Neuron, Volume 61

Supplemental Data

Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes

Mounya Elhilali, Ling Ma, Christophe Micheyl,
Andrew J. Oxenham, Shihab A. Shamma

Physiological cortical responses of alternating and synchronous two-tone sequences.

Supplementary Figures 1 and 2 show an example of a single unit in primary auditory cortex, in responses to an alternating (synchronous) sequence as defined in experiment 1.

Model simulations with Complex Stimuli

The model simulations are based on the analysis of temporal response coherence across tonotopic channels can correctly account for the percepts evoked by synchronous as well as non-synchronous tone sequences. This already represents a substantial improvement on earlier models of auditory streaming, which would not correctly predict that tones that are widely separated along the tonotopic dimension are nevertheless grouped perceptually into a single stream if they are synchronous. However, the guiding principle of temporal coherence across neural channels is far more general than that of tonotopic separation, and can be applied to more complex and more natural sound scenes than two-tone sequences. Examples of application of the model to more complex sounds are presented in Supplemental Figures 3 and 4. In these examples, we focus on two

particularly interesting types of spectrally and temporally complex but well controlled stimuli, which have been used in studies of two very important aspects of auditory perception: informational masking and speech recognition.

Informational masking refers to an inability to detect sounds that are not masked in the auditory periphery, but are rendered inconspicuous by the presence of concomitant sounds, which contain more salient features such as fast spectral changes. This phenomenon, which has been the object of increasing research in the past ten years, is currently thought to play an important role in everyday life, where sounds of interest (e.g., speech or music) are often accompanied by other, distracting sounds. In the laboratory, the phenomenon can be demonstrated using a sequence of isochronous (i.e., regularly repeating) "target" tones, which the listener must detect amid a cloud of randomly varying, dispersed and desynchronized "distractor" tones, as is illustrated in Supplementary Figure 3.

The stimuli used in this simulation followed the design used by (Micheyl, 2007). Specifically, the stimuli consisted of 40 multi-tone bursts. Each burst was 100 ms in duration (including 20-ms onset and offset ramps). The total stimulus duration was 4 seconds. Each burst consisted of 5 simultaneous tone masker components, with frequencies randomly chosen from a list of 16 frequencies spaced 2 semitones (about 12%) apart. A target frequency was chosen at 1050Hz, and was presented at every other burst. The relative maskers-to-target level was fixed at 0dB SPL. A protection zone around this target frequency was chosen so that no masker tone could exist in that spectral band. The half-width of this protection zone was varied from 2 to 12 semitones.

As shown in Supplemental Figure 3, the target sequence is separated from the distracters by a range of "silent" channels or "protected zone", which serves a role analogous to the frequency separation (ΔF) between the alternating tones in the simpler auditory streaming paradigm illustrated in Fig. 1: if the protection zone is wide enough, the regularly repeating target tones form a separate stream, which emerges perceptually from the randomly varying tonal background (Micheyl, 2007). This stimulus has been widely used as a simplified simulation of the detection of sound sources in complex, spectro-temporally varying acoustic environments (Durlach et al., 2003a; Durlach et al., 2003b; Kidd et al., 2002; Kidd et al., 2003; Kidd et al., 1998; Oxenham et al., 2003), and it was recently used to investigate neural correlates of sound awareness in auditory cortex (Gutschalk et al., 2008). In the absence of detailed physiological data on AI responses to such "informational masking" stimuli, we simulated cortical responses to the target and distracter tones by constructing a "cortical" channel array in which each channel integrated inputs from a moderate range of frequencies (ranging from ¼ octaves to 2 octaves in steps of ¼ octaves), which represents a population of neurons with average bandwidth of 1.125 octaves, reflecting the average bandwidth of cortical cells (as in the tuning curve of Fig.5A). We then analyzed the coherence of responses among the channels, exactly as described earlier for the two-tone stimuli. The simulation results are shown in Supplementary Fig. 3. Consistent with psychophysical data in the literature (Micheyl, 2007), the model predicted that the likelihood that the target tones were heard as a separate stream increased with the width of the protected zone, from a 50% or less predicted chance of perceived segregation at the $1/6^{th}$-octave width, to nearly a 90% chance at the one-octave width. This behavior of the model results from a decrease in the

temporal coherence of cortical responses in target and non-target channels with increasing frequency separation between the target and masker tones, due to decreasing overlap between the broadly-tuned cortical channels.

