{-}$, paired t-test, sidak correction, p = 0.019, 0.007, 0. 699 for $TS^{-}TC^{-}$, $TS^{-SIG}TC^{-}$ and $TS^{+SIG}TC^{-}$ respectively). Interestingly, while neither the $TS^{+}TC^{+}$ or $TS^{+SIG}TC^{-}$ conditions were significantly different from the neutral condition ($TS^{0}TC^{0}$), they were significantly different from one another (6.39dB difference, paired t-test, sidak correction, p = 0.049), suggesting they produce small changes from the neutral condition that only become significant when contrasted. In addition, $TS^{+}TC^{+}$ thresholds were significantly lower than the other two unstable conditions (paired t-test, sidak correction, p = 0.00031 and 0.0046, $TS^{+}TC^{+}$ vs $TS^{-}TC^{-}$ and $TS^{-SIG}TC^{-}$ respectively). Together these results demonstrate that the preceding sound statistics can significantly affect subsequent sound in noise detection. Interference is largest when the signal bands temporal statistics are unstable, through jittering ($TS^{-}TC^{-}$ and $TS^{-SIG}TC^{-}$, i.e. disrupting use of the temporal stability grouping cue in the signal band). Smaller changes are observed when temporal stability in the build-up is present ($TS^{+SIG}TC^{-}$, i.e. when only disrupting temporal coherence) or providing both cues ($TS^{+}TC^{+}$).

Discussion

We employed a simple paradigm to assess the relative contribution of temporal coherence and temporal stability, during the build-up period to subsequent sound segregation. We found that the grouping cues available during build-up critically shaped subsequent perception. Introducing temporal instability (i.e. jitter, $TS^{-SIG}TC^{0}$ and $TS^{-}TC^{-}$) during build-up significantly hampered sound segregation, and elevated tone detection thresholds, relative to conditions with temporal stability ($TS^{+SIG}TC^{0}$ and $TS^{+}TC^{+}$). In addition, disrupting temporal coherence during the build-up ($TS^{+SIG}TC^{-}$ and $TS^{-SIG}TC^{-}$ vs $TS^{-}TC^{-}$) also produced significant increases in thresholds. The benefit attributable to adding temporal stability was a 2.62dB reduction in thresholds by ($TS^{0}TC^{0}$ vs $TS^{+}TC^{+}$, individual comparison was not significant) and the benefit attributable to temporal coherence was 3.72dB ($TS^{+SIG}TC^{0}$ vs $TS^{+}TC^{+}$ and $TS^{-SIG}TC^{0}$ vs $TS^{-}TC$). The pattern of results demonstrates the importance of build-up to subsequent perception: in all cases the masker for the tone detection task was both stable and coherent, yet the influence of the precursor was strong enough as to over-ride the the introduction of beneficial grouping prior to tone onset.

Build-up promotes object permanence through active interference

Previous work has suggested that surrounding (i.e. before and after) a period of temporal coherence with random temporal structure (similar to $TS^{-}TC^{-}$) disrupts the benefits of temporal coherence[32,33].

This would be expected as pre[34] and post-cursor[30] sounds influence the use of temporal coherence. While our study includes only a precursor, these studies represent the closest work to this performed here. Grose et al. presented 4 continuous narrow bands of noise (20 Hz wide centred at 804, 1200, 1747 and 2503Hz) with a random temporal envelope (i.e. temporally irregular when noise bands were sufficiently spaced) that was switched from being comodulated (temporally coherent) to random (random fringes surrounding a comodulated period), while trained participants (n=9) were asked to detect pure tones[33]. The authors found that incoherently modulated fringes produced a small interference with comodulation masking release (elevated thresholds by ~2dB). Subsequent work by, largely, the same authors[32] demonstrated a similar effect in highly trained participants (n=4), though with a much larger effect size (~10dB). Similarly, a large interference (9.55dB) was observed in this study when replacing a comodulated build-up with a temporally jittered one (n=19, minimal training provided).

