The ability of the auditory system to cope with temporal subsampling depends on the hierarchical level of processing

Benedikt Zoefel[a,b], Naveen Reddy Pasham[c], Sasskia Brüers[a,b] and Rufin VanRullen[a,b]

Evidence for rhythmic or 'discrete' sensory processing is abundant for the visual system, but sparse and inconsistent for the auditory system. Fundamental differences in the nature of visual and auditory inputs might account for this discrepancy: whereas the visual system mainly relies on spatial information, time might be the most important factor for the auditory system. In contrast to vision, temporal subsampling (i.e. taking 'snapshots') of the auditory input stream might thus prove detrimental for the brain as essential information would be lost. Rather than embracing the view of a continuous auditory processing, we recently proposed that discrete 'perceptual cycles' might exist in the auditory system, but on a hierarchically higher level of processing, involving temporally more stable features. This proposal leads to the prediction that the auditory system would be more robust to temporal subsampling when applied on a 'high-level' decomposition of auditory signals. To test this prediction, we constructed speech stimuli that were subsampled at different frequencies, either at the input level (following a wavelet transform) or at the level of auditory features (on the basis of LPC vocoding), and presented them to human listeners. Auditory recognition was significantly more robust to subsampling in the latter case, that is on a relatively high level of auditory processing. Although our results do not directly demonstrate perceptual cycles in the auditory domain, they (a) show that their existence is possible without disrupting temporal information to a critical extent and (b) confirm our proposal that, if they do exist, they should operate on a higher level of auditory processing. *NeuroReport* 26:773–778 Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved.

NeuroReport 2015, **26**:773–778

Keywords: auditory, high level, perceptual cycles, robustness, sampling

[a]Université Paul Sabatier, Toulouse, France, [b]Centre de Recherche Cerveau et Cognition (CerCo), CNRS, UMR5549, Pavillon Baudot CHU Purpan, Toulouse Cedex, France and [c]Indian Institute of Technology, Bhubaneswar, Odisha, India

Correspondence to Benedikt Zoefel, Centre de Recherche Cerveau et Cognition (CerCo), CNRS, UMR5549, Pavillon Baudot CHU Purpan, BP 25202, 31052 Toulouse Cedex, France
Tel: +33 562 746 131; fax: +33 562 172 809; e-mail: zoefel@cerco.ups-tlse.fr

Received 9 June 2015 **accepted** 15 June 2015

Introduction

Recent research suggests that the visual system does not continuously monitor the environment, but rather samples it, cycling between 'snapshots' at discrete moments in time (perceptual cycles; for a review, see VanRullen *et al.* [1]). Interestingly, most attempts at discovering analogous perceptual cycles in the auditory system failed [2,3], indicating crucial differences between the visual and the auditory systems. A reason for this becomes evident when comparing the temporal structure of visual and auditory stimuli: whereas visual scenes are relatively stable over time, auditory input changes rapidly over time. In fact, whereas the visual system might rely particularly on the spatial dimension, time might be the most important factor for the auditory system [4] – and thus, subsampling auditory input in the time domain might destroy essential information [1]. Does this mean that perceptual cycles cannot be found in the auditory domain because it is impossible to subsample the auditory stream without losing important information? In this article, we argue that this is not necessarily the case – rather, it is possible that subsampling does take place in the auditory system, but on a relatively 'high' level of auditory processing: auditory information might be more temporally stable after a certain amount of feature extraction, enabling auditory subsampling without a significant loss of information. Thus, in this study, temporal subsampling was not only applied to the direct input to the auditory system, but we also subsampled the auditory stream on a higher-level representation (i.e. on the output level of a vocoder extracting auditory features by the use of linear predictive coding, LPC [5]). We predicted that the auditory system may prove significantly more robust to subsampling on the level of auditory features than when a similar subsampling was applied on the input level. We tested auditory vulnerability in a two-back recognition task (see Methods). An improved performance in this task for stimuli subsampled on a higher-level representation than for those subsampled at the input level would support the possibility that auditory perceptual cycles operate on a hierarchically high level of auditory processing.