Temporal response coherence across auditory (cortical) channels can also explain some essential aspects of the perception of a totally different type of stimuli known as "sinewave speech". Sinewave speech stimuli are produced by summing a few (usually, two or three) pure tones, the frequencies and amplitudes of which are independently modulated in such a way that they follow the frequencies and amplitudes of spectral peaks (or "formants") in a natural speech segment (e.g., a sentence) (Remez et al., 1981). Naïve listeners presented for the first time with such stimuli usually fail to identify them as speech; they report having heard simultaneous whistles, indicating that the tones that make up the stimulus are heard as separate streams. After being told that the sounds can be heard a speech, and with some practice, listeners usually become able to better understand sinewave speech; however, the percept seems to remain one of separate streams. According to the "coherence" hypothesis, the reason why sinewave speech stimuli are not perceptually integrated into a coherent auditory stream is that the tones that make it up are modulated *independently* in frequency and amplitude, and thus evoke temporally *incoherent* responses across auditory channels along the tonotopic axis. Based on this hypothesis, it is predicted that that *coherent* amplitude modulation of the tones in sinewave speech stimuli should promote their perceptual grouping. Indirect evidence for this is provided by results in the psychoacoustics literature, which demonstrate that sinewave speech is more easily intelligible when the tones are modulated coherently in amplitude compared to when they are incoherently modulated (Barker and Cooke, 1990;

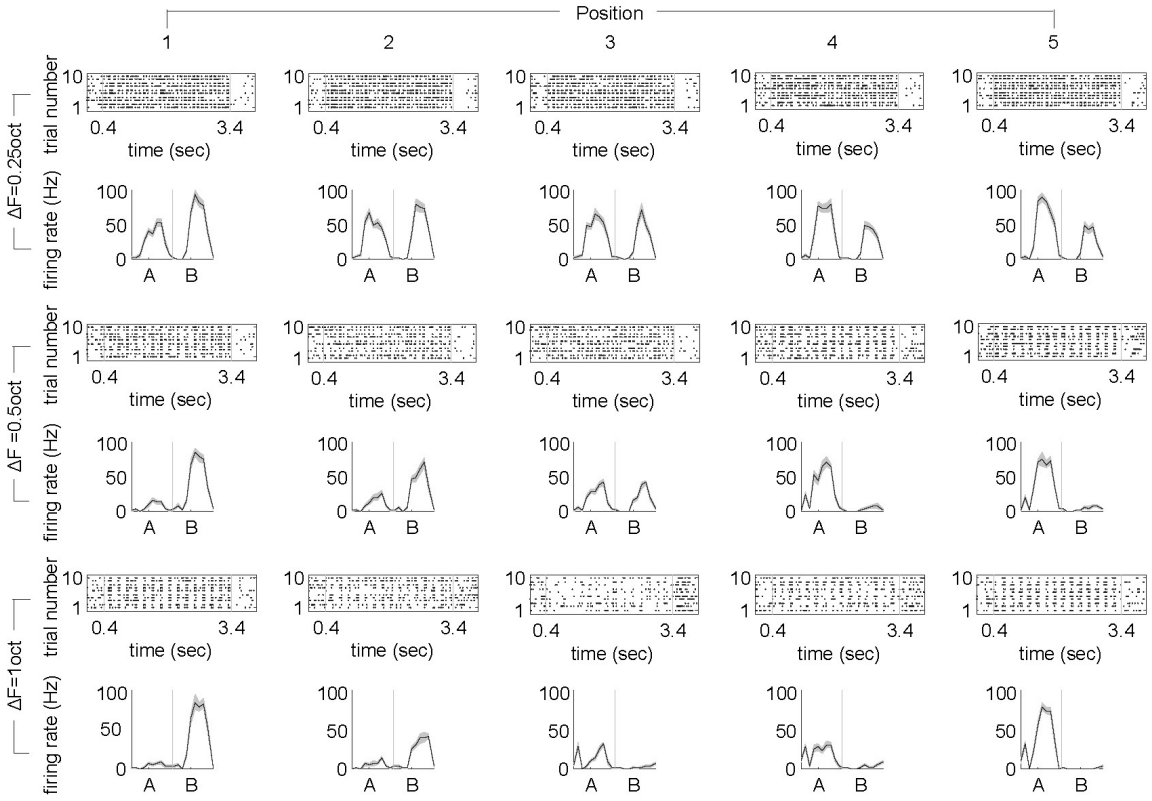
Carrell and Opie, 1992). Supplementary Fig. 4 shows simulation results obtained by using sinewave speech stimuli (constructed from 2 sinewaves following the first and second formants). In this case, the model's input consisted of two sinusoidal waveforms tracking the first and second formant of a speech utterance. The pulsed sinewave was generated by modulating the speech utterance using square modulations at a rate of 10Hz. The model performs a temporal coherence analysis across channels, and yields a ratio of $\lambda 2/\lambda 1 = 0.89$ for sinewave speech (Suppl. Fig 4A), which can explain the lack of perceived coherence in sinewave speech. Alternatively, when channels are coherently modulated (by gating the signal with a square wave at 10 Hz), the model predicts increased perceptual grouping as indicated by the singular value ratio of $\lambda 2/\lambda 1 = 0.5$ (Suppl. Fig 4B).