Based on this effect Grose et al. suggested that random fluctuations in the envelope (equivalent to our jitter) before and after the comodulated period could impede the formation of an internal model of across-channel coherence, i.e. temporal coherence. However, in our manipulations we found very little evidence for interactions between temporal coherence and temporal stability in the build-up. This relationship was tested directly in a 2x2 design (Fig. 1) and we observed that the temporal coherence advantage did not change significantly when temporal stability was disrupted and vice versa. In addition, no significant interaction was found between the two. However, there were a number of differences between the work here and this earlier work that may explain these differences. The first is that Grose and colleagues present temporally coherent periods in continuous noise, this meant there were pre and post-cursors and also that the background had more time to build-up. As mentioned above post-cursors have been shown to disrupt temporal coherence[30] and so it is not possible to rule out their influence in Grose et al. Another possibility is the use of a second grouping cue in the present study (i.e. temporal stability). In the previous studies temporal coherence was varied but there was no intended control of temporal stability. However, narrowband (20Hz) noise was used instead of tones, this stimulus creates low frequency envelope fluctuations meaning they had temporal envelope statistics (i.e. the envelope fluctuations fell over a restricted range of amplitude modulation frequencies). In this way it could be argued this stimulus has temporal stability, as their envelope statistics are stable over time, though to a much lesser extent than in the current study where sinusoidal envelopes were applied (i.e. perfect regularity in temporal structure). After a jittered build-up period ($TS^{-}TC^{-}$) it is possible that the sudden introduction of temporal stability and temporal coherence formed a strong cue that a new object had arrived, i.e. the transition between build-up and masker represents a large change in the ongoing sound statistics and so would be more likely to be treated as the introduction of a new sound than the continuation of an existing one. Finally, in their “jittered” conditions the temporally coherent background and the target tone were onset at identical times whereas there was an onset asynchrony when there was temporal stability (the comodulation was present in the build-up and the masking period). This means there was a grouping cue between tone and masker in all manipulations except those where the temporal coherence was continuous. This would reduce segregation and increase thresholds possibly accounting for the changes that the authors suggest might be evidence of a jittered build-up blocking comodulation masking release and, hence, the temporal coherence grouping cue. Overall, the balance of evidence suggests that jitter does not affect the subsequent use of temporal coherence as a grouping (or at least not for very long). It seems more probable this impacts the use of temporal stability cues (stability in the modulation statistics) as this produces a large and significant interference in subsequent detection.

Implications for the temporal coherence theory

The strong version of temporal coherence theory states that temporal coherence is the primary grouping cue. Here we investigated the influence of varying both temporal coherence and temporal stability. We did not find evidence for the primacy of temporal coherence in our paradigm design. In Fig. 1 temporal stability appears to exert a greater influence on grouping as disrupting temporal stability (Fig. 1C) produced greater unmasking than adding temporal coherence (Fig. 1D). This, however, is an unfair comparison because, as this work shows, disrupting cues appears to produce a larger deficit than adding a cue (Fig. 2), e.g. adding both cues ($TS^{+}TC^{+}$ vs $TS^{0}TC^{0}$) produced a smaller effect than disrupting them ($TS^{-}TC^{-}$ vs $TS^{0}TC^{0}$). Therefore, the comparison should be like for like (measuring either enhancement or disruption alone). Within our design we can measure the size of the effect of disrupting coherence (in the build-up) by measuring the effect of adding incoherent flanking bands to the narrowband condition ($TS^{+SIG}TC^{0}$ vs $TS^{+SIG}TC^{-}$). This suggests the effect of adding disruption of temporal coherence is a 2.34dB decrease in threshold. We can also infer the effect of disrupting temporal stability by calculating the effect of disrupting all channels and subtracting the effect of adding disrupted flanking bands, i.e. ($TS^{-}TC^{-}$ - $TS^{0}TC^{0}$) - ($TS^{+SIG}TC^{0}$ – $TS^{+SIG}TC^{-}$). This suggests disrupting temporal stability in the signal channel is 4.7db. It is unlikely that disrupting coherence produces a larger effect given disrupting temporal stability produced twice the effect size and there was no statistical evidence for a difference between the two (t-test, $p = 0.9973$).