Methods

Participants

Seven participants (four women, mean age 26.2 years), all fluent in English, volunteered to participate in the

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (*www.neuroreport.com*).

0959-4965 Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved. DOI: 10.1097/WNR.0000000000000422

experiment. All participants provided written informed consent, reported normal hearing, and received compensation for their time. The experimental protocol was approved by the relevant ethical committee at Centre National de la Recherche Scientifique (CNRS).

Stimulus construction

One original 10-min audio sequence [sampling rate (SR) = 44 100 Hz], a recording of a male native English speaker reading parts of a classic novel, was used as the primary stimulus in our experiment. The audio recording was cut into 200 3-s long 'snippets'. These snippets were then subsampled, either at the input level ('input condition'; i.e. at the level of the very input to the auditory system, such as in the cochlea; Fig. 1, top) or at the level of auditory features ('feature condition'; i.e. at a level beyond cochlear processing; Fig. 1, bottom). 'Subsampling' a given input stream does not necessarily mean 'forgetting' or 'ignoring' information. It might just be that the temporal order of information is lost, while the information itself is preserved. Thus, in our study, for both conditions (input and feature), we simulated 'subsampling' of the auditory system by shuffling auditory samples within a given time interval: for a SR of 4 Hz, for instance, all samples within a 250-ms window were shuffled. Of course, the larger this interval, the more difficult for the system to restore the exact (order of) information. However, we hypothesized that this restoration would be easier if the subsampling takes place in the auditory feature domain than when the input is subsampled at the input level as the former is temporally more stable. For every snippet, to prevent the use of static information for recognition, two subsampled versions were created by starting the shuffling interval either on the first sample or on the nearest sample to 1 + SR/SF/2. Sample sound files are available for both conditions as Supplemental digital content, 1–18 (*http://links.lww.com/WNR/A322*, *http://links.lww.com/WNR/A323*, *http://links.lww.com/WNR/A324*, *http://links.lww.com/WNR/A325*, *http://links.lww.com/WNR/A326*, *http://links.lww.com/WNR/A327*, *http://links.lww.com/WNR/A328*, *http://links.lww.com/WNR/A329*, *http://links.lww.com/WNR/A330*, *http://links.lww.com/WNR/A331*, *http://links.lww.com/WNR/A332*, *http://links.lww.com/WNR/A333*, *http://links.lww.com/WNR/A334*, *http://links.lww.com/WNR/A335*, *http://links.lww.com/WNR/A336*, *http://links.lww.com/WNR/A337* *http://links.lww.com/WNR/A338*, *http://links.lww.com/WNR/A339*).

Subsampling at the input level

For the input condition (Fig. 1, top), snippets were converted into the wavelet domain to approximate cochlear transduction (continuous Morlet wavelet transform of order 6). Snippets were divided into intervals, with the reciprocal of this interval corresponding to the desired subsampling frequency (SF). The amplitudes of the complex wavelet coefficients within the respective interval were shuffled. The phase information at the first sample of each interval was preserved and interpolated to avoid artifacts created by discrete phase transitions. After shuffling, final snippets for the time condition were obtained by applying the inverse wavelet transform.

Subsampling at the feature level

For the feature condition (Fig. 1, bottom), auditory features for each snippet were extracted using an LPC vocoder [5]. More precisely, linear prediction coefficients α_k were constructed such that each auditory sample s at time t can be seen as a linear combination of past p samples (p is the order of prediction):

$$s'(t) = -\sum_{k=1}^{p} a_k \times s(t-k), \tag{1}$$

where $s'(t)$ is the predicted auditory sample. A pre-emphasis filter [6] was applied on s before $s'(t)$ was calculated. 11 α_k (among which the first is unity) were calculated for each frame of 30 ms, with 20 ms between centers of subsequent frames (resulting in an overlap of 10 ms between frames). For each frame fr, after a Hamming window was applied, α_k were constructed using the method of least squares, that is the following total prediction error E was minimized:

$$E(fr) = \sum_{t=-\infty}^{\infty} e^2(fr,t), \tag{2}$$

where

$$\begin{aligned} e(fr,t) &= s(fr,t) - s'(fr,t) \\ &= s(fr,t) + \sum_{k=1}^{p} a_k(fr) \times s(fr,t-k). \end{aligned} \tag{3}$$

This was done using the Levinson–Durbin algorithm, which we do not explain in detail here, but which is described thoroughly in the relevant literature [7,8].

