Abstract

Human speech and vocalizations in animals are rich in joint spectrotemporal (S-T) modulations, wherein acoustic changes in both frequency and time are functionally related. In principle, the primate auditory system could process these complex dynamic sounds based on either an inseparable representation of S-T features, or alternatively, a separable representation. The separability hypothesis implies an independent processing of spectral and temporal modulations. We collected comparative data on the S-T hearing sensitivity in humans and macaque monkeys to a wide range of broadband dynamic spectrotemporal ripple stimuli employing a yes-no signal-detection task. Ripples were systematically varied—as a function of density (spectral modulation-frequency), velocity (temporal modulation-frequency), or modulation depth—to cover a listener's full S-T modulation sensitivity; derived from a total of 87 psychometric ripple detection curves. Audiograms were measured to control for normal hearing. Determined were hearing thresholds, reaction time distributions, and S-T modulation transfer functions (MTFs); both at the ripple detection thresholds, and at supra-threshold modulation depths. Our psychophysically derived MTFs are consistent with the hypothesis that both monkeys and humans employ analogous perceptual strategies: S-T acoustic information is primarily processed separable. Singular-value decomposition (SVD), however, revealed a small but consistent, inseparable spectral-temporal interaction. Finally, SVD analysis of the known visual spatiotemporal contrast-sensitivity function (CSF) highlights that human vision is space-time inseparable to a much larger extent than is the case for S-T sensitivity in hearing. Thus, the specificity with which the primate brain encodes natural sounds appears to be less strict than is required to adequately deal with natural images.

NEW & NOTEWORTHY We provide comparative data on primate audition of naturalistic sounds comprising hearing thresholds, reaction time distributions, and spectral-temporal modulation transfer functions. Our psychophysical experiments demonstrate that auditory information is primarily processed in a spectral-temporal independent manner by both monkeys and humans. Singular-value-decomposition of known visual spatiotemporal contrast-sensitivity—in comparison to our auditory spectral-temporal sensitivity—revealed a striking contrast in how the brain encodes natural sounds as opposed to natural images, as vision appears to be space-time inseparable.

naturalistic sounds; primate audition; psychophysics; spectrotemporal modulation transfer functions; spectrum-time separability

INTRODUCTION

Biological sounds are characterized by statistical regularities in their dynamic spectral modulations, in which the frequency content changes over time. The ability to faithfully encode spectrotemporal (S-T) modulations is not only important for sound recognition, but also for sound segregation in environmental noise—like listening to a conversation at a cocktail party (1-4). Similar problems arise when animals attempt to distinguish mating or echolocating calls from ambient noises (5, 6). Examples include species-specific communication signals in animals as diverse as mammals, birds, amphibians, reptiles and insects (7-10). The auditory system faces the challenge to distinguish sounds based on their S-T modulation content. In particular, humans rely on the speed and direction of covarying S-T amplitude modulations to derive meaning from spoken words (4, 11).

Neurophysiological experiments in macaques implicate an ancient cortical system processing S-T modulations (12-16). The mechanisms by which monkeys process vocalizations could also extend to humans (17-22). With this comparative hypothesis in mind, we exposed humans and monkeys to a wide range of dynamic S-T ripples to characterize their S-T perceptual abilities (Fig. 1). Ripples (*Eqs. 1-2*) are naturalistic broadband signals with inseparable spectral and temporal modulations (Fig. 1*A*). They form a two-dimensional Fourier basis for sound, whereby any acoustic pattern can be composed by the superposition of a particular set of ripples (23, 24). Their importance in hearing research lies in the parametric assessment of auditory processing of complex sounds. Ripples have proven their audiological value as parametric non-speech stimuli, responses to which are predictive for speech perception (25-27). Moreover, measuring auditory-evoked responses to ripples—at either perceptual or neurophysiological level—allows assessment of S-T (in)separability of, or within, the auditory system.

--- Figure 1 about here ---

Separable, or alternatively, inseparable S-T sensitivity can be determined through singular value decomposition (SVD) analysis of the two-dimensional S-T modulation transfer function (MTF; Fig. 1*C*) encompassing the product of a time-dependent—temporal modulation: velocity *ω (in Hz)*—and a frequency-dependent—spectral modulation: density *Ω (cycles/octave, or c/o)*—transfer function (*Eq. 7*). Separable S-T sensitivity is characterized by the inseparability index α_{SVD} (*Eq. 8*) equaling zero and the SVD MTF correlation coefficient r^2_{SVD} equaling unity, when separability is complete (see Fig. 2 for explanation, left-hand column). In this case spectral and temporal modulations are processed independently.

In contrast, inseparable S-T sensitivity is characterized by $\alpha_{SVD} > 0$ and $r^2_{SVD} < 1$ (Fig. 2, right-hand column), highlighting that spectral and temporal modulations are processed dependently to some extent. Finally, S-T sensitivity can be biased to a particular ripple movement direction—upward vs. downward S-T modulations—in which case the MTF sensitivity distribution is asymmetric along the horizontal-dimension and could give rise to a $r^2_{up/down} < 1$ (Fig. 2, bottom row).

Quantitative analysis of S-T receptive fields (STRFs) of auditory neurons has demonstrated an increased proportion of neurons with inseparable STRFs ranging from midbrain *inferior colliculus* to primary auditory cortex (13, 23, 24, 28-37). While it is evident that separable and inseparable S-T encodings are manifest at different processing stages within the auditory pathway, it is not straightforward to predict what happens at the perceptual level. Psychophysical measurements in humans (38)—assigning detection thresholds to a wide range of dynamic ripples—are consistent with an up/down symmetric, separable processing model (top-left, Fig. 2). In this special case, the perceptual MTF is mirror-symmetric around the zero-density axis and oriented orthogonal to the spectral modulation axis.

Given S-T separability of human hearing at threshold (38, 39), it is perhaps surprising to learn that the region with highest sensitivity is not optimized to the S-T modulations that dominate speech (4, 11, 12). Likewise, zebra finches show ripple detection thresholds (40) that do not correspond to the dominant modulation spectra of their own vocalization calls (37, 40). This is unexpected, since the forebrain of songbirds appears to be specialized for processing vocalizations (41).

--- Figure 2 about here ---

Two hypotheses could explain these apparent discrepancies. First, preferential sensitivity to conspecific vocalizations may not be evident at the modulation detection threshold, as intelligible vocalizations are typically produced well above threshold (42). If so, supra-threshold MTFs could mirror the asymmetric nature of the S-T decompositions of e.g., English speech (*'intelligible'*, Fig. 1*B*), wherein the strongest modulations are downward moving (38). Supra-threshold S-T hearing is then asymmetric, resembling the S-T sensitivity pattern of the right-hand panels in Fig. 2. Alternatively, the processing of S-T modulations may be based on information efficiency principles (43, 44), instead of neuro-ethological ones (40). In this case, increased S-T sensitivity for vocalizations over other classes of biological sounds and perceptual levels is no longer expected and may give rise to a separable and symmetric MTF, also for supra-threshold sounds (top-left, Fig. 2). To dissociate between