References

1 Carlyon, R. P. How the brain separates sounds. *Trends Cogn Sci* **8**, 465-471, doi:10.1016/j.tics.2004.08.008 (2004).

2 Cherry, R. & Rubinstein, A. A comparison of two approaches to assessment of speech recognition ability in single cases. *J Am Acad Audiol* **16**, 54-62; quiz 63-54, doi:10.3766/jaaa.16.1.6 (1953).

3 McDermott, J. H. The cocktail party problem. *Curr Biol* **19**, R1024-1027, doi:10.1016/j.cub.2009.09.005 (2009).

4 Shamma, S. A., Elhilali, M. & Micheyl, C. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* **34**, 114-123, doi:10.1016/j.tins.2010.11.002 (2011).

5 Nelken, I., Rotman, Y. & Bar Yosef, O. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **397**, 154-157, doi:10.1038/16456 (1999).

6 Best, V., Ozmeral, E. J., Kopco, N. & Shinn-Cunningham, B. G. Object continuity enhances selective auditory attention. *Proc Natl Acad Sci U S A* **105**, 13174-13178, doi:10.1073/pnas.0803718105 (2008).

7 Brandewie, E. & Zahorik, P. Prior listening in rooms improves speech intelligibility. *J Acoust Soc Am* **128**, 291-299, doi:10.1121/1.3436565 (2010).

8 Sohoglu, E. & Chait, M. Detecting and representing predictable structure during auditory scene analysis. *Elife* **5**, doi:10.7554/eLife.19113 (2016).

9 Cumming, R., Wilson, A. & Goswami, U. Basic auditory processing and sensitivity to prosodic structure in children with specific language impairments: a new look at a perceptual hypothesis. *Front Psychol* **6**, 972, doi:10.3389/fpsyg.2015.00972 (2015).