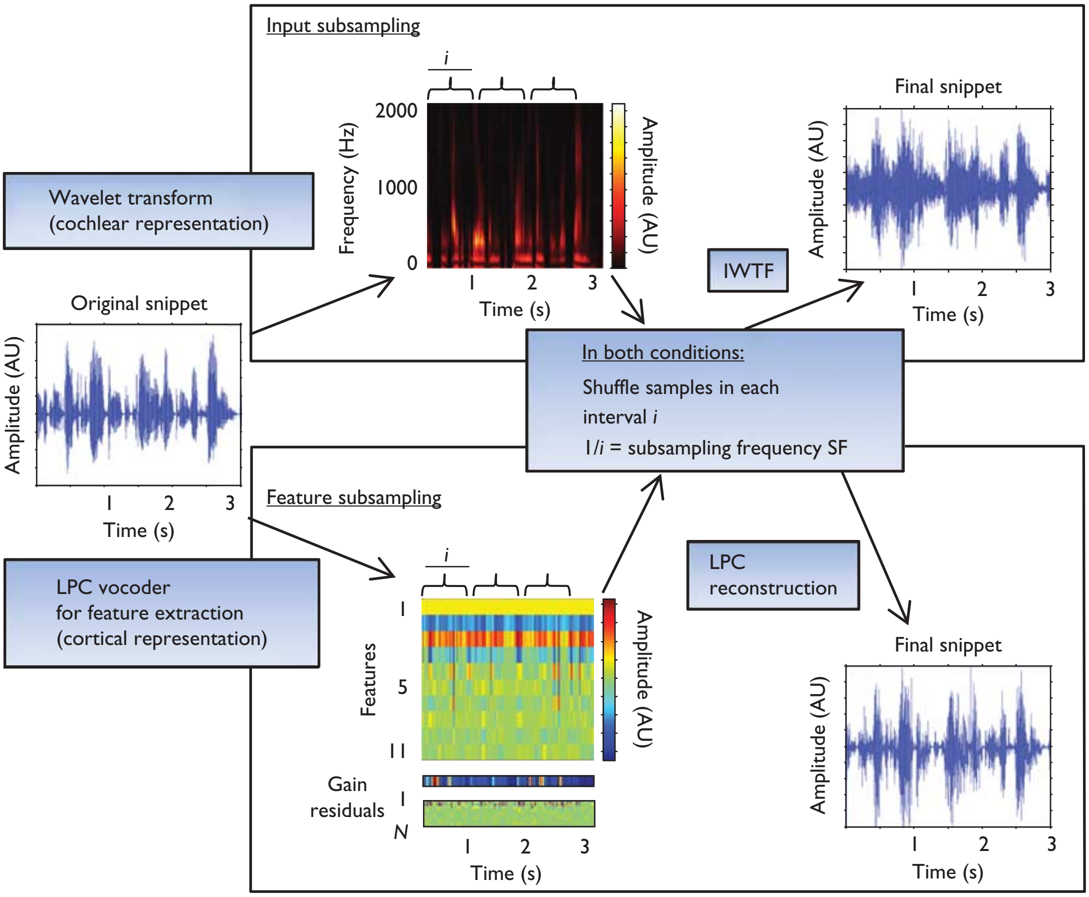
Two more parameters were extracted for each frame: the gain g (defined as the power of the speech signal in each frame) and the residual r:

$$r(fr,t) = \frac{e(fr,t)}{g(fr)}. \tag{4}$$

Discrete cosine transform (DCT) was applied to the residual of each frame; all except the first 50 DCT coefficients were discarded (as energy of the speech signal is concentrated in those 50 coefficients) and the inverse DCT was applied to obtain residuals with an improved signal-to-noise ratio [9]. Gaussian noise was added to the residuals (signal–noise ratio ~1:1) to improve the final sound quality of the reconstructed speech snippets.