Notes on coherence and dimensionality reduction

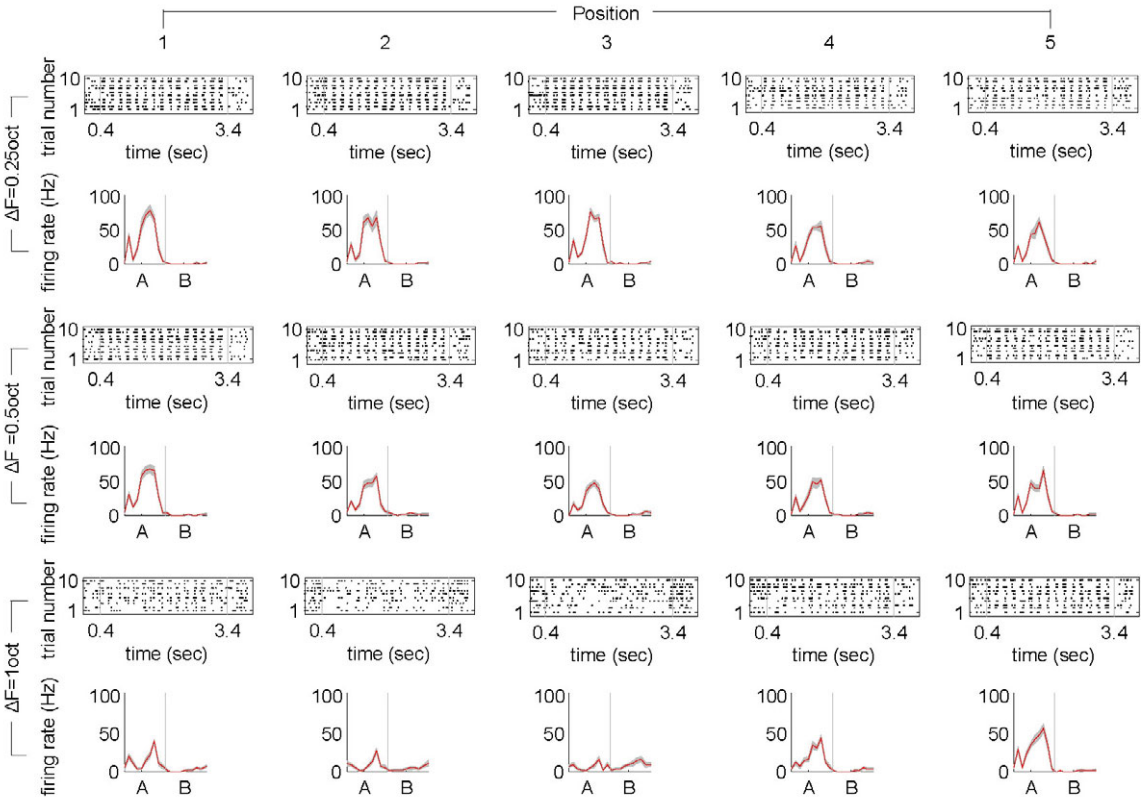
The computational model described here is based on the idea that temporal coherence across auditory channels promotes perceptual grouping. This is realized in the model by computing the short-term correlation between all pairs of channel, and grouping together coherent channels. This procedure is closely related to the classic problem of feature reduction, in which redundant channels (or features) are eliminated (Duda et al., 2000). The singular-value-decomposition implementation that was used here is only one of many possible forms (Bau and Trefethen, 1997; Golub and Van Loan, 1996; Press et al., 1992), some of which are more biologically plausible than others. In fact, this coherence model can also be recast as a pattern clustering problem that does not even require explicit measurements of coherence (Elhilali and Shamma, 2008).