10 Myers, B. R., Lense, M. D. & Gordon, R. L. Pushing the Envelope: Developments in Neural Entrainment to Speech and the Biological Underpinnings of Prosody Perception. *Brain Sci* **9**, doi:10.3390/brainsci9030070 (2019).
11 Andreou, L. V., Kashino, M. & Chait, M. The role of temporal regularity in auditory segregation. *Hear Res* **280**, 228-235, doi:10.1016/j.heares.2011.06.001 (2011).
12 Ben-David, B. M., Tse, V. Y. & Schneider, B. A. Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker? *Hear Res* **290**, 55-63, doi:10.1016/j.heares.2012.04.022 (2012).
13 Bregman, A. S. Auditory streaming is cumulative. *J Exp Psychol Hum Percept Perform* **4**, 380-387, doi:10.1037//0096-1523.4.3.380 (1978).
14 McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nat Neurosci* **16**, 493-498, doi:10.1038/nn.3347 (2013).
15 McFadden, D. Comodulation masking release: effects of varying the level, duration, and time delay of the cue band. *J Acoust Soc Am* **80**, 1658-1667, doi:10.1121/1.394277 (1986).
16 Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J. & Chait, M. Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proc Natl Acad Sci U S A* **113**, E616-625, doi:10.1073/pnas.1508523113 (2016).
17 Beeston, A. V., Brown, G. J. & Watkins, A. J. Perceptual compensation for the effects of reverberation on consonant identification: evidence from studies with monaural stimuli. *J Acoust Soc Am* **136**, 3072, doi:10.1121/1.4900596 (2014).
18 Boubenec, Y., Lawlor, J., Gorska, U., Shamma, S. & Englitz, B. Detecting changes in dynamic and complex acoustic environments. *Elife* **6**, doi:10.7554/eLife.24910 (2017).
19 Nguyen, Q. A., Rinzel, J. & Curtu, R. Buildup and bistability in auditory streaming as an evidence accumulation process with saturation. *PLoS Comput Biol* **16**, e1008152, doi:10.1371/journal.pcbi.1008152 (2020).
20 Thompson, S. K., Carlyon, R. P. & Cusack, R. 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, doi:10.1037/a0021925 (2011).
21 Anstis, S. S., S. Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance* **11(3)**, 257–271, doi:https://doi.org/10.1037/0096-1523.11.3.257 (1985).
22 Beauvois, M. W. & Meddis, R. Time decay of auditory stream biasing. *Percept Psychophys* **59**, 81-86, doi:10.3758/bf03206850 (1997).
23 Haywood, N. R. & Roberts, B. Build-up of the tendency to segregate auditory streams: resetting effects evoked by a single deviant tone. *J Acoust Soc Am* **128**, 3019-3031, doi:10.1121/1.3488675 (2010).
24 Haywood, N. R. & Roberts, B. Build-up of auditory stream segregation induced by tone sequences of constant or alternating frequency and the resetting effects of single deviants. *J Exp Psychol Hum Percept Perform* **39**, 1652-1666, doi:10.1037/a0032562 (2013).
25 Roberts, B., Glasberg, B. R. & Moore, B. C. J. Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. *J Exp Psychol Hum Percept Perform* **34**, 992-1006, doi:10.1037/0096-1523.34.4.992 (2008).
26 Rogers, W. L. & Bregman, A. S. An experimental evaluation of three theories of auditory stream segregation. *Percept Psychophys* **53**, 179-189, doi:10.3758/bf03211728 (1993).
27 Sollini, J. & Chadderton, P. Comodulation Enhances Signal Detection via Priming of Auditory Cortical Circuits. *J Neurosci* **36**, 12299-12311, doi:10.1523/JNEUROSCI.0656-16.2016 (2016).
28 Noorden, v. Temporal coherence and the perception of temporal position in tone sequences. *Perception research (IPO)* (1975).
29 Christiansen, S. K. & Oxenham, A. J. Assessing the effects of temporal coherence on auditory stream formation through comodulation masking release. *J Acoust Soc Am* **135**, 3520-3529, doi:10.1121/1.4872300 (2014).

30 Dau, T., Ewert, S. & Oxenham, A. J. Auditory stream formation affects comodulation masking release retroactively. *J Acoust Soc Am* **125**, 2182-2188, doi:10.1121/1.3082121 (2009).
31 Verhey, J. L., Ernst, S. M. & Yasin, I. Effects of sequential streaming on auditory masking using psychoacoustics and auditory evoked potentials. *Hear Res* **285**, 77-85, doi:10.1016/j.heares.2012.01.006 (2012).
32 Grose, J. H., Buss, E. & Hall, J. W., 3rd. Within- and across-channel factors in the multiband comodulation masking release paradigm. *J Acoust Soc Am* **125**, 282-293, doi:10.1121/1.3023067 (2009).
33 Grose, J. H., Hall, J. W., 3rd, Buss, E. & Hatch, D. R. Detection of spectrally complex signals in comodulated maskers: effect of temporal fringe. *J Acoust Soc Am* **118**, 3774-3782, doi:10.1121/1.2108958 (2005).
34 McFadden, D. & Wright, B. A. Temporal decline of masking and comodulation masking release. *J Acoust Soc Am* **92**, 144-156, doi:10.1121/1.404279 (1992).