For subsampling, α_k, gain, and residual were always shuffled together (i.e. α_k, gain and residual for a given frame were never separated). The step size of 20 ms

Fig. 1



Overview of the experimental approach. Original snippets were subsampled, either at the input level (input condition; top) or at the level of auditory features (feature condition; bottom). In both conditions, subsampling was realized by shuffling auditory samples in a certain time interval, with the reciprocal of this interval corresponding to the SF in the respective snippet. This shuffling was performed in the wavelet domain (corresponding to a cochlear representation of the sound) for the input condition and in the feature domain (obtained by an LPC vocoder; corresponding to a cortical representation of the sound) for the feature condition (see Methods). Note that subsampling in the input domain corresponds to simulating perceptual cycles on the level of the cochlea, whereas subsampling in the feature domain simulates perceptual cycles on a higher level (beyond cochlear processing) of the auditory pathway.

between frames restricted our maximal SF to 50 Hz, without any additional subsampling/shuffling. After shuffling, final snippets for the feature condition were obtained by filtering, in each frame, the residual, multiplied by the gain in the respective frame, by the obtained α_k:

$$s_{\mathrm{final}}(fr,t) = \sum_{k=1}^{p} a_k(fr) \times r_g(fr,t-k), \qquad (5)$$

where $r_g(fr,t)$ is $r(fr,t) \times g(fr)$.

Experimental paradigm

For both conditions, snippets were presented (separated by 1 s blank intervals) in a randomized order to our participants, who were instructed to perform a two-back task: they were asked to indicate by a button press any snippet that matched the one presented two snippets ago. These two-back repeats occurred randomly with a probability of 33%. Whenever a two-back repeat occurred in the sequence, it was always between nonidentical subsampled versions (with the same subsampling interval durations, but differing in the exact delay at which subsampling intervals were applied; see above). Stimuli were presented in separate blocks of 30 snippets. In each block, a different SF was applied. Participants completed 60 trials for each SF and condition. The two conditions ('input' and 'feature') were tested on separate days.

Data analyses

We hypothesized that the auditory system is more robust to temporal subsampling at the level of auditory features (feature condition) than at the input level (input condition). This robustness was tested in an auditory recognition task for snippets of different SF. Of course, with decreasing SF, performance will decline in both conditions. However, if our hypothesis is true, the precise SF where auditory recognition starts to decline will be lower for the feature than for the input condition.

We defined auditory recognition as d', the sensitivity of our participants in the two-back task. d' takes into account both correct responses (participants' response 'repeat' when there was actually a repeat) and false alarms (participants' response 'repeat' when there was no repeat):

$$d' = z(\text{hits}) - z(\text{false alarms}),$$

where $z(p)$, $p \in [0,1]$, is the inverse of the cumulative Gaussian distribution [10]. Performance in these signal detection tasks usually results in psychometric curves that have sigmoidal shapes [10]. We thus defined the lowest sustainable SF as the inflection point of those psychometric curves in both conditions. To test whether perception was significantly more robust against temporal subsampling in the feature condition, for each participant, we fitted a sigmoidal curve to the performance in both conditions and calculated its inflection point (in Hz). Inflection points were then compared across conditions using Student's t-test to test whether the robustness of auditory perception against temporal subsampling differs between the two conditions.

Results

In this study, participants were presented with speech stimuli that were subsampled (at different temporal SF) either at the input level (input condition) or at the level of auditory features (feature condition). Using a two-back task (see Methods), we tested the robustness of the auditory system to this subsampling – if perceptual cycles do exist in the auditory system, they can only occur at frequencies that prove to be robust against temporal subsampling. Of course, positive results would not imply that they actually do occur, but our approach can inform us whether sampling on a hierarchically high level of auditory processing (i.e. in the feature condition) can reduce the detrimental effects of environmental sampling (i.e. loss of information) and thus 'keep alive' the notion of perceptual cycles in the auditory system.