Supplementary Figure Legends

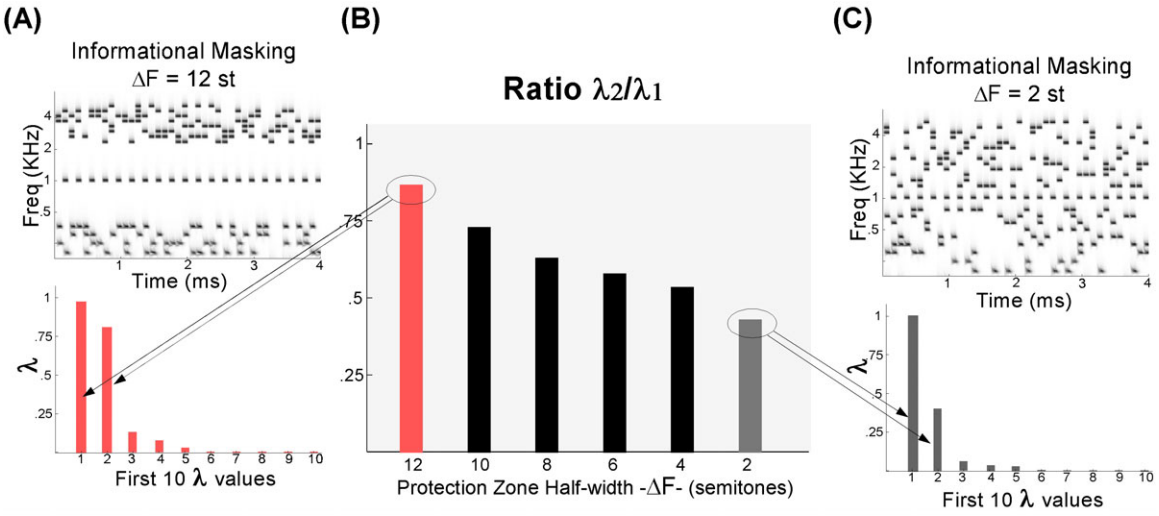
Supplementary Figure 1. Single unit example for alternating sequence in experiment 1. Raster and period histogram in two-tone alternating mode for all conditions (3 ΔF x 5 positions). Each condition has 10 repetitions. Each trial includes 0.4 second pre-stimulus silence, 3 second two-tone sequences, and 0.6 second post-stimulus silence. Grey lines in raster indicate stimuli onset and offset. Grey area in period histogram indicates standard error.



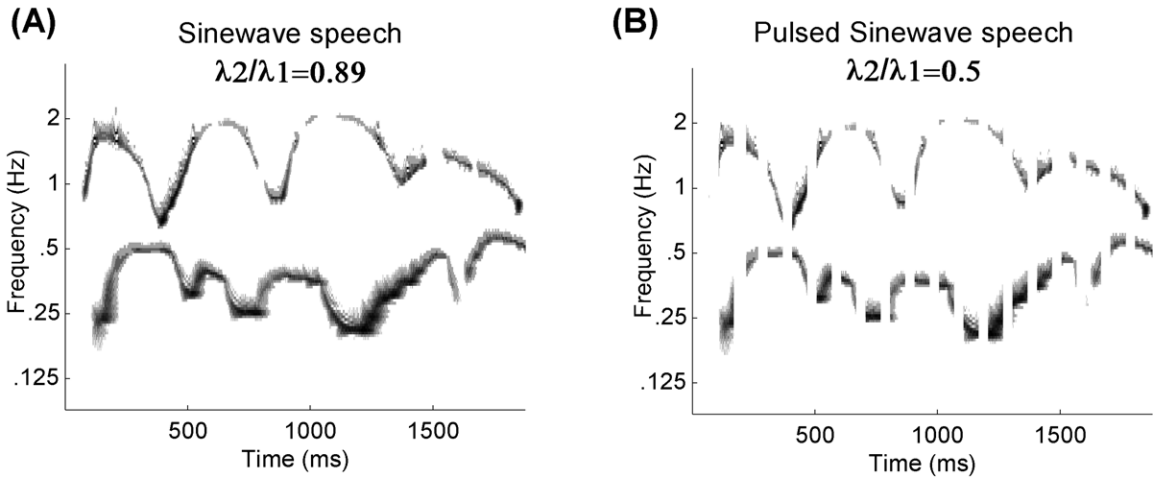
Supplementary Figure 2. Single unit example for synchronous sequence in experiment 1. Raster and period histogram in two tone synchronous mode for all conditions (3 ΔF x 5 positions). Each condition has 10 repetitions. Each trial includes 0.4 second pre-stimulus silence, 3 second two-tone sequences, and 0.6 second post-stimulus silence. Grey lines in raster indicate stimuli onset and offset. Grey area in period histogram indicates standard error.



