Comparative Primate Audition

Spectral-temporal processing of naturalistic sounds in monkeys and humans

Robert F. van der Willigen\textsuperscript{1,2,3,*}, Huib Versnel\textsuperscript{1,4}, A. John van Opstal\textsuperscript{1}

Running Title: Comparative Primate Audition

* To whom correspondence should be addressed.

E-mail: r.f.van.der.willigen@hr.nl

\textsuperscript{1}Department of Biophysics, Donders Institute for Brain, Cognition and Behaviour; Radboud University, Nijmegen, The Netherlands

\textsuperscript{2}School of Communication, Media and Information Technology, Rotterdam University of Applied Sciences, Rotterdam, The Netherlands

\textsuperscript{3}Research Center Creating 010, Rotterdam University of Applied Sciences, Rotterdam, The Netherlands

\textsuperscript{4}Department of Otorhinolaryngology and Head & Neck Surgery, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
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.

Keywords: 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. 1A). 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. Dynamic rippled-noise parameters and threshold S-T MTF psychophysics.**

A: The 5x5 stimulus grid represents a subset of 25 ripple spectrograms with varying temporal (vertical axis) and spectral (horizontal axis) modulation rates. Three parameters define the amplitude envelope of each spectrogram: (1) Velocity, $\omega$ [Hz]: temporal modulation; (2) Density, $\Omega$ [c/o]: spectral modulation; (3) Modulation depth, $\Delta M$ [%] (linear scale). The $\omega/\Omega$ ratio specifies (4) upward ($< 0$) or downward ($> 0$) direction of spectral motion. $\Omega = 0$: pure temporal modulations; $\omega = 0$: pure spectral modulations. The (0,0) stimulus has a flat spectrogram representing static noise.

B: Ripple-onset detection performance as function of $\Delta M$, for the encircled downward ripple in (A). Detection threshold (Eq. 5) is defined at the half-point of the fitted (Eq. 4) psychometric curve (black line). The parameter $\alpha$ specifies the detection threshold response criterion—defining the function’s relative position along the x-axis. The guess rate ($\gamma$) and lapse rate ($\lambda$) specify the lower—close to 0%—and upper—close to 100%—bounds of the function, respectively. The here estimated threshold level (arrow) equates to a modulation depth of ~28%. C: The 87 ripples thus yield the threshold modulation transfer function (MTF) as function of $(\Omega, \omega)$, here shown for human listener h1. Grey scale represents $\Delta M$ threshold levels, which is not defined for the (0,0) sound (hatched). Inner white rectangle circumscribes stimulus grid shown in (A). The black outlined square at (8,1,2) represents the detection threshold as obtained from the fitted curve shown in (B).

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. 1C) encompassing the product of a time-dependent—temporal modulation: velocity $\omega$ (in Hz)—and a frequency-dependent—spectral
Comparative Primate Audition

modulation: density $\Omega$ (cycles/octave, or c/o)—transfer function (Eq. 7). Separable S-T sensitivity is characterized by the inseparability index $a_{SVD}$ (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 $a_{SVD} > 0$ 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 $r^2_{up/down} < 1$ (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).
Comparative Primate Audition

Figure 2. Separable vs. inseparable characterization of S-T hearing.

Four theoretical sensitivity S-T MTF matrices (multi-colored panels), equivalent to the gray-scaled one shown in Fig. 1C. Red colors specify low ripple sensitivity, blue colors, high sensitivity. MTFs are categorized according to up/down symmetry—symmetric (top row) vs. asymmetric (bottom row)—, or extent of separability—separable (left column) vs. inseparable (right column)—. Note, the upper MTFs are mirror-symmetric around the zero-density (white vertical) axis and oriented orthogonal to the spectral modulation (horizontal) axis. The grey insets provide quantitative analysis of both the extent of inseparability and the up/down symmetry of the four displayed MTFs. Here, $a_{SVD}$ reflects the degree of inseparability, with zero corresponding to full separability across the entire S-T domain. The $r^2_{SVD}$ statistic reflects the proportion of variance accounted for when assuming separability. Fully separable means: only the 1st eigenvalue—$\lambda_1$—is significant (Eq. 7). Inseparable means: more than one eigenvalue is significant. The number of blue bars in each grey inset represent the number of significant eigenvalues needed to fully reconstruct their respective MTFs. The length of each bar is a measure of magnitude [a.u.]. The $r^2_{up/down}$ statistic reflects perfect symmetry when equaling unity, and the gain, $g$, of the relation, $M(\Omega < 0) = g \times M(\Omega > 0)$, equals one. Note, the highly asymmetric up/down MTFs (lower row) display a sensitivity biased towards downward moving ripple sounds, with a gain of 2.5.

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. 1B), 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 preferential and non-preferential sensitivity to naturalistic S-T modulations, and to enable a direct comparison between species, we studied five humans and five monkeys responding to a wide range of ripples under identical psychophysical conditions, and we determined their S-T sensitivities at threshold and supra-threshold levels.

Our results demonstrate that monkeys and humans share a largely unbiased up/down perceptual strategy—based on separable sensitivities to spectral and temporal amplitude modulations—when processing inseparable sounds. However, our analysis also indicated a small but significant, contribution of inseparability to the S-T sensitivity of both species. To conclude, we also demonstrate—by means of SVD-analysis of the known visual spatiotemporal CSF (45)—that human vision is predominantly governed by inseparable processing of naturalistic stimuli.
MATERIALS AND METHODS

Ethics Statement

Our tests were purely behavioral and involved no distress or discomfort to our human volunteers or our monkeys. All experimental procedures complied with the European Communities Council Directive of September 22, 2010 (2010/63/EU). The local ethics committee for the use of laboratory animals (DEC) of the Radboud University reviewed and approved all experimental protocols. To ensure the animals’ health and welfare, their general appearance was monitored daily and recorded in a welfare diary, along with their daily food and fluid intake.

Human psychophysics on five healthy volunteers was performed after they had been informed about the behavioral procedures and their consent was taken. Protocols were approved by the local ethics committee of the Faculty of Social Sciences of the Radboud University (ECSW 2016-2208-41).

As previously described (46, 47), the monkeys were pair-housed to stimulate normal social behavior. About 24 hours before the start of a test session, water intake was limited to 20 ml/kg. The monkey earned a small water reward of 0.2 ml per successful test trial. After a test session, if needed, water was supplemented to the required minimum of 20 ml/kg, and in addition, the animal received pieces of fruit. In weekends, the animals’ fluid intake was increased to 400 ml daily.

To monitor the animal’s health status, body weight, and water and food intake were daily recorded. Expert veterinarian assistance was available on site. Quarterly recording of hematocrit values ensured that the animal’s kidney function remained within the normal physiological range. When signs of discomfort or illness were observed, experiments were stopped, and the animal was treated until he was recovered. Our procedures follow the water-restriction protocol of the Animal Use and Care Administrative Advisory Committee of the University of California at Davis (UC Davis, AUCAAC, 2001).

Participants, Animal Care and Training

Five adult male rhesus monkeys (Macaca mulatta; weights 6.5 to 9.5 kg; m1-m5) and five adult humans (age between 23 and 43 years; one female; two naive volunteers; h1-h5), participated in our experiments. Monkeys could move their head
freely while seated in a custom-made primate chair within a sound attenuated room. Each monkey learned to release a down-pressed bar upon the onset of an audible stimulus to receive a water reward. Psychophysical testing started when performance had stabilized, and all monkeys could successfully complete a daily session with at least 1500 trials.

**Pure Tone Adaptive-Tracking Procedure and Stimulus Control**

Tones (0.250, 0.375, 0.500, 0.750, 1.0, 1.5, 2, 3, 4, 6, 8, 12, 16, 24 and 32 kHz) were digitally synthesized and delivered online (260 kHz sampling rate) to a loudspeaker in the free field at the straight-ahead position (distance ~80 cm), using Tucker Davis Technology’s hardware (TDT, Alachua, FL, USA – RX6 Systems 3). Attenuation occurred through custom-built amplifiers. Loudspeaker output (Visaton GmbH SC5.9) was sine-onset/cosine-offset ramped (5 ms rise/fall time) and defined by a flat frequency characteristic (to within 3 dB) from 0.1 up to 50 kHz after equalization (Behringer International GmbH, Willich, Germany – Ultra-Curve PRO DSP8000). Sound intensity was calibrated by adjusting its root mean square (RMS) voltage with respect to a reference voltage (1 kHz at 80 dB sound pressure level (SPL)) and measured at the approximate position of the subject’s head with a calibrated Brüel and Kjær sound amplifier and microphone (B&K, Norcross, GA, USA – BK2610/BK4134). Ambient background noise levels varied between 30-35 dB SPL. Reflections above 500 Hz were effectively attenuated by acoustic foam (Uxem, Lelystad, the Netherlands – Redux AX2250) covering the walls, floor, ceiling, and every large object present.

Speaker-derived pure tone thresholds were determined for all subjects—except for monkeys m4 and m5—through a single-interval adaptive-tracking staircase procedure. Each staircase run commenced at 65 dB SPL and was adjusted according to the psychophysical transformed-rule \((48)\). That is, the intensity of a given tonal frequency was decreased by 10 dB after three consecutive hits, whereas it was increased by 10 dB after two consecutive misses. After four (monkeys) or two (humans) reversals the adaptive step size was reduced to 2 dB. Testing continued until at least 13 (monkeys) or 11 (humans) reversals had occurred for which the
averaged intensity level was stable within 2 dB. Examples for five tones presented to monkey m1 are shown in Fig. 3A.

The ability to perceive the onset of a pure tone was assessed by having listeners release a bar as soon as they heard the tone. We randomly varied the inter-stimulus time between 500 and 3100 ms. All tones lasted 600 ms. Lapses in attention were monitored through catch trials, comprising a tone well above threshold. Catch trial tones had the same frequency as the staircase test stimulus with which it was randomly interleaved. Monkeys received ≈35%, and humans ≈5% catch trials. Through this high percentage of catch trials the probability of being rewarded was 0.6, which ensured the monkey’s motivation to perform at high level. Staircase runs with lapse rates above 10% were discarded. Hearing thresholds were measured daily, with the 15 tonal frequencies presented in a random order to avoid bias. The final threshold estimates combined the data from 6 x 15 (per monkey: m1-m3) or 2 x 15 (per human:h1-h5) staircase runs that did not deviate more than 10% from the mean value (Fig. 3B).