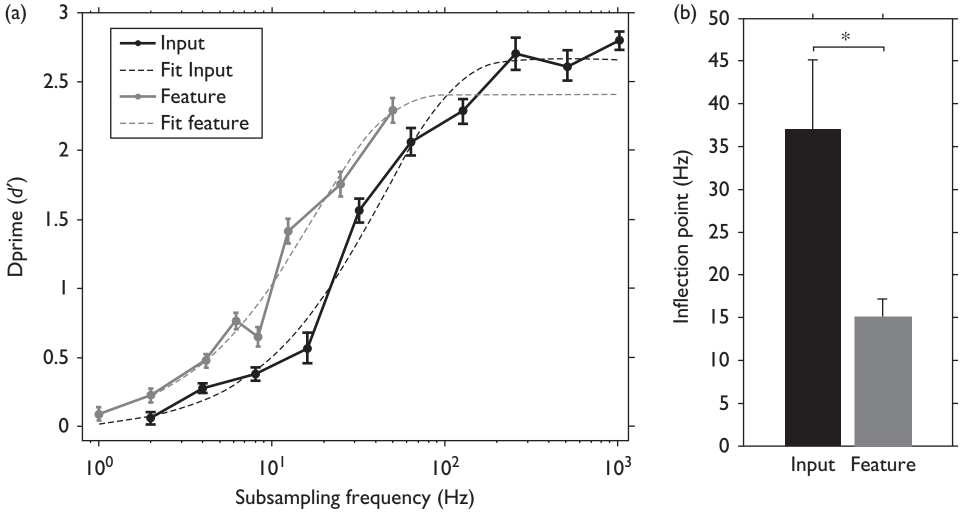
The performance (measured in d'; see Methods) of our participants ($N=7$) in the two-back task for both conditions is shown in Fig. 2. For both conditions, of course, performance increased with increasing SF, and both curves resemble sigmoid psychometric functions. However, this curve is shifted toward lower SF for the feature condition, indicating that the auditory system is more robust against subsampling at the level of auditory features than at the input level (Fig. 2a). When we define the lowest sustainable SF as the inflection point of those psychometric curves under both conditions (Fig. 2b; see Methods), this SF is significantly lower [$t(6)=2.61$, $P=0.023$] for the feature condition (15.1 ± 5.3 Hz; mean and SD across participants) than for the time condition (37.0 ± 21.5 Hz). Thus, subsampling on a hierarchically higher level of auditory processing would occur with a clear advantage for the auditory system as the SF could be reduced without a significant loss of information.

Discussion

In a previous study, we showed that subsampling the auditory stream in the input domain has detrimental effects on stimulus processing already at SF below 32 Hz [1]. This result suggests that, if the auditory environment is indeed monitored in a discrete manner, this sampling cannot take place on the very input to the system as important information would be lost (or the SF would have to be so high that subsampling would be useless). Instead of rejecting the idea of perceptual cycles in audition, we proposed an alternative idea, which is supported by experimental findings in the present study: subsampling in the auditory system is less detrimental to performance if it takes place on a hierarchically higher level of auditory processing – on the level of auditory features. Our results are in line with the work published by Suied *et al.* [11]: although they used short vocal sounds instead of speech sound, they were able to show that participants could still recognize those sounds even though they were reduced to a few perceptually important features, with the number of features (10 features/s) similar to our perceptual 'threshold'. We extend their findings by systematically testing different SF and by comparing perceptual consequences for subsampling at the level of the cochlea with those for a hierarchically higher stage of auditory processing.