Supplementary Figure 3. Model simulation for informational-masking stimuli. **(A)** A stimulus consisting of masker tones and repeating target tone with 12 semitones protection zone is presented to the model. The decomposition of the coherence matrix reveals two main singular values (lower panel). **(B)** The ratio of λ2/λ1 is shown as we vary the protection zone around the target tone from 12 to 2 semitones. As the protection zone increases, the ratio λ2/λ1 decreases, hence the likelihood of grouping streams present in the stimulus increases. **(C)** The informational masking stimulus is shown for a 2 semitone protection zone. The lower panel reveals the singular values of the coherence matrix C. λ2 is considerably smaller than the 12 semitone case (shown in panel A), indicated an increased likelihood of grouping the maskers and target streams.



Supplementary Figure 4. Model simulation using sinewave speech. **(A)** A sinewave utterance is generated using two sinusoidal waves following the first and second formant of a normal speech sentence. The coherence analysis and subsequent singular value decomposition yields an eigen-value ratio $\lambda 2/\lambda 1 = 0.89$. **(B)** The same utterance is pulsed at a rate of 10Hz, and yields a singular-value ratio of $\lambda 2/\lambda 1 = 0.5$.



Supplementary References

Barker, J., and Cooke, M. (1990). Is the sine-wave speech cocktail party worth attending? Speech Communication *27*, 159-174.

Bau, D., and Trefethen, L.N. (1997). Numerical linear algebra (Philadelphia: SIAM: Society for Industrial and Applied Mathematics).

Carrell, T.D., and Opie, J.M. (1992). The effect of amplitude comodulation on auditory object formation in sentence perception. Percept Psychophys *52*, 437-445.

Duda, R., Hart, P., and Stork, D.G. (2000). Pattern Classification and Scene Analysis., 2 edn (John Wiley and Sons.).

Durlach, N.I., Mason, C.R., Kidd, G., Jr., Arbogast, T.L., Colburn, H.S., and Shinn-Cunningham, B.G. (2003a). Note on informational masking. J Acoust Soc Am *113*, 2984-2987.

Durlach, N.I., Mason, C.R., Shinn-Cunningham, B.G., Arbogast, T.L., Colburn, H.S., and Kidd, G., Jr. (2003b). Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. J Acoust Soc Am *114*, 368-379.

Elhilali, M., and Shamma, S.A. (2008). A cocktail Party Problem - with a Cortical Twist: How cortical mechanisms contribute to sound segregation. Under review in J Acoust Soc Am.

Golub, G.H., and Van Loan, C.F. (1996). Matrix Computations, 3 edn (Baltimore, MD: Johns Hopkins University Press).

Kidd, G., Jr., Mason, C.R., and Arbogast, T.L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. J Acoust Soc Am *111*, 1367-1376.

Kidd, G., Jr., Mason, C.R., and Richards, V.M. (2003). Multiple bursts, multiple looks, and stream coherence in the release from informational masking. J Acoust Soc Am *114*, 2835-2845.

Kidd, G., Jr., Mason, C.R., Rohtla, T.L., and Deliwala, P.S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. J Acoust Soc Am *104*, 422-431.

Micheyl, C., Shamma, S, Oxenham, AJ (2007). Hearing out repeating elements in randomly varying multitone sequences: a case of streaming? In Hearing - from basic research to applications., B. Kollmeier, Klump, G, Hohmann, V, Langemann, U, Mauermann, M, Uppenkamp, S, Verhey, J, ed. (Berlin: Springer), p. in press.

Oxenham, A.J., Fligor, B.J., Mason, C.R., and Kidd, G., Jr. (2003). Informational masking and musical training. J Acoust Soc Am *114*, 1543-1549.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1992). Numerical recipes: The art of scientific computing (Cambridge University Press).

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. Science *212*, 947-949.