Finally, we performed Monte Carlo simulations to emulate the performance of an ideal observer—not limited by internal noise constraints (49)—responding to a single-interval hold-release task version of our three-down/two-up transformed-rule. These simulations are needed because our single interval task—equivalent to the one shown in Fig. 4A—essentially equates to a simple non-forced yes-no task for which there is no expected probability of the stimulus appearing at a given point in time, as opposed to a two-interval forced choice task where the probability of the stimulus presence equals 0.5 (48). For 100,000 simulations each containing up to 100 adaptive steps, the mean proportions of correct responses were found to range from 55% to 65%, with an average of 60%.

Ripple Sound Design and Parameterization

Our test sequences with the ripple stimuli comprised a flat broadband noise of duration, \( D \), followed by a S-T modulated broadband complex. Each ripple, \( S(t) \), included 126 simultaneously presented tones equally spaced—20 per octave—along the logarithmic frequency scale, ranging from \( f_0 = 250 \text{ Hz} \) to \( f_{126} = 19 \text{ kHz} \) (spanning 6.25 octaves):
Comparative Primate Audition

\[ S(t) = \sum_{n=1}^{126} R(t,x) \cdot \sin(2\pi \cdot f_n \cdot t + \varphi_n) \quad \text{for} \quad -\pi < \varphi_n < +\pi \]

with \( f_n = f_0 \cdot 2^{(n-1)/20} \) for \( 1 \leq n \leq 126 \) (1)

Apart from the \( f_0 \) component, which had its phase fixed at maximum amplitude (\( \Phi_0 = \pi/2 \)), tonal phase, \( \Phi_n \), was randomized between \(-\pi\) and \(+\pi\). Noise amplitude was modulated by a single sinusoidal envelope, \( R(t,x) \):

\[ R(t,x) = \begin{cases} 
1 & \text{for } 0 \leq t < D \\
1 + \Delta M \cdot \sin(2\pi (\omega \cdot t + \Omega \cdot x)) & \text{for } 0 \leq t < D \\
1 + \Delta M & \text{for } 1 \leq n \leq 126 
\end{cases} \]

Here, \( t \) is time [seconds]; \( x \) is the position on the frequency axis in octaves above \( f_0 \); \( \omega \) is the temporal modulation rate, called ripple velocity [Hz]; \( \Omega \) the spectral modulation rate, called ripple density [cycles/octave, or c/o]; \( \Delta M \) is the envelope amplitude modulation-depth of the ripple on a linear scale from 0 up to 100%; \( D \) defines the duration of the static noise (\( \omega \) and \( \Omega \) are set to zero) at the onset of the stimulus sequence. In the modulated second part (\( t > D \)), the sign of the \( \omega/\Omega \) ratio sets the upward (< 0) or downward (> 0) direction with which the amplitude envelope sweeps the S-T domain. As illustrated in Fig. 1A, pure temporal amplitude modulations, \( \Omega = 0 \), give rise to vertically oriented modulations, called amplitude modulated (AM) noises. Pure spectral modulations, \( \omega = 0 \), give rise to horizontally oriented modulations, called static ripples (50). Sound level (RMS) was fixed at 56 dB SPL for both the static noise and the ripple. As detailed in Fig. 4A, \( D \) was varied from 1.0 to 3.0 s in humans and from 1.5 to 3.5 s in monkeys. Modulation duration equaled 0.8 s in humans or 1.0 s in monkeys. The longer duration for the monkeys was needed to ensure stimulus control at low modulation-depth levels. Also other studies have found that monkeys perform better when exposed to longer durations in temporally-based auditory tasks (51).

All stimuli were selected from a matrix of 88 combinations of (n=11) ripple densities \( \Omega (-3.0, -2.4, -1.8, -1.2, -0.6, 0, +0.6, +1.2, +1.8, +2.4 \text{ and } +3.0 \text{ c/o}) \) and (n=8) ripple velocities \( \omega (0, 4, 8, 16, 32, 64, 128 \text{ and } 256 \text{ Hz}) \). A subset of this matrix is shown in Fig. 1A. Up to 11 \( \Delta M \) levels were used (0, 5%, 7.5%, 10%, 15%, 20%, 30%, 40%, 50%, 70% and 100%). As the catch trial stimulus (\( \Omega(0) = (0,0) \)) was not modulated,
subjects heard 10x87 + 1 = 871 different audible S-T sounds.

Sound synthesis, digitalization, deliverance (50 kHz sampling-rate) and acoustic conditions were identical to those described in the tone audiogram section above, except that each stimulus sequence was stored off-line as a waveform audio file prior to the experiment proper. Sound level was 56 dB SPL. The following methodological requirements were met: First, each subject received a unique set of n x 871 (D, ΔM, ω, Ω) combinations distributed evenly over the recording sessions. Second, D was uniformly distributed over the (ΔM, ω, Ω) combinations (Fig. 5C). Third, the order in which the test sequences were presented was unique for each subject. Fourth, sound intensity and total power of the flat broadband noise equaled that of the ripple.

Finally, we calculated the normalized 4th moment, $M_4$, of our ripple stimuli as defined by:

$$M_4 = \frac{\frac{1}{T} \int_T s^4(t) dt}{\left(\frac{1}{T} \int_T s^2(t) dt\right)^2}$$  (3)

where $s(t)$ is the time-domain representation (Eq. 1), and T is the duration. $M_4$ is of behavioral relevance because it provides a measure of instantaneous amplitude fluctuations for which humans are known to be quite sensitive (52).

The averaged log$_{10}(M_4)$ value pooled over 11 ΔM values (mean [95% CI]) of unmodulated noise, dynamic ripples and static ripples were: (2.99 [2.98 – 2.99]), (2.98 [2.98 – 2.99] and (4.2 [3.96 – 4.55]), respectively. Thus, static ripples stand out from the dynamic ripples in the sense that they could in principle be discriminated based on their higher $M_4$. However, for $\Delta M \leq 55\%$ (still well above threshold, see Fig. 6A), the $M_4$ of static ripples did not deviate significantly from those obtained from flat noise or dynamic ripples, thus listeners could not use instantaneous fluctuations as a potential cue.

**Ripple Detection Paradigm, Stimulus Control Monitoring and Number of Trials**

We assessed perceptual performance by requiring listeners to release a response bar upon detection of an audible change in an otherwise flat broadband noise. A trial started by pushing down a response bar and terminated when the bar was released upon ripple-onset detection. Responses between 220 and 800 ms after ripple onset were defined as hits. When subjects failed to detect the modulation (latency > 800
ms), the response was counted as a miss. Early bar release trials (latency < 220 ms) were discarded (Fig. 4A).

We used the method of constant stimuli to measure the S-T modulation detection performance to all 88 (Ω, ω) combinations as a function of 11 stimulus levels, ΔM. Daily, stimulus levels were presented in a randomly intermixed sequence from a predefined subset of randomly selected (Ω, ω) combinations to form a single recording session. Daily recording sessions contained ≈1,600 responses for monkeys, as opposed to ≈600 for humans.

To monitor stimulus control in monkeys, pure static noise catch trials—(Ω, ω) = (0,0)—, presented at a probability p ≈ 0.35, were randomly interleaved with the test sequence trials. Human listeners received catch trials at p ≈ 0.15. Catch trial performance—i.e., a measure of the listeners’ guess rate—ranged from 16% up to 35% in monkeys. For humans, this range was 1% to 12%. We observed that guess rates were roughly constant for both species and within the expected range known from the literature. In addition, we found that procedural and perceptual learning did not have a long-lasting effect on performance in both humans and monkeys. This is shown in Fig. A1 of the Appendix section on the robustness of the detection threshold estimation.

In total, each of the 968 ripple stimulus parameter combinations—ΔM, ω, and Ω—was repeated at least 16 times for monkeys and 8 times for humans. Measurements were terminated when the 95%-CI of all 88 (Ω,ω) thresholds (Eq. 5) was less than 10%.

Overall, monkeys performed in more than 19,000 trials, which were spread out over 20-30 daily recording sessions. Humans, on the other hand, performed in about 8,000 trials or more, which were spread out over 13-16 recording sessions. The total number of responses required to obtain reliable threshold estimates (95%-CI < 10%) was higher for the monkeys (m1-m5: 19721-23291) than for the human listeners (h1-h5: 8481-8811).
Psychometric Function Parameterization and Fitting

Our single-interval psychometric functions—\( P(x; \alpha; \beta; \gamma; \lambda) \)—were parameterized as cumulative Weibull distribution functions \( F(x; \alpha; \beta) \):

\[
P(x; \alpha; \beta; \gamma; \lambda) = \frac{\gamma + (1 - \gamma - \lambda)}{2} \cdot F(x; \alpha; \beta) \] (with \( F(x; \alpha; \beta) = 1 - e^{-(x/\alpha)\beta} \) for \( 0 \leq x \leq 100 \)) \tag{4}

Here \( x \) is the dependent variable \( \Delta M \); \( \gamma \) is the guess rate—false positives—representing the fraction of trials where listeners released the bar at random, but within the hit-window time interval (as defined in Fig. 4A); \( \lambda \) is the lapse rate—misses—calculated from the difference between 100% correct and the actual performance at near maximum \( \Delta M \) values. Thus, \( \gamma \) and \( \lambda \) define the lower—close to 0%—and upper bound—close to 100%—of the psychometric function, respectively (as indicated graphically in Fig. 1B). The parameter \( \alpha \) specifies the threshold response criterion—defining the function’s relative position along the \( x \)-axis—and \( \beta \) specifies the slope—lateral spread—of the cumulative Weibull distribution function. The detection threshold was defined as the \( \Delta M \) value of \( x \) for which responses fell on the midpoint between the lower and upper bound of the fitted psychometric curve (Fig. 1B):

\[
P(x; \alpha; \beta; \gamma; \lambda) = \gamma + (1 - \gamma - \lambda)/2 \] \tag{5}

The four parameters—\( \alpha, \beta, \gamma \) and \( \lambda \)—were treated as free parameters. As Bayesian constraining prior functions (58) we chose Beta distributions for \( \lambda \) and \( \gamma \), normal distributions for \( \alpha \), and log-normal distributions for \( \beta \). The log-likelihood ratio, based on 10,000 Monte-Carlo simulations, allowed verification of the goodness-of-fit: two-sided \( \chi^2 \) deviance > 20, \( p < 0.003 \). That is, the likelihood of finding a deviance greater than 20—given 11 stimulus levels and 4 free parameters (i.e., number of degrees of freedom equals 7)—by chance alone for all of the 880 fitted psychometric functions (pooled across all 10 subjects) was less than 0.3% (57, 59). Cross-validation analysis by means of Bayesian inference and model-free estimation (59) on 10% (randomly selected) of the performance data collected did not produce thresholds and slopes with significantly different 95%-CIs. In other words, the estimated thresholds as derived from the raw performance data did not depend on the statistical method used.

Blocks of trials wherein monkey listeners did not reach 100% detection level at easily detectable ripple modulation levels—i.e., lapses in attention judged on misses (lapse-
rate) to $\Delta M > 80\%$—were discarded $\left(53, 54\right)$. This was determined through visual inspection of the upper bound of the fitted psychometric curves $\left(Eq. 4\right)$. About 40% to 20% of the daily recorded monkey responses were discarded. This level of inattention is not uncommon for trained macaque monkeys $\left(51, 55, 56\right)$.

Threshold S-T MTF Construction

From the ripple detection performance data of each listener—as obtained for each of the 87 $(\Omega, \omega)$ detectable combinations—we fitted a psychometric function $\left(Eq. 4\right)$ using the constrained maximum-likelihood algorithm $\left(57\right)$. Ultimately, all 87 functions were used to construct the threshold S-T MTF matrix $M(\Omega, \omega)$; as shown graphically in Fig. 1. The threshold-value for the stimulus at $(0,0)$—for which the threshold cannot be determined because it is inherently indistinguishable—was determined through interpolation.

Suprathreshold S-T MTF construction

We obtained supra-threshold MTFs by constructing iso-$\Delta M$ MTFs from the complete data base of psychometric functions. That is, instead of using the performance scale—the threshold response criterion $\left(Eq. 5\right)$, used was the stimulus scale, $\Delta M$, as the independent measure for constructing supra-threshold MTFs.

MTF normalization

To enable a visual and direct quantitative comparison across-subject and between-species (Fig. 8A, Fig. 12A, Fig 13 and Fig 14) we normalized all values of MTF matrix $M(\Omega, \omega)$:

$$M_{\text{norm}}(\Omega, \omega) = \frac{M(\Omega, \omega) - \min[M(\Omega, \omega)]}{\max[M(\Omega, \omega)] - \min[M(\Omega, \omega)]}$$

(6)

with max$[M(\Omega, \omega)]$ and min$[M(\Omega, \omega)]$ representing the highest and lowest values of the MTF of each listener. In this way, each value was scaled onto a $[0, 1]$.

Inseparability Index $a_{\text{svd}}$

The degree of separability was quantified for each $M(\Omega, \omega)$ through singular value decomposition—SVD; $\left(30, 36\right)$—expressing $M(\Omega, \omega)$ as the product of three matrices:

$$M(\Omega, \omega) = G(\omega) \cdot K(\lambda_i) \cdot H(\Omega)$$

(7)
G(ω) and H(Ω) are orthogonal matrices, and K(λ_i) is the singular matrix with the dimensionless eigenvalues λ_i on its diagonal, and zeros elsewhere. If the singular matrix has only one significant eigenvalue — λ_1 > 0 and λ_{i>1} = 0 — then M(Ω,ω) is fully explained by the product of two orthogonal vectors. These are the first singular vectors in Gω(Ω) and Hω(Ω), representing the temporal (TMTF) and the spectral (SMTF) modulation transfer functions, respectively. In other words, M(Ω,ω), is then said to be fully separable, when every row is a scaled version of every other row, and columns are scaled versions of each other too.

The degree of separability was quantified using the inseparability index (21, 30, 36, 60):

$$\alpha_{SVD} = 1 - (\lambda_1^2 / \sum_{i=1}^{n} \lambda_i^2)$$  \hfill (8)

with summation over the number of tested velocity values, n = 8, as prescribed by matrix M(Ω,ω). Thus, $\alpha_{SVD}$ represents the proportion of the total power in M(Ω,ω) that is accounted for by its highest inseparable approximation. If $\alpha_{SVD} = 0$, the power in the MTF is only determined by the first eigenvalue and thus separable. If $\alpha_{SVD} > 0$, however, then G(ω) and H(Ω) may interact.

To test the statistical significance of $\alpha_{SVD} > 0$, the $\alpha_{SVD}$ values computed from the actual M(Ω,ω) were plotted against those computed from randomly permuted versions of M(Ω,ω), see e.g., Fig. 9A. This was achieved by generating 100,000 bias-corrected percentile bootstrap (61) samples of $\alpha_{SVD}$ for both the actual and randomized data.

**Singular Value Decomposition S-T MTF: Coefficient of Determination: $r^{2}_{SVD}$**

To quantify how a given $\alpha_{SVD} > 0$ (Fig. 2, right plots: inseparable) relates to the degree with which the actual measured M(Ω,ω) can be reconstructed, we replaced the singular eigenvalue matrix K(λ_i) of Eq. 7 with only its first eigenvalue λ_1. This yields the predicted or recovered MTF: M_{rec}(Ω,ω), under the assumption of full separability. A quantified measure for the degree of separability, $r^{2}_{SVD}$, was calculated by performing a Spearman’s rank correlation between each of the 88 elements of M_{rec}(Ω,ω) and the measured M(Ω,ω). Thus, a fully separable MTF gives rise to an $r^{2}_{SVD}$ that equals unity (Fig. 2, left plots: separable).
Up/Down symmetry MTF: Coefficient of Determination: $r^2_{\text{up/down}}$

To quantify the degree of up/down symmetry, the MTF of a single listener was divided into a pair of half-matrices: one containing the thresholds—iso-$\Delta M$ scores—to upward, $M(\Omega < 0, \omega > 0)$, and the other to downward $M(\Omega > 0, \omega > 0)$ moving ripples (see Fig. 2). The derived measure of squared correlation between upward and downward sensitivity, $r^2_{\text{up/down}} = 1$, reflects perfect symmetry when the gain, $g$, of the relationship: $M(\Omega < 0) = g \cdot M(\Omega > 0)$ equals one. Bootstrap analysis (100,000 bias-corrected percentile bootstrap samples (61) of randomly permuted MTF matrices gave rise to $r^2_{\text{up/down}} \leq 0.8$. Thus, $r^2_{\text{up/down}} > 0.8$ signifies a high degree of up/down symmetry that did not arise spontaneously.

Mutual information

We applied a mutual information analysis (see e.g., Figs 8B, 9B, and 12B, C) to obtain a quantifiable measure of the geometric relationship between a pair of $M(\Omega, \omega)$ matrices, like the ones shown in Fig. 8A (62).

Mutual information is defined as follows. For $X$, a discrete random variable with probability distribution, $p(X)$, the Shannon entropy (63, 64)—in bits—is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

(9)

Here $X$ can take $n$ discrete values $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$. Note that $H(X) \approx 0$ when $p \approx 0$, or $p \approx 1$, otherwise $H(X) > 0$. The mutual information of $H(A)$ and $H(B)$ is then defined by:

$$I(A; B) = H(A) + H(B) - H(A, B)$$

(10)

where $H(A,B)$ is the conditional entropy of $A$ given $B$. If $A$ and $B$ are dependent variables, the total entropy is reduced. Mutual information is sensitive to both the size and the information content of the overlap between $A$ and $B$ (64). For further details, see Appendix: Mutual information.

Probability Density Estimation

Non-parametric kernel density-estimation methods (65) allow for optimal interpolation of finite data to construct a continuous representation (66). We used an
adaptive MATLAB (The MathWorks, Inc., Natick, MA) implemented algorithm—based on the smoothing properties of linear diffusion processes (67)—to compute probability density functions.

**Random Permutation**
Data were randomly permuted by means of the MATLAB `randperm` function. That is, the index numbers—row and/or column indices—of a given 1D data vector or 2D data matrix were reshuffled randomly.

**Kolmogorov-Smirnov (K-S) Test**
Two-sample Kolmogorov-Smirnov (non-parametric) testing was performed to compare the empirical distribution functions of two continuous random variables (with sample size n) under the null hypothesis, $H_0$, that both are from the same continuous distribution.

**Kendall's Rank Correlation**
Kendall’s rank correlation is a nonparametric test of independence. We calculated the Kendall’s tau correlation coefficient, $tau-b$, under the null hypothesis that there is no ordered relationship.

**Confidence Intervals**
Confidence intervals (CI) were estimated using a nonparametric, bias-corrected bootstrapping algorithm (61).

**Auditory spectrotemporal MTF (S-T MTF) modeling**
Details on the computation of human and monkey S-T MTF surfaces of Fig. 14 (audition: two top-left plots) are provided by Eq. A2 in the Appendix section on: Auditory Spectrotemporal MTF Modelling.

**Visual Spatiotemporal MTF modeling**
Details on the computation of the human space-time surface, $G(\alpha, \nu)$ of Fig. 14 (vision: top-right plot) are provided by Eq. A3 in the Appendix section on: Visual Spatiotemporal MTF Modelling.
RESULTS

Pure Tone Hearing Sensitivity

We determined free field pure tone audiograms, to ensure that the listeners had normal hearing. Figure 3A shows an example of our psychophysical adaptive-tracking procedure with monkey m1 for five different tones. Note the stable behavior around the different pure tone thresholds, demonstrating that the animal was under full stimulus control. We obtained similar results for monkeys m2 and m3.

Figure 3B shows the averaged hearing-thresholds of all human listeners (h1-h5, left panel) and the three monkeys (m1-m3, right panel) for all tested frequencies (0.125 ≤ f ≤ 32 kHz). Note that the monkeys’ hearing range extends to frequencies that are inaudible to humans (monkeys: threshold ~40 dB at f = 32 kHz; humans: threshold > 90 dB). In conclusion, the mean range of our subjects’ audiograms corresponds well with normal hearing (68).

Figure 3. Tracking procedure and free field audiograms.
A: Graphical representation of the adaptive-tracking procedure for monkey m1, as recorded on a single day. The thin black line shows 5 successive staircase runs produced by onset detection judgments to pure tones of 1000, 375, 8000, 3000, and 24000 Hz, respectively. Stimulus level was decreased after three consecutive hits and increased following two consecutive misses. A single run ended when at least 13 reversals (circles) occurred in the direction of the change in stimulus level. The average provided an estimate for the detection threshold as indicated by the arrows (dashed lines). This high level of stimulus control was observed in all monkeys tested (m1-m3). B: Audiograms showing the hearing thresholds of human (h1-h5: left panel) and monkey (m1-m3: right panel) listeners, presented on a logarithmic scale. Grey areas indicate the 95%-confidence intervals (95%-CI) as assessed by bias-corrected percentile bootstrap resampling on 100,000 evaluations. Our data concurs with the known literature, showing that monkeys can hear sounds at frequencies that are much higher than humans can. Their hearing range extends to over 20 kHz, while humans can only hear sounds up to about 16 kHz.
Comparative Primate Audition

**Ripple Stimulus Paradigm and Latency Analysis**

Listeners were trained (monkeys) or instructed (humans) to release a response bar upon detection of an audible change—i.e., ripple onset—in an otherwise static broadband noise of random duration. In total, we employed 87 combinations of spectral and temporal modulation rates (Fig. 1C), across 10 modulation-depths, $\Delta M$ (Fig. 1B), plus a catch stimulus without S-T and amplitude modulation $[(\Omega, \omega, \Delta M) = (0,0,0)]$. As such, each listener was exposed to a (pseudo) randomized sequence of 871 unique $(\Delta M, \Omega, \omega)$ ripple combinations, for which the timing of the ripple onset—dictated by duration, $D$, of the static noise—was unpredictable (horizontal grey bars, Fig. 4A and vertical grey bars Fig. 5C). During testing, each ripple was repeated at least $n = 16$ (monkey) or $n = 8$ (human) times.

Figure 4B illustrates the complete latency distributions of human h1 (8,811 responses) and monkey m1 (19,721 responses). Both histograms reveal a clear bimodal distribution. The first peak—Hits—corresponds to correctly detected ripples. The averaged hit latency—median plus confidence interval [95%-CI] in ms—in our monkey (m1-m5) and human (h1-h5) listeners was 400 [366-412] ms and 443 [323-472] ms, respectively. These data correspond well to reaction times of sound-evoked hand/arm movements (69). The second peak—Misses—corresponds to ripples that listeners failed to hear.

The pooled latency data of Fig. 4C were selected for hits only and are displayed as a function of cumulative trial number across all recording sessions. Compared to our human listeners (h1-h5, top panel), the monkeys (m1-m5, bottom panel) were on average 43 ms faster in releasing the response bar upon ripple-onset detection. However, neither the mean (white lines) nor the variability (grey areas) of the reaction times changed over time for both monkeys and human listeners. This stable performance indicates the absence of perceptual learning during the experiments and permitted pooling of the data across different recording sessions.
Comparative Primate Audition

Figure 4. Behavioral paradigm and hit latency analysis.

A: Definition of hits and misses, trial duration and catch trials. A trial started by holding down a response bar and terminated when the bar was released (vertical arrows) upon ripple-onset detection. Static noise duration was randomized. Ripple duration remained constant. Responses between 220 and 800 ms after ripple onset were defined as hits (Hit window). When subjects failed to detect the modulation (latency > 800 ms), the response was counted as a miss. Early bar release trials (latency < 220 ms) were discarded. Note, catch trials comprised static noise only. B: Distribution of reaction times for human h1 (top) and monkey m1 (bottom). The first peak corresponds to hits. The latencies around the second peak are due to misses. Data were collected over a period of 3-6 months. The inset (dark gray, lower plot) shows a 2x2 contingency table summarizing the four possible stimulus-response outcomes. Note, the probability of a hit response on catch trials when the ripple-onset cannot be detected—false positive—is a measure of a listeners’ guess rate defining the lower bound—γ—of the fitted psychometric curve (see Fig. 1B). C: Hit reaction times as function of cumulative trial number, pooled across stimuli: human (top: h1-h5) vs. monkey (bottom: m1-m5) listeners. The data cover the last 4,200 (top panel) or 8,500 (bottom panel) recorded hits of each listener. Solid white lines represent the 30-trial running average as a function of cumulative trial number. Grey areas reveal the variability in the underlying mean latencies. Note absence of a learning effect since the running averages do not decay over time.

Ripple Detection Performance and Stimulus Variability

Figure 5 illustrates two psychometric response data sets: performance—% correct; Fig. 5A—and response latency—Fig. 5B—for one human (h1, left) and monkey (m1, right) listener. Both responded to the same dynamic ripple (Ω = -3.0 c/o, ω = 32 Hz),
presented under various modulation-depths, $\Delta M$, and randomized static noise durations, $D$.

The estimated thresholds—$\Delta M$ at vertical midpoint between lower and upper bound of the black fitted curves; Fig. 5A—were comparable for the two listeners, as indicated by the crossings between the vertical and horizontal grey lines (h1: 27 [23 - 33] % vs. m1: 24 [20 - 31] %). The estimated slopes ($\beta$ [95%-CI]), however, differed significantly (human h1: 3.5 [2.5 - 3.9] vs. monkey m1: 2.1 [1.1 - 2.4]).

The response reaction time decreased systematically with increasing $\Delta M$ (Fig. 5B). Here, the upper and lower limits (horizontal grey lines) of the fitted curves correspond to the peaks of hits and misses in Fig. 4B, respectively. Note, however, other studies have found that for a given $\Delta M$, reaction time changes systematically as a function of ripple velocity, as well as ripple density (18, 46). That is, the reaction time is determined by all the three parameters defining the amplitude envelope of a ripple stimulus: (1) Velocity, $\omega$ [Hz]: temporal modulation; (2) Density, $\Omega$ [c/o]: spectral modulation; (3) Modulation depth, $\Delta M$, [%].

Stimulus variability can be a confounding factor in the sense that longer delays in ripple onset, $D$, may induce a more liberal placement of the internal decision criterion, resulting in different response latencies (69). To check for this possible, confound, we analyzed hit latency against $D$, but did not obtain any systematic relationship (Fig. 5C). This was verified by Kendall's rank correlation, one-tailed test: h1: $\tau-b < 0.1$, $p > 0.1$ (left panel); m1: $\tau-b < 0.07$, $p > 0.2$ (right panel). Comparable non-significant $p$ values were obtained for the other listeners, h2-h5 and m2-m5.
Comparative Primate Audition

Figure 5. Ripple detection threshold and latency analysis. A: Fitted psychometric curves (Eq. 4) of listeners h1 and m1 for dynamic ripple $\Omega$: -3.0 c/o, $\omega$: 32 Hz. Vertical dashed-lines define the 95%-CI of the detection thresholds. The grey areas comprise 10,000 evaluations of the expected performance function, obtained through Monte-Carlo resampling. Trial sample-sizes: $n_{h1} > 11 \times 12$; $n_{m1} > 11 \times 18$. B: Ripple-onset detection latency as a function of $\Delta M$ to the same ripple. The black lines were based on a shape-preserving interpolation algorithm for illustrative purposes only. Grey areas define the 95%-CIs, as assessed by bias-corrected percentile bootstrap resampling (100,000 samples). The upper and lower horizontal grey lines mark peaks of hits and misses of the reaction time distributions in Fig. 4B. C: Static noise duration—$D$—and ripple-onset reaction time—Latency—of human h1 and monkey m1 are unrelated. Histograms indicate near-uniform distributions of $D$ (vertical grey bars), and non-uniform distributions of latency (horizontal grey bars).

Statistical Analysis Psychometric Parameters: Threshold ($\alpha$) vs. Slope ($\beta$)

The expected performance functions of the fitted psychometric curves (Fig. 5A) were parameterized as a cumulative Weibull distribution function $F(x; \alpha; \beta)$ (Eq. 4), wherein $\alpha$ determines the threshold and $\beta$ determines the slope. Fig. 6 summarizes an across-subject characterization of the fitted psychometric data, pooled across all combinations of ripple densities and velocities.

Figure 6A shows the probability density distributions of the $\alpha$ and $\beta$ values, pooled across human (h1-h5, light shading) and monkey (m1-m5, dark shading) listeners, respectively. An across-subject analysis of the $\alpha$ or $\beta$ distributions for testing within species differences did not reveal any significant difference (two-sample: $n_1 = 87$, $n_2 = 435$, one-tailed K-S statistic: human h1-h5 $\alpha$: $k \leq 0.16$, $p > 0.1$; $\beta$: $k \leq 0.12$, $p > 0.05$ vs. monkey m1-m5 $\alpha$: $k \leq 0.15$, $p > 0.1$; $\beta$: $k \leq 0.18$, $p > 0.05$). Next, we established that the species-specific $\alpha$ distributions (human vs. monkey) did not differ in overall shape either (two-sample: $n_1 = 435$, $n_2 = 435$, two-tailed K-S statistic: $k = 0.09$, $p >$
0.05), as can be inferred from their corresponding cumulative distributions (inset; left plot Fig. 6A).

In contrast, the slopes of the pooled monkey data were consistently lower, when compared to those of the pooled human data (inset; right panel, Fig. 6A): the peak of the human $\beta$ probability density function is centered at 3.6 (bandwidth: 4.5), vs. that of the monkeys at 2.6 (bandwidth: 2.4). K-S testing confirmed that these distributions were significantly different (two-sample: $n_1 = 435$, $n_2 = 435$, $k = 0.44$ $p < 0.0001$). Thus, ripple detection thresholds were determined with a higher discriminating power—i.e., steeper slopes—in humans than in monkeys. Importantly, the narrow bandwidths of $\beta$ suggest that the ripple-based S-T sensitivity is characterized by a relatively constant slope of the psychometric curves.

In Fig. 6B, we compared the ripple thresholds of each listener with those pooled and averaged across humans (h1-h5; left panel) and monkeys (m1-m5; right panel), respectively. The large overlap between the 95%-CIs of the squared correlation coefficients and their proximity to unity reveals a close correspondence between the averaged and the respective individual threshold data for both humans (left inset box) and monkeys (right inset box).
Comparative Primate Audition

Figure 6. Psychometric threshold and slope value distributions. A: Shown are distribution plots of the fitted psychometric function (Eq. 4) parameters $\alpha$ (threshold: left plot) and $\beta$ (slope: right plot). Data are pooled across human (h1-h5; light shading) or monkey (m1-m5; dark shading) listeners; representing 2 x 5 x 87 data points per plot. Insets represent Log-scaled cumulative distributions of the associated probability density functions. Notice that the $\alpha$ distributions fully overlap; the $\beta$ distributions have a narrow width (< 5), whereby monkeys yield systematically lower slopes. B: Comparison between $\alpha$ values of individual subjects and the pooled averaged thresholds (humans: left; monkeys: right). Grey shaded areas represent the convex hull enclosing all 87 data points as obtained from each listener. Thresholds are scaled onto the [0,1] range (Eq. 6). The diagonal (black) corresponds to the identity line. Insets represent the 95%-CIs (grey lines) of the coefficients of determination (black dots) corresponding to the data of each listener shown below. None of the 95%-CIs has a value below 0.8, highlighting a close relationship between the individual and the averaged datasets.

Comparative characterization of raw performance data

Figure 7 provides a complete overview of the relationship between the raw (i.e., non-fitted) performance data and the S-T parameters of all dynamic ripple stimuli. Each colored contour plot shows a two-dimensional performance pattern for a particular ripple velocity, whereby the performance levels belonging to a unique $(\Omega, \omega)$ combination are ordered vertically as a function of $\Delta M$. We observed several striking similarities and differences between the pooled raw performance patterns of human (h1-h5, Fig. 7A) and monkey (m1-m5, Fig. 7B) listeners. First, the blue-yellow-colored
contours shift progressively upwards with increasing ripple velocity, ranging from 4 up to 256 Hz. Its progression, however, is more prominent in humans than in monkeys, signifying that monkeys are more sensitive—i.e., better performance at low modulation-depths—to ripple velocities above 16 Hz. Second, performance decreases with increasing ripple density. This trend occurs in both species and is most prominent for the high velocities.

The human performance patterns, however, have dark-red contours that are not seen in the monkey results. This demonstrates that monkeys rarely reached 100% performance for modulation values up to 50%, as was shown in Fig. 6A for monkey m1. Moreover, the human response patterns show less variability—i.e., clearer transitions in coloring—. This characteristic is consistent with the observation that the averaged guess rate of the monkeys was higher than that of humans (See the Appendix section on: Robustness of Detection Threshold Estimation). Also note that the iso-density contours at 0 c/o (vertical midlines) in the 0 Hz velocity plots are dark-blue. Thus, the catch trials evoked adequate low performance levels in all listeners,

**Figure 7. Performance characterization to dynamic ripples.** A-B: Averaged contour performance plots of modulation depth ($\Delta M$: 0 to 50 %) against ripple density ($\Omega$: -3.0 to +3.0 c/o), for all 8 ripple velocities ($\omega$: 0 to 256 Hz). Performance levels are color coded. Within each subplot, iso-density lines represent raw psychometric functions, like the fitted ones in Fig. 5A. Contours were smoothed for illustrative purposes. (A) Pooled human data (N = 43,115 trials). (B) Pooled monkey data, (N = 107,787 trials).
thereby signifying their non-modulated acoustic content. Overall, the raw performance data of Fig. 7 agree well with the statistical analysis of the fitted psychometric data summarized in Fig. 6. Within the same species, ripple detection performance shows a low degree of variability, whereas between species systematic differences may be observed.

**Comparative Characterization of the S-T MTF at threshold**

The threshold-based MTFs of Fig. 8A were obtained by pooling and averaging the normalized MTF matrix, \( M_{\text{norm}}(\Omega, \omega) \) (Eq. 6; see also Fig. 1C), for all human (h1-h5; top panel) and monkey (m1-m5; bottom panel) listeners, respectively.

The MTFs can be best characterized as follows. First, both species reach their peak sensitivity (dark-red contours) around zero density (-0.6 to +0.6 c/o, human vs. -1.2 to +1.2 c/o, monkey). Along the (vertical) temporal modulation axis, however, peak sensitivity is shifted toward higher ripple velocities in the monkey MTF (30-60 Hz) when compared to the human MTF (6-20 Hz).

Second, the limit of the S-T modulation-rate is expressed by the slope of the velocity-density sensitivity (dashed lines). The steepness of this slope—determined through linear regression of the .38 (yellow) contour—in the monkey MTF (107 [105 - 109] log(Hz)/c/o) is 1.8 times higher compared to the slope of the human MTF (60 [57 - 61] log(Hz)/c/o). Note also that their respective offsets at zero density (as [95%-CI] Hz) are shifted by almost one octave with respect to each other: (monkeys: 287 [284 - 291] Hz) vs. (humans: 163 [162 - 165] Hz).

Third, Fig. 8B shows the probability density distribution functions of the correlation coefficient between normalized MTF for each subject individually (humans: h1 to h5; monkeys: m1 to m5) and the respective averaged human or monkey MTF as shown in (A). None of the distributions have a correlation coefficient lower than 0.8, and overlap extensively (see grey insets), highlighting a close relationship between the individual and their respective pooled MTFs.

The main finding from the human and monkey threshold-based MTF is a systematically ordered, but quantitatively dissimilar pattern of spectral-temporal modulation sensitivities. We used two distinct metrics on the MTFs to quantify the similarity of the two species statistically: mutual information (Eq. 11), and linear
correlation. The left panel of Fig. 8C emphasizes that for temporal modulation rates below 20 Hz the human and monkey MTFs do not differ significantly (purple: high correlation; dark red: high mutual information), whereas above 100 Hz the MTFs differ markedly. The right panel of Fig. 8C shows the same measures plotted as function of the spectral modulation rate. Note that both the mutual information and correlation remain high and independent of ripple density, indicating that across ripple velocities the MTFs for monkeys and humans are highly similar in shape for all ripple densities tested.

Figure 8. Comparative characterization of the S-T MTF.

A: Normalized MTFs based on pooled and averaged threshold data of human (top) and monkey (bottom) listeners. Vertical axis: ripple velocity; horizontal axis: ripple density. Thresholds are normalized (Eq. 6) and color coded (bar), for visualization purposes only. Dashed lines indicate the fall-off in sensitivity at high ripple velocities. Contours are smoothed for illustrative purposes only. The transparent rectangles are inferred from the guess rates obtained from catch stimuli (Ω: 0 c/o, ω: 0 Hz). Notice the high congruence in shape of the MTFs, despite a vertical shift of approximately one octave, between the high sensitivity regions (dark red contours). B: Squared correlation probability density distribution functions between normalized MTF as determined for each subject individually (humans: h1-h5; monkeys: m1-m5) and the respective group human or monkey MTF as shown in (A). Color coded histograms (shades of purple) show the variation of the squared Spearman rank correlation coefficient r² across 100,000 bias-corrected percentile bootstrap samples, as obtained for each subject separately. The grey inset-boxes represent the 95%-confidence intervals of the respective color-coded distributions shown below.
Notice the extensive overlap and similarity in shape signifying a close relationship between the individual and the pooled MTFs. C: Quantitative comparison between human and monkey MTFs. Mutual information (purple) and $r^2$ (red) as a function of log-velocity (left panel), and ripple density (right panel). Colored areas indicate 95%-CI. Mutual information is scaled onto the $[0,1]$ range for illustrative purposes only. Note, however, $r^2$ signifies a linear dependence, whereas mutual information measures general dependence (including non-linear relations) (64, 94). Mutual information highlights that human sensitivity to temporal ripple modulation deviates significantly and in a consistent manner from that of monkeys. The latter is not evident for sensitivity to spectral ripple modulations. The squared correlation analysis does not signify this significant difference in sensitivity to ripple velocity between humans and monkeys.

**Testing for (In)separability and Up/Down Symmetry of the S-T MTF at threshold**

Figure 9A summarizes our statistical analysis on the inseparability indices derived from SVD analysis of the ten threshold-based MTFs, one for each listener. Here, $\alpha_{SVD}$ reflects the degree of inseparability of the measured data, with zero corresponding to full separability across the entire S-T domain. The $r^2_{SVD}$ statistic reflects the proportion of variance accounted for when assuming separability. We compared $\alpha_{SVD}$ to $r^2_{SVD}$ by means of bootstrap resampling for the individual human (h1-h5, left panel) and monkey (m1-m5, right panel) listeners. In the perfectly separable case, the data would be concentrated at $(r^2_{SVD}, \alpha_{SVD}) = (1,0)$ as demonstrated in Fig. 2.

Despite small quantitative differences, the bootstrap analysis gave identical results. In all subjects, the processing of S-T modulations appears to be predominantly separable. Convex-hulls corresponding to the measured data (purple) lie close to the (1,0) point, and do not overlap at all with the simulated convex-hulls determined by chance alone (green). The latter were generated by randomly permuted MTFs. Yet, the small, but systematic, deviations from the (1,0) point, could be explained by the contribution of a small inseparable component in the MTFs.

Figure 9B compares the mutual information and correlation measures to assess up/down symmetry of the MTFs. In both monkeys and humans, the S-T sensitivity patterns for upward ($\Omega < 0$) moving ripples closely resembles the pattern obtained for downward ($\Omega > 0$) moving ripples. First, peak density (bright yellow) of bootstrap samples derived from the measured MTF data is centered at (0.95, 0.83), which is
close to the ideal (1,1) point, signifying perfect up/down symmetry as demonstrated in Fig. 2. Second, the latter do not coincide with the peak densities that arise by chance alone (derived from permuted data): white inset boxes with the highest densities at (0.18, 0.04), which is close to (0, 0); the point representing a total absence of symmetry. Third, despite the slightly higher variability of the monkey data, the peak densities of both species lie closely together.

**Figure 9.** Intersubject separability vs. symmetry S-T MTF analysis. A: Separability analysis: human (h1-h5, left panel) vs. monkey (m1-m5, right panel). Shown are inseparability (a<sub>SVD</sub>) versus separability (r<sup>2</sup><sub>SVD</sub>) parameter plots. Each shaded area represents the convex hull enclosing 100,000 bootstrap samples, drawn from the original MTF data of single subjects. CI-95% intervals from randomly permuted MTFs (green) do not overlap with the those of the measured data (purple). The latter are positioned close to (1,0), in line with spectral-temporal separable MTFs (see fig.2 for explanation). (B) Up/down MTF symmetry analysis: human (h1-h5, left) vs. monkey (m1-m5, right). Shown are color-coded 2D probability density plots—coefficients of determination: r<sup>2</sup><sub>up/down</sub> vs. mutual information: I(up;down)__. Permuted data represents samples from randomly reshuffled MTF values. The highest densities of the permuted data clusters close to (0,0), which arise due to chance alone. By contrast, highest densities from the original human and monkey data cluster around (0.8, 0.9), indicative of nearly perfect symmetry.

In addition, we determined the first singular vectors from the SVD analysis to assess the general shape of the spectral (red curves, Fig. 10A), and temporal MTF (red curves, Fig. 10B), H(Ω) and G(ω), respectively (Eq. 7). These λ<sub>1-SVD</sub> reconstructed one-dimensional transfer functions were compared with the averages of the actual measured transfer functions (black curves, Fig. 10, A-B): note the close similarity in shape of the red and black curves. These results are consistent with the pooled and normalized MTFs of Fig. 8A and the Intersubject MTF analysis of Fig. 9A. First, the threshold-based S-T MTF can be generally characterized as spectrally low-pass and
temporally band-pass. Second, the separable portion of the threshold-based S-T MTF is a viable descriptor of the original data.

Figure 10. Comparative analysis of spectral vs. temporal MTFs at threshold.
A-B: One-dimensional spectral (A) and temporal (B) MTFs of humans (left) and monkeys (right). Black lines represent the averages of the individual G_Ω(ω) and H_Ω(Ω) vectors, of either the pooled human (N=5) or monkey (N=5) MTFs of Fig. 8A. Red lines represent the 1st eigen value SVD reconstructions (Eq. 7) of the averaged data. Modulation index (right y-axis) is given by -20 x log_{10}(ΔM/100) (in dB). Curves are smoothed for illustrative purposes only. Notice that compared to humans, the spectral and temporal modulation sensitivity of macaques is shifted to the lower-end of the frequency domain and the higher-end of the time domain, respectively. Grey lines: cut-offs of the respective SMTFs and TMTFs. Despite these cross-species differences, the close match between the (red) λ₁-SVD reconstructed data and the (black) averaged data strongly supports full separability of S-T sensitivity in both humans and monkeys.

Although Figs. 9 and 10 convincingly demonstrate that the spectral-temporal MTFs of humans and monkeys are strongly governed by separable spectral and temporal processing mechanisms, small, but systematic deviations from the ideal separability points at (1,0) and (1,1) were also noticeable in Fig. 9. These deviations could be largely explained by including the contribution of a small inseparable component in the MTFs, through the second singular value, λ₂. To demonstrate this, we compared the reconstructed MTFs from the separable analysis (Eq. 7, with K(λ₀) = λ₁, i.e., a scalar), with the reconstruction that included the first two singular values, K(λ_i) = [λ₁ 0; 0 λ₂], i.e., a 2x2 diagonal matrix.

Figure 11 compares the r²SVD coefficients of determination for the two reconstructions for each individual listener. Although the inclusion of only the first singular value already accounted for more than 93% of the variability in the
reconstructions for every participant (horizontal dimension, Fig. 11), adding the second singular value improved the reconstruction for all individuals, as now all $r^2_{SVD} > 0.98$ (vertical dimension, Fig. 11). Interestingly, this result held equally for the human and monkey participants. This leaves open the possibility that apart from a dominant fully separable S-T processing mechanism in the primate auditory system, there is a small, but consistent contribution from an inseparable mechanism as well. Indeed, the relative value $\lambda_2/\lambda_1$ varied between 5.1% and 8.8% for the humans, and between 7.7% and 17.0% for the monkeys. Including also the third singular value led to a minor, just significant, further improvement of the fit (not shown).

Figure 11. Relative contribution of singular values to the 2D-shape of the S-T MTF.

Shown is a $r^2_{SVD} \lambda_1$ versus $r^2_{SVD} \lambda_1 + \lambda_2$ scatter-plot comprising MTF data of all human (red, h1-h5) and monkey (blue, m1-m5) listeners. All plotted coefficients of determination were calculated by performing a Spearman’s rank correlation between each of the 88 elements of $M_{rec}(\Omega, \omega)$—reconstructed from either the 1st SVD eigen values or, alternatively the SVD of 1st + 2nd eigen values—and the original $M(\Omega, \omega)$. The vertical and horizontal error bars represent 95%-CIs. The diagonal (black) corresponds to the identity line. The 2nd eigen value of all the measured S-T MTFs adds a small, but consistent, contribution to the reconstructed $M_{rec}$. Note, the averaged $\lambda_2$ is ≈7% and ≈11% of the $\lambda_1$ values for human (brown bars) and monkey (blue bars) listeners, respectively. This is illustrated in the inset (lower left corner). In other words, the first singular value, $\lambda_1$, can be used to recover the overall 2D-shape of the original MTF for up to 95% in humans and 90% in monkeys. Apparently, there is a significant inseparable component to the S-T sensitivity tuning in both species.

Comparative Characterization of the S-T MTF at Suprathreshold.

So far, we constrained our analysis to the perceptual detection thresholds of dynamic ripples. Here, we examine the important question to what extent the threshold-based MTFs generalize to the “ineligibility” of clearly audible—suprathreshold—ripple modulation-depths (for explanation see Fig. 1B). Figure 12
summarizes our suprathreshold analysis of the S-T MTF.

**Figure 12. Suprathreshold characterization of the S-T MTF.** A: Incremental ΔM chronology of 15 iso-ΔM MTF contour plots for human—ΔM: 11-25%; shaded areas left panels in (B) and (C)—and monkey (right column) listeners—ΔM: 6-20%; shaded areas right panels in (B) and (C)—, respectively. Each colored plot represents the pooled and averaged data of human (left, N=5) or monkey (right; N=5) listeners. Colors code represents performance level (% correct). Contours were normalized and smoothed for illustrative purposes only. Notice the striking 2D-shape congruency of many iso-ΔM plots with the threshold contour plots of Fig. 8A. B: Normalized iso-ΔM MTF α_{SVD} plotted as a function of ΔM (black lines). CI-95% intervals from randomly permuted MTFs (green) do not overlap with the measured data (purple), except for monkey ΔM values above 40%. Data are normalized to facilitate comparison with threshold defined MTFs. Because of normalization, abscissa values <1 (dashed red lines) denote iso-ΔM MTFs with a lower degree of separability as was estimated for the corresponding MTFs at threshold. Note, at unity (cross-section, vertical white lines), however, the iso-ΔM MTFs mirror the threshold-MTF in terms of their inseparability index. Normalized separability indices at unity occur at ΔM ≈17% (human) and ≈ 7% (monkey). C: Normalized iso-ΔM mutual information (MI) is plotted as a function of ΔM (black lines). CI-95% intervals from randomly permuted MTFs (green) do not overlap with the measured data (purple), except for ΔM values below 10%. Data are normalized to facilitate comparison with threshold defined MTFs. Because of normalization, maximized MI equals unity. That is, iso-ΔM MI was divided by the maximal obtainable MI at threshold. The latter was obtained for each listener by computing the mutual information between identical pairs of threshold-MTFs. Dark grey shaded areas denote the ΔM values for which the respective iso-ΔM MTFs are comparable to the threshold-MTF: ΔM~16-21% (humans) vs. ΔM~18-23% (monkeys).

Figure 12A shows a subset of the iso-ΔM MTF contour plots for ΔM =11%-25% (human, N=5) data, h1-h5), and ΔM = 6%-20% (monkey, N=5). Note the systematic changes in both the human (left panel) and monkey (right panel) iso-ΔM MTF
Comparative Primate Audition

chronology. First, the regions of higher performance levels (red coloring) gradually increase in size as function of $\Delta M$. Second, irrespective of its size, the overall shape of this region appears to be conserved up to $\Delta M$ levels that supersede the measured thresholds for most stimuli (cf. left panel, Fig. 6A). In Fig. 12B, we quantified the shape of the iso-$\Delta M$ MTFs—up to 45% modulation depth, well above threshold—with respect to the threshold-based MTFs of Fig. 8A in terms of their full separability. Note, a value below 1.0 (red dashed lines) indicates a higher degree of separability at suprathreshold than was the case for the threshold MTFs. Despite quantitative differences, the iso-$\Delta M$ MTF analysis shows that both the human and monkey auditory systems preserve S-T separability and direction sensitivity symmetry at suprathreshold levels.

In Fig. 12C, the ranges of maximal mutual information for stimulus-modulation levels between $\Delta M$~16-21% in humans and between $\Delta M$~18-23% in monkeys compare well to the values obtained from the threshold MTFs, indicating that the overall 2D-shape of the MTF is preserved for these suprathreshold stimuli.
DISCUSSION

Psychoacoustic measurements of perceptual detection thresholds to dynamic ripples have so far been performed only in humans and songbirds (38, 40), and have not included supra-threshold analyses. Here we report on the perceptual spectrum/time sensitivity to inseparable naturalistic acoustic stimuli in normal hearing humans and rhesus monkeys across their dynamic hearing range. Using the same psychophysical methods for both species, we collected free field hearing thresholds to pure tones, and a large data set of manual reaction times for the full range of audible spectral-temporal modulations. Together, these results provide a unique data base to assess and compare the hearing capabilities of human and non-human primates for near-threshold and above-threshold sensation levels. The data show that rhesus monkeys have poorer low-frequency and superior high-frequency hearing than humans (Fig. 3) confirming previous reports (68), and a higher sensitivity to temporal modulations >100 Hz than humans (Figs. 7,8,10) as reported for pure temporal modulations (51).

Our central new finding is that in both species spectral-temporal processing is well understood by largely independent unbiased contributions from two separable components: the averaged spectral, \( H_\omega(\Omega) \), and the averaged temporal, \( G_\Omega(\omega) \), modulation transfer functions (black curves, Fig. 10), respectively. These two components could explain more than 90% of the response data (Fig. 11). This finding did not only hold near the modulation detection-threshold but extended to supra-threshold modulation-depths as well (Fig. 12). Interestingly, a significant contribution from inseparable, processing channels, in which neural populations are tuned to ripples, was also identified to all individuals of both species (Fig. 11).

Psychoacoustics

By applying dynamic ripples that cover the S-T sensitivity range, we avoided testing humans and monkeys to an arbitrary, and possibly biased, set of biological sounds (e.g., conspecific vocalizations), or natural sounds (e.g., recorded environmental noises). Our approach deviates from previous psychoacoustic studies (11, 38-40, 70-73) in that listeners were exposed to a large variation in the stimulus parameters, while determining complete psychometric functions for 87 different spectral-temporal \((\Omega, \omega)\)
combinations (Fig. 1C). Studies that applied S-T modulated sounds determined threshold but not suprathreshold performance (38-40), or discrimination performance (25-27, 74) on a limited set of S-T combinations.

In our study, listeners could never predict which ripple to expect. As such, they could only respond consistently to the sound when attending to the S-T amplitude modulations, rather than to some random spurious event that could have been present in the transition from static noise to ripple. The consistent reaction time distributions (Fig. 4), the low across-subject response variability for both humans and monkeys (Figs. 6 and 7), and the consistent misses for catch trials (Fig. 7) confirm the validity of our experimental approach.

**S-T MTFs versus MPS of Speech**

Figure 13A (left panel) provides a direct comparison of human S-T hearing from our MTF data (Fig. 8A, top) with the reported human modulation power spectrum (MPS) of speech (11). From the overlaid outer and inner contour lines—delineating the 90% and 95% of the modulation power in male (American English) speech—it is obvious that the ripple-based S-T window in humans extends well beyond the dominant modulation spectra of speech. Unfortunately, a direct comparison of our monkey MTFs with the MPS of their vocalizations (12) is not possible, as the ripple density of available MPS is provided in cycles/kHz, whereas our MTFs were based on ripples with logarithmic density (cycles/octave). Irrespective of this scaling difference, the monkey vocalization MPS correlates strongly with that of human speech — \( r = 0.82 \); (12) — . We thus expect that the monkey MTF (Fig. 13A, right) is likely to extend beyond the dominant modulation spectra of its vocalizations too.

When comparing our results (colored contours, Fig. 13B, left) to the psychophysical MTF data from (38) (black contour lines), it is clear that in both studies the perceptual MTF can be characterized as temporal band-pass and spectral low-pass. The only apparent difference is that the highest sensitivity in our MTF is observed at higher temporal modulations (dark red area around 10 Hz) than in the monkey study (38) data (centered around 4 Hz within the 0.06 contour line). Study (11) applied an alternative filtering method, closely related to the use of ripples, from which they derived the S-T MTF for speech intelligibility. They compared their data with the
Comparative Primate Audition

psychoacoustic MTF from (38), showing a high degree of similarity. Given the considerable methodological differences, the similarity in shape between respective MTFs suggests that ripple-based MTFs provide a robust objective measure for S-T hearing in humans. Finally, notice that despite the shift towards higher frequencies of the temporal modulation axis, the monkey MTF (Fig. 13B, right) bears considerable resemblance to the human MTF (left panel).

Figure 13. S-T hearing in humans and monkeys compared to known human MPS data. A: S-T MTF results of the present study (colored contours) remapped in a format equivalent to that reported by (11) for the modulation power spectrum (MPS) of male speech. The MPS forms a second-order representation derived from spectrogram analysis, that is commonly used to visualize and quantify the joint S-T modulations of human speech. For comparative analysis, we show the remapped data of the monkey MTF as well. B: Subset of the data (shaded areas) shown in (A), in a format equivalent to that reported by (38). The overlaid, black iso-performance lines (left panel) represent the data from (38).

Spectral versus Temporal Modulation

Here, we discuss how the dissection of the MTF into one-dimensional spectral and temporal modulation transfer functions (Fig. 10) compares to earlier measurements of pure spectral—here denoted by $H_0(\Omega)$—, or pure temporal—$G_0(\omega)$—transfer functions, respectively. From the data (black curves) in Fig. 10 it can be seen that our averaged SMTFs (A) and TMTFs (B) corroborate the band-pass/low-pass characteristics as is typically found in comparative studies on vertebrate hearing (75).

$H_0(\Omega)$-defined SMTFs generally show a low-pass characteristic with comparable
cut-off frequencies. Spectral modulation detection is most sensitive from 0.5 up to 3 cycles/octave, with a roll-off of about 3 dB per octave. $G_\omega(\omega)$-defined TMTFs generally display a band-pass filter characteristic with a pronounced decrease in sensitivity at very low-frequency modulations (<3 Hz). Temporal modulation detection is most sensitive from 2 up to 20 Hz with a roll-off of about 3 dB per octave. Moreover, the high consistency among our averaged TMTFs ($Macaca mulatta$) and the $G_\omega(\omega)$-defined TMTFs reported by (72) ($Macaca fuscata$) and ($Macaca mulatta$) confirms that manual reaction times to dynamic-ripples provide a robust objective measure of S-T sensitivity in monkeys.

**Threshold versus Suprathreshold S-T Hearing**

The majority of meaningful biological sounds are well above hearing threshold (42, 76). Thus, it is not self-evident that the threshold-based MTF (Fig. 8A and C) provides an adequate description of how the auditory system processes S-T amplitude modulations in general. Nor is it self-evident that dynamic ripples—covering the full S-T perceptual range—are processed approximately linearly over a wide range of modulation-depths. In particular, under many conditions linear models cannot account for cortical responses of the vertebrate auditory system to ripples (13, 28, 31, 77). It is therefore of particular relevance to determine how dynamic ripples are perceived at supra-threshold modulation-depths (Fig. 12).

At supra-threshold $\Delta M$ values between $\Delta M$~16-21% in humans and between $\Delta M$~18-23% in monkeys, the respective iso-$\Delta M$ MTFs closely match the 2D-shape of threshold-MTFs (Fig. 12C). This result may be due to the approximately constant slope of the psychometric curves across ripple modulations, which resulted from our statistical analysis of estimated slopes, $\beta$, of the Weibull curves (Fig. 6A), while 85% of the thresholds, $\alpha$, were at $\Delta M$<20%.

A representation of independent spectral and temporal processing in threshold and supra-threshold acoustic regimes may explain why the S-T window of hearing in humans and macaques extends beyond the dominant modulation spectra of their own (conspecific) vocalizations (11, 12). That is, if S-T processing puts a premium on the statistical properties of natural sounds to obtain an efficient representation of spectrum and time (10, 43, 44), it is to be expected that hearing does not show an obvious...
perceptual bias towards particular (e.g., conspecific) sounds. This unifying hypothesis of S-T hearing in humans and monkeys goes against the specialization hypothesis that the auditory brain is specifically adapted to represent speech or vocalizations, in order to explain why animal brains are better at representing conspecific vocalizations than those of other animals (78). Secondary adaptations to specific behaviorally relevant sounds, leading to a hybrid hypothesis, are discussed in the following section.

**(In)separability**

The presence of both separable and inseparable S-T neurons in the auditory processing stream may underlie our finding that spectrum-time auditory perception is nearly separable (Figs 8A, 9, and 10), but still shows an additional inseparable characteristic (Fig. 11). We also found that the S-T MTFs of humans and monkeys are to a large extent spectrum/time independent, with no preference for either upward vs. downward moving ripples (as illustrated in Fig. 2, top). The dominant separable system may be expressed in separate rostral and caudal streams (79) and/or along right and left hemispheric differences in spectral and temporal sensitivity, respectively (2). Our data thus suggest that the auditory system may primarily process sounds through independent channels, corresponding to a separable system that represents the spectral and temporal eigenvectors of Fig. 10, and complemented by an inseparable system, which accounts for a higher sensitivity to S-T modulations. Neurophysiological recordings suggest a systematic increase in the percentage of inseparable STRFs of auditory neurons from midbrain IC to primary auditory cortex (28-36). Thus, although sound processing could occur in a separable way at an early level (say, up to the IC (31, 32, 36), inseparable S-T filters can still be present at higher areas and reflect their contribution to the percept as an inseparable component. Although it may seem wasteful to have both separable spectral and temporal filter banks, and tuned S-T filters, it should be kept in mind that a limited set of such filters may represent a vast acoustic world (80).

Furthermore, separable, and inseparable neural processing streams allow the system to flexibly develop highly selective and adaptive tuning to specific behavioral needs (e.g., attending to conspecific vocalizations in the presence of background masking sounds), without interfering with overall (separable) spectral and temporal...
processing and sensitivity. In this way, the perceptual response to the acoustic environment can rapidly adjust its sensitivity to meet task requirements, or optimization needs in complex, unpredictable acoustic scenes. Adaptation to new context has been demonstrated in several behavioral animal experiments (1, 81, 82) and in humans, even specifically for spectrotemporal modulations (74).

**Audition vs. Vision**

Several studies (21, 37, 81-86) have fueled the idea that the primary auditory cortex responds to dynamic ripples in a way that is analogous to responses of primary visual cortex to moving visual gratings. The point of view adopted is a representation of natural sounds and images that is consistent with efficient statistical principles to extract features of fundamental importance to the respective sensory systems (87-90). The implication is that at the cortical level the behaviorally relevant acoustic attributes of dynamic ripples —spectral modulation frequency, temporal modulation frequency, modulation depth—are represented in much the same way as those of moving visual gratings, i.e., spatial frequency, temporal frequency, luminance contrast.

A clear indication of how natural stimuli are encoded within the auditory and visual systems may be found in the way they process spectrum-time vs. space-time, respectively. Equivalence within these sensory systems would then be expressed by their (in)separability. Note that separability is a continuous variable (see Eq. 8, and Fig. 2), and the degree of separability depends strongly on the metric used to represent the data (38, 91).

To provide a fair assessment of visual and auditory sensitivity to naturalistic stimuli we therefore performed identical statistical tests on two separability indices—$\alpha_{SVD}$ and $r^2_{SVD}$—to compare separability of the spatial-temporal visual CSF (45) with the auditory S-T MTF. The top panels in Fig. 14 show the sensitivity surfaces for audition (top-left panels) and vision (top-right panel; data, with permission, from Kelly). The auditory plots run nearly parallel to the spectral-temporal axes, which is in line with a dominant spectral-temporal separability (cf. Fig. 2). The visual data, however, indicate a clear oblique orientation of the spatial-temporal sensitivity surface, which suggests strong inseparability, as reported previously (92, 93). This is further corroborated by the bootstrap analysis of the separability indices (bottom row panels),
which show that the randomly permuted visual MTF (green convex hull, bottom-right) overlaps considerably with the measured CSF (purple convex hull). In other words, the original data do not deviate significantly from those that arise by chance alone. This statistical analysis demonstrates that vision is space-time inseparable to a much larger degree than S-T sensitivity of audition. Thus, the specificity with which the auditory brain encodes natural sounds may be less stringent than the specificity needed to adequately deal with natural images.

**Figure 14. Frequency–time hearing vs. known human space-time vision.** The single surfaced shapes derived from human audition (left upper panel) and monkey (right panel) listeners represent their averaged auditory MTFs, which define the frequency-time window of S-T hearing. We used a 5th order 3D polynomial model to describe the S-T MTFs. The coefficients representing the relationship between the fitted data points of the model are provided in Table A1. For comparative analysis (vision: upper-right panel), we show the human spatiotemporal contrast-sensitivity function (CSF), defining the space-time window of vision (adapted from Kelly, 1979; see Appendix section on Visual Spatiotemporal MTF Modelling). The vertical axis of an auditory MTF represents the depth modulation sensitivity index, which is the reciprocal of the minimum perceptual amplitude modulation required to detect a rippled noise. The axes of a visual MTF, however, represent the spatial and temporal frequencies of a contrast-reversing pattern, and the observer’s contrast sensitivity. Color encodes iso-sensitivity regions with a
resolution of 10%, ranging from light-blue (10% to 20%) up to dark-red (90% to 100%). Bottom panels show inseparability ($\alpha_{SVD}$) versus separability ($r^2_{SVD}$) parameter plots in the same format as Fig. 9A, but now for normalized and pooled S-T MTF data, as shown in Fig. 8A. Note the prominent overlap between permuted—green convex hulls—and actual data—purple convex hulls—for the visual MTF, indicating inseparability of space and time in perceptual human vision.
APPENDIX

Robustness of Detection Threshold Estimation

To monitor the accuracy with which each detection threshold could be estimated throughout the recording sessions, we calculated their respective 95%-CIs and displayed this measure as a function of cumulative trial number on a log-log scale. Fig. A1 shows that the accumulation of data from subsequent recording sessions improved the estimates of the extracted thresholds in both humans (left panel) and monkeys (right panel). Notice that the data shown cover the last 14,080 trials of each monkey, and the last 7,040 trials of each human listener.

Compared to humans (≈8,600 on average), we needed 2.5 times as many responses from the monkeys (≈21,600 on average) to converge to a stable 95%-CI below 10%. A possible source for this difference is the monkeys’ higher guess rate (26% vs. 4%), along with a much greater proportion of catch trial stimuli needed to keep the monkeys under stimulus control (35% vs. 15%).

Figure A1. Solid lines represent the mean 95%-CIs of 5 x 87 ripple threshold values as a function of the number of cumulative trials for human (left: h1-h5) and monkey (right: m1-m5) listeners. The cumulative trial number can be regarded as a measure of “practice time”. Grey areas define the 95%-CI, as assessed by bias-corrected percentile bootstrap resampling (100,000 samples). Data were smoothed for illustrative purposes only. Inset boxes show the same data used to compute the solid white lines, but now the chronological order with which the data were collected was reversed. Notice that in this case (solid black lines) the same number of trials was required to converge to the same confidence levels. These graphs show that performance was stable over time, and that the decline of CIs with time resulted from the accumulation of data.
Artificially reversing the chronology with which the data were obtained did not alter this result (insets). This confirms that procedural and perceptual learning did not influence performance over time (48). Instead, the data show that the improvement in estimated thresholds over time results from an increase in the total number of responses per threshold estimation.

**Mutual Information Computation**

To compute the mutual information measure $I(A;B)$ from a pair of spectral-temporal $M(\Omega,\omega)$ matrices, we constructed a joint histogram (64), defined as a function of two variables, with $A = M_1(\Omega,\omega)$ and $B = M_2(\Omega,\omega)$. To obtain $h$, the values of $A$ and $B$, were mapped onto the $[A_{\text{min}}, A_{\text{max}}]$ and $[B_{\text{min}}, B_{\text{max}}]$, range, respectively, using equally spaced bins as determined through the interpolation algorithm of Chen and Varshney (94) (for review, see (95)). The joint probability function used in the calculation of $I(A;B)$ of a given $M(\Omega,\omega)$ pair was then obtained by normalizing $h$: 
\[ I(A; B) = \begin{cases} \sum_{a \in A} \sum_{b \in B} p_{A,B}(a, b) \cdot \log \frac{p_{A,B}(a,b)}{p_A(a)p_B(b)} \\ \text{with } p_{A,B}(a, b) = \frac{h(a,b)}{\sum_{a \in A, b \in B} h(a,b)} \\ p_A(a) = \sum_{b \in B} p_{A,B}(a, b) \\ p_B(b) = \sum_{a \in A} p_{A,B}(a, b) \end{cases} \] (A1)

Mutual information between a pair of MTFs is maximal when they have identical shapes (i.e., linearly dependent), whereas \( I(A; B) \) is zero if the MTFs are completely dissimilar.

To test the statistical significance of \( I(A; B) > 0 \), the mutual information values computed from the actual \( M(\Omega, \omega) \) measurements were plotted against those computed from randomly permuted, but shuffle-corrected (94) versions of \( M(\Omega, \omega) \)— see Fig. 10B. This was achieved by generating 100,000 bias-corrected percentile bootstrap samples of \( I(A; B) \) for both the actual and randomized data (61).

**Auditory Spectrottemporal MTF Modelling**

The human and monkey S-T MTF surfaces (audition: two top-left plots of Fig. 14) were approximated by fitting a 5\(^{th}\) order 3D polynomial model to the normalized MTFs shown in Fig. 8, applying the surface fitting tool (poly55) of MATLAB (The MathWorks, Inc., Natick, MA).

The coefficients of the 3D surfaces were found for the following polynomial equation, with \( x = \omega \) (in Hz), and \( y = \Omega \) (in cycles/octave):

\[ M(x, y) = \sum_{n,m = 0}^{5} p_{nm} \cdot x^n y^m \] (A2)

Goodness-of-fit statistics: sum of squares due to error < 5; \( R^2 > 0.88 \). The two sets of 21 coefficients are provided in Table A1. Note that the fitted surfaces only represent the modulation thresholds for downward moving ripples. The data were smoothed by a factor of 10 through bivariate linear interpolation for illustrative purposes only. Finally, to ensure a fair assessment of the degree of inseparability of audition vs. visual, the dimensions of the matrices defining the S-T MTFs (5 x 8) were identical to those defining the spatiotemporal visual MTF (5 x 8). Goodness-of-fit statistics: sum of
Comparative Primate Audition

squares due to error < 5; R² > 0.88. The two sets of 21 coefficients are provided in Table A1. Note that the fitted surfaces only represent the modulation thresholds for downward moving ripples. The data were smoothed by a factor of 10 through bivariate linear interpolation for illustrative purposes only. Finally, to ensure a fair assessment of the degree of inseparability of audition vs. visual, the dimensions of the matrices defining the S-T MTFs (5 x 8) were identical to those defining the spatiotemporal visual MTF (5 x 8).

Each coefficient $p_{nm}$ belongs to polynomial term $x^n y^m$ of Eq. A2, which includes terms up to order 5, under the constraint that $0 \leq n+m \leq 5$.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Human MTF</th>
<th>Monkey MTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p00$</td>
<td>+7.196e-01</td>
<td>+6.527e-01</td>
</tr>
<tr>
<td>$p01$</td>
<td>+6.075e-03</td>
<td>+1.407e-02</td>
</tr>
<tr>
<td>$p02$</td>
<td>-1.453e-04</td>
<td>-2.436e-04</td>
</tr>
<tr>
<td>$p03$</td>
<td>+1.118e-06</td>
<td>+1.772e-06</td>
</tr>
<tr>
<td>$p04$</td>
<td>+3.862e-09</td>
<td>-5.977e-09</td>
</tr>
<tr>
<td>$p05$</td>
<td>+5.023e-12</td>
<td>+7.646e-12</td>
</tr>
<tr>
<td>$p10$</td>
<td>+6.676e-04</td>
<td>-6.000e-02</td>
</tr>
<tr>
<td>$p11$</td>
<td>-4.487e-04</td>
<td>+2.025e-04</td>
</tr>
<tr>
<td>$p12$</td>
<td>+2.994e-06</td>
<td>-6.307e-06</td>
</tr>
<tr>
<td>$p13$</td>
<td>+1.636e-08</td>
<td>+4.354e-08</td>
</tr>
<tr>
<td>$p14$</td>
<td>+3.416e-11</td>
<td>-8.366e-11</td>
</tr>
<tr>
<td>$p20$</td>
<td>-1.059e-02</td>
<td>-3.133e-02</td>
</tr>
<tr>
<td>$p21$</td>
<td>-9.817e-04</td>
<td>-1.051e-03</td>
</tr>
<tr>
<td>$p22$</td>
<td>+5.713e-06</td>
<td>+6.308e-06</td>
</tr>
<tr>
<td>$p23$</td>
<td>-1.092e-08</td>
<td>-1.173e-08</td>
</tr>
<tr>
<td>$p30$</td>
<td>-1.743e-04</td>
<td>+2.162e-02</td>
</tr>
<tr>
<td>$p31$</td>
<td>+5.116e-05</td>
<td>-2.110e-05</td>
</tr>
<tr>
<td>$p32$</td>
<td>-1.633e-07</td>
<td>+3.456e-08</td>
</tr>
<tr>
<td>$p40$</td>
<td>+1.017e-03</td>
<td>+2.486e-03</td>
</tr>
<tr>
<td>$p41$</td>
<td>+1.103e-05</td>
<td>+3.404e-06</td>
</tr>
<tr>
<td>$p50$</td>
<td>+1.096e-05</td>
<td>-1.662e-03</td>
</tr>
</tbody>
</table>
Visual Spatiotemporal MTF Modelling

The human space-time CSF surface, $G(\alpha, \nu)$ of Fig. 14 (vision: top-right plot) was defined by:

$$G(\alpha, \nu) = \begin{cases} \kappa \cdot \nu \cdot \alpha^2 \cdot e^{-\left(\frac{2\alpha}{\alpha_{\text{max}}}\right)} \\ \text{with } k = 6.1 + 7.3 \cdot |\log(v/3)|^3 \\ \alpha_{\text{max}} = \frac{45.9}{(v + 2)} \end{cases}$$

(A3)

Here, $\alpha$ is the temporal frequency [Hz] and $\nu$ [cycles/octave] is the spatial frequency. The parameters $\alpha_{\text{max}}$ and $k$ are scaling factors. For details see page 1345.
GLOSSARY

h1-h5 human listeners one to five
m1-m5 monkey listeners one to five

CSF Contrast Sensitivity Function
FM Frequency-Modulated
MI Mutual Information
MPS Modulation Power Spectrum
MTF or $M(\Omega, \omega)$ Modulation Transfer Function
SMTF Spectral MTF
TMTF Temporal MTF
S-T SpectroTemporal

CI Confidence Interval
D Stimulus Duration
SVD Singular Value Decomposition
SPL Sound Pressure Level
RMS Root Mean Square

$S(t)$ ripple Stimulus equation $\text{Eq. 1}$
$R(t,x)$ sinusoidal Ripple amplitude envelope $\text{Eq. 2}$
$M^4$ normalized 4$th$ Moment $S(t)$ $\text{Eq. 3}$

$P(x;\alpha;\beta;\gamma;\lambda)$ Psychometric function $\text{Eqs. 4-5}$

$\lambda$ lapse rate (misses)
$\gamma$ guess rate (false positives)
$\alpha$ threshold
$\beta$ slope

$M_{\text{norm}}(\Omega, \omega)$ normalized MTF $\text{Eq. 6}$
$K(\lambda)$ singular eigenvalue matrix $\text{Eq. 7}$
$H_{\omega}(\Omega)$ Spectral MTF $\text{Eq. 7}$

$G(\omega)$ Temporal MT $\text{Eq. 7}$
$\alpha_{\text{SVD}}$ Inseparability index $\text{Eq. 8}$

$r^2_{\text{SVD}}$ statistic measure of MTF separability
$r^2_{\text{up/down}}$ statistic measure of up/down MTF symmetry

$I(A;B)$ Mutual Information $\text{Eqs. 9-10}$
$M(x,y)$ Spectro-Temporal MTF model $\text{Eq. A1}$
$G(\alpha,\nu)$ Space-Time MTF model $\text{Eq. A2}$

49
Comparative Primate Audition

Ω
ripple density [c/o]: spectral modulation frequency

ω
ripple Velocity [