Whereas there is plenty of (psychophysical and electrophysiological) evidence for perceptual cycles in vision [12–15], equivalent experiments consistently fail for the auditory domain ([2,3]; reviewed in VanRullen *et al.* [1]) or remain debated [16–19]. Thus, it is possible that the reported experimental approaches for the visual system are not appropriate for the investigation of the auditory system. Our study directly contributes toward resolving this discrepancy by providing an answer to why this could be the case: perceptual cycles might operate on different hierarchical levels in the two systems and thus might not be captured with the same experimental methods (although this does not necessarily mean that the visual system is not influenced by high-level factors; for instance, some evidence for perceptual cycles in vision is modulated by attention [15]). The visual world is relatively stable over time and subsampling does not disrupt essential information, even when the very input to the system is discretized. This was verified in our previous study [1], where, in a two-back recognition paradigm similar to the one used here, human observers could robustly recognize visual inputs at subsampling rates below 5 Hz. In contrast, the fluctuating nature of auditory stimuli makes it necessary to extract features that are

Fig. 2



Auditory perception is more robust against temporal subsampling at the level of auditory features than when the same subsampling is applied at the input level. (a) Performance in a two-back task in the input (black) and feature condition (gray). Note that the sigmoidal curve (a fit is shown using dashed lines) is shifted toward lower SF for the feature condition, indicating that, for a given SF, performance was better (and perception more robust) in the latter condition. (b) Inflection points for both conditions, averaged across sigmoidal curves fitted on the data of individual participants. The inflection point is significantly higher for the input condition than for the feature condition, indicating that subsampling has more detrimental effects when it is performed at the level of the cochlea (i.e. at the input level) than at a hierarchically higher level of auditory processing (i.e. in the feature domain). SEM across participants is shown by error bars.

both relevant and more stable before subsampling can be applied (see Results [1,20]). Indeed, whereas early auditory representations of speech seem to entail all acoustic details, representations at later hierarchical stages are rather categorical (i.e. relatively independent of acoustic information) and thus more stable in time [21], and therefore, more robust to subsampling. This property of an increasing abstraction of auditory representation along the pathway begins beyond the primary auditory cortex and is particularly outstanding in the anterior temporal cortex (ventral stream) [21], making it, although speculatively, a good candidate for perceptual cycles in the auditory system: 'auditory objects' are 'built' within this stream [22], transforming spectrotemporal (i.e. time-resolved) properties of the input stream into more abstract 'identities' (i.e. relatively independent of the time domain). The superior temporal sulcus – in which certain neurons respond more strongly to speech than to other sounds [23] – is part of this stream.

Auditory perception in our study did not prove as robust to temporal subsampling as observed previously for vision. This does not necessarily imply that auditory perceptual cycles, if they exist, must be faster than visual ones. Instead, it may just be that our 'feature' decomposition did not fully capture the complexity of the auditory representation at which subsampling occurs. The higher in the hierarchy of the auditory pathway, the slower are the 'preferred' frequencies of the auditory system [24]. It thus remains to be shown in future studies whether subsampling an even more complex decomposition of the auditory signal can result in performance that equals that obtained in vision.

More studies are necessary to find an appropriate experimental approach for perceptual cycles in audition and to characterize them with respect to their location in the auditory pathway. One step forward toward auditory perceptual cycles has been published recently by our group [25]: in that study, specifically constructed noise was mixed with speech sound to counterbalance fluctuations in low-level features of the latter (i.e. fluctuations in amplitude and spectral content). Importantly, these mixture speech/noise stimuli remained intelligible, indicating that high-level features of speech (including phonetic information) were preserved and fluctuated rhythmically. We could show that the detection of tone pips is modulated by this 'high-level rhythm' and, consequently, that phase entrainment, the brain's adjustment to regular stimulation, indeed involves a high-level component. This finding is in line with the present study, suggesting a periodic mechanism on a high level of auditory processing.

To conclude, our data suggest that, even in the auditory world of continuous, rapid temporal fluctuations, the idea of discrete perceptual processing can be kept alive: discretization on a hierarchically high level of auditory processing is possible without disrupting essential information. Of course, our experiment does not prove that perceptual cycles do exist in audition; however, we conclude that (a) there is a possibility that they exist and (b) if so, they are a high-level phenomenon.