Figures

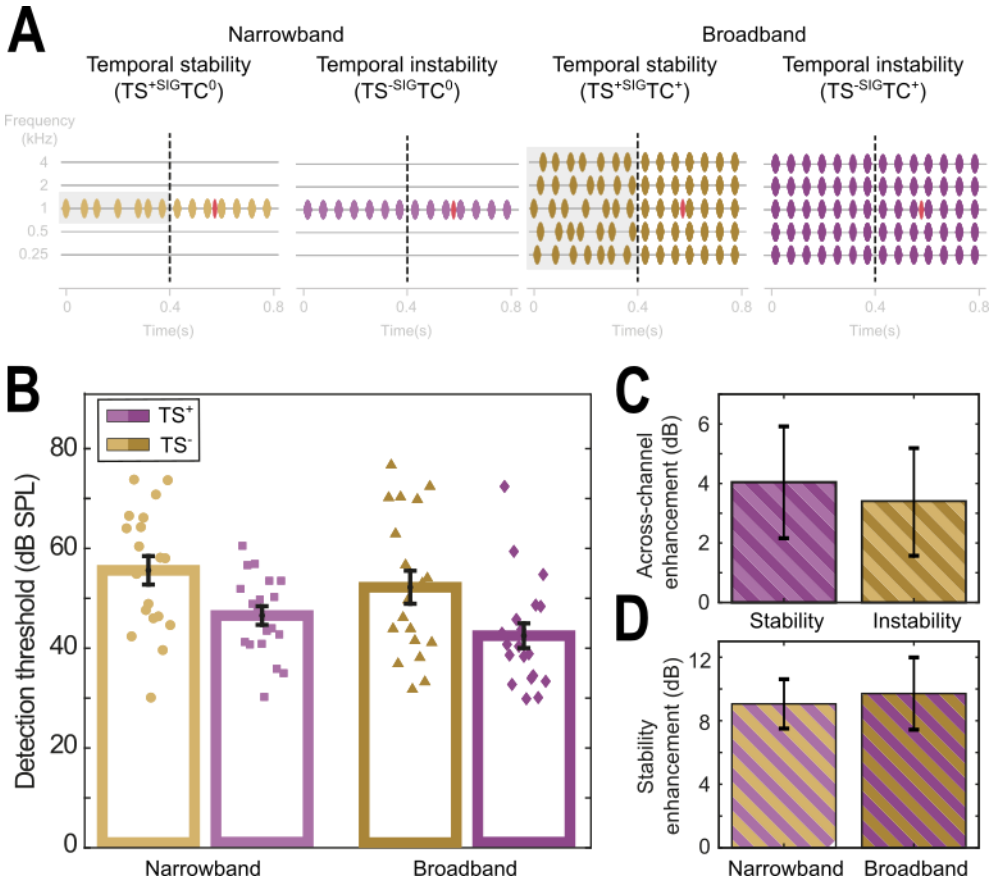

Fig. 1. Temporal coherence and temporal stability during build-up significantly affect grouping with no evidence of an interaction between them. A) Stimulus schematic. Each stimulus was composed of a background (purple/green) and a pure tone signal (red). The backgrounds were a series of pure tone pips presented in either: Left two panels) a narrowband (TC^0 notation) or Right two panels) broadband ($TS^{+/-}$ notation, 5 x 1 octave spaced tones centred at 1kHz) configuration. Each background was composed of a pre-cursor (left of the vertical dashed line) and a masker (right of the vertical dashed line). The pre-cursor could either promote temporal stability (purple and TS+ notations, including TS^{+SIG}) or disrupt temporal stability (green and TS^- notation, including TS^{-SIG}). B) Pure-tone detection thresholds for each condition in (A). C) Across-channel unmasking, i.e. the decrease in masked threshold attributable to adding tones with the same temporal structure as in the build-up and masker ($TS^+ = TS^{+SIG}TC^0$ vs $TS^{+SIG}TC^0$ and $TS^+ = TS^{-SIG}TC^0$ vs TS^-TC^-) was similar regardless of pre-cursor (TS^+ = ~3.5 dB, TS^- = ~3.8 dB). D) Stability unmasking, i.e. the decrease in masked threshold attributable to the build-up containing temporally stable vs instable structure (Narrowband = $TS^{+SIG}TC^0$ vs $TS^{-SIG}TC^0$ and Broadband = TS^+TC^+ vs TS^-TC^-) was similar regardless of pre-cursor (Narrowband: ~9.7 dB, Broadband: ~9.4 dB).

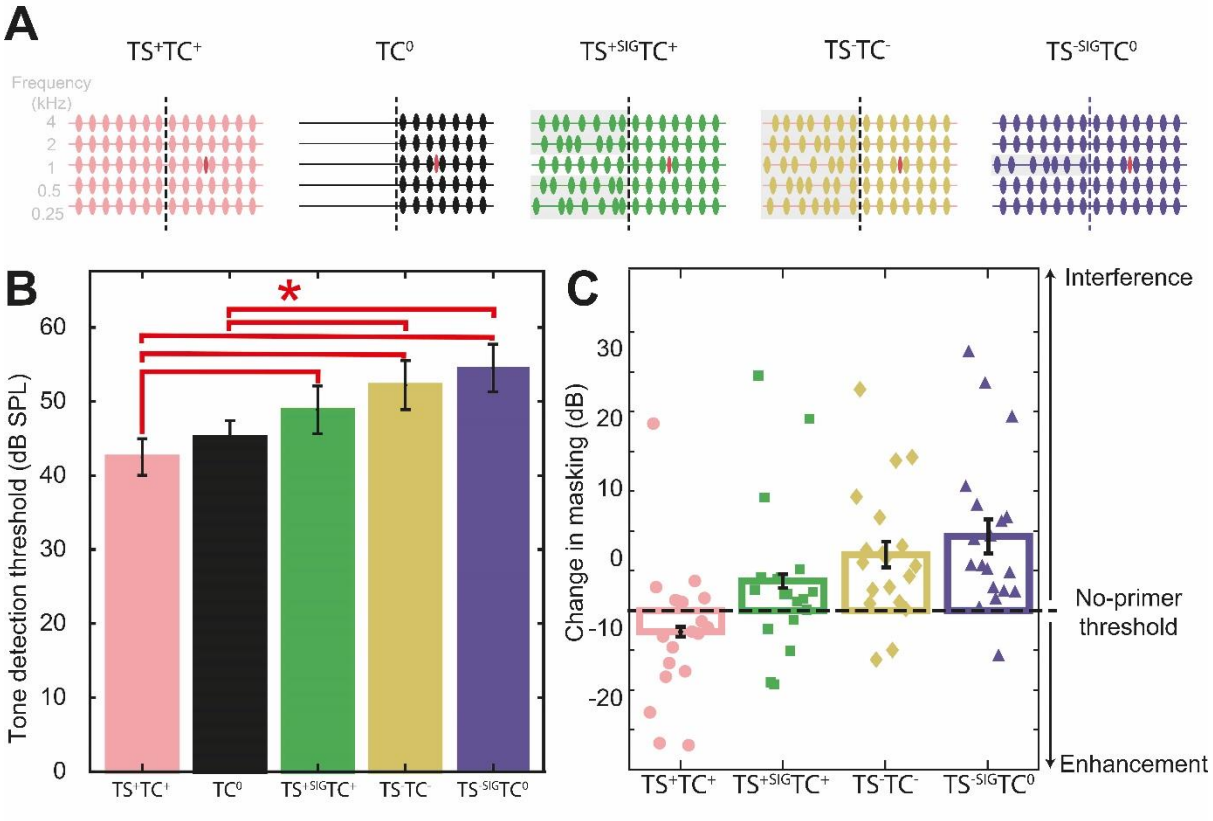

Fig. 2. Instability in sound statistics (either temporal coherence or stability) cause interference in the grouping process. A) Stimulus schematic. Five conditions were contrasted: a neutral condition (i.e. no build-up: TS^0TC^0) and four where the build-up was varied). These were designed to promote the use of just temporal stability ($TS^{+SIG}TC^-$), disrupt temporal stability in all channels (TS^-TC^-) and disrupt temporal stability in the signal channel and disrupt the use of temporal coherence ($TS^{-SIG}TC^-$). B) Changes in masked threshold compared to the neutral condition (>0 = pre-cursor interferes with detection of the tone, <0 = pre-cursor enhances tone detection).