Acknowledgements

The authors are grateful to Daniel Pressnitzer for helpful comments and discussions. This study was supported by a Studienstiftung des deutschen Volkes (German National Academic Foundation) scholarship to B.Z. and an ERC grant P-CYCLES number 614244 to R.V.

Conflicts of interest

There are no conflicts of interest.

References

1 VanRullen R, Zoefel B, Ilhan B. On the cyclic nature of perception in vision versus audition. *Philos Trans R Soc Lond B Biol Sci* 2014; **369**:20130214.

2 İlhan B, VanRullen R. No counterpart of visual perceptual echoes in the auditory system. *PloS One* 2012; **7**:e49287.

3 Zoefel B, Heil P. Detection of near-threshold sounds is independent of EEG phase in common frequency bands. *Front Psychol* 2013; **4**:262.

4 Kubovy M. Should we resist the seductiveness of the space:time::vision: audition analogy? *J Exp Psychol Hum Percept Perform* 1988; **14**:318–320.

5 Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Commun* 2010; **52**:12–40.

6 Chung K, McKibben N. Microphone directionality, pre-emphasis filter, and wind noise in cochlear implants. *J Am Acad Audiol* 2011; **22**:586–600.

7 Levinson N. The Wiener RMS error criterion in filter design and prediction. *J Math Phys* 1947; **25**:261–278.

8 Durbin J. The fitting of time series models. *Rev Inst Int Stat* 1960; **28**:233–243.

9 Soon IY, Koh SN, Yeo CK. Noisy speech enhancement using discrete cosine transform. *Speech Commun* 1998; **24**:249–257.

10 Macmillan NA, Creelman CD. *Detection theory: a user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2004.

11 Suied C, Drémeau A, Pressnitzer D, Daudet L. Auditory sketches: sparse representations of sounds based on perceptual models. In: Aramaki M, Barthet M, Kronland-Martinet R, Ystad S, editors. *From sounds to music and emotions*. Berlin Heidelberg: Springer; 2013. pp. 154–170.

12 VanRullen R, Reddy L, Koch C. The continuous wagon wheel illusion is associated with changes in electroencephalogram power at approximately 13 Hz. *J Neurosci* 2006; **26**:502–507.

13 Busch NA, Dubois J, VanRullen R. The phase of ongoing EEG oscillations predicts visual perception. *J Neurosci* 2009; **29**:7869–7876.

14 Mathewson KE, Gratton G, Fabiani M, Beck DM, Ro T. To see or not to see: prestimulus alpha phase predicts visual awareness. *J Neurosci* 2009; **29**:2725–2732.

15 Busch NA, VanRullen R. Spontaneous EEG oscillations reveal periodic sampling of visual attention. *Proc Natl Acad Sci U S A* 2010; **107**:16048–16053.

16 Henry MJ, Herrmann B. A precluding role of low-frequency oscillations for auditory perception in a continuous processing mode. *J Neurosci* 2012; **32**:17525–17527.

17 Henry MJ, Obleser J. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc Natl Acad Sci U S A* 2012; **109**:20095–20100.

18 Ng BSW, Schroeder T, Kayser C. A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *J Neurosci* 2012; **32**:12268–12276.

19 Vanrullen R, McLelland D. What goes up must come down: EEG phase modulates auditory perception in both directions. *Front Psychol* 2013; **4**:16.

20 Thorne JD, Debener S. Look now and hear what's coming: on the functional role of cross-modal phase reset. *Hear Res* 2014; **307**:144–152.

21 Davis MH, Johnsrude IS. Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hear Res* 2007; **229**:132–147.

22 Bizley JK, Cohen YE. The what, where and how of auditory-object perception. *Nat Rev Neurosci* 2013; **14**:693–707.

23 Overath T, McDermott JH, Zarate JM, Poeppel D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* 2015; **18**:903–911.

24 Edwards E, Chang EF. Syllabic (~2-5 Hz) and fluctuation (~1-10 Hz) ranges in speech and auditory processing. *Hear Res* 2013; **305**:113–134.

25 Zoefel B, VanRullen R. Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. *J Neurosci* 2015; **35**:1954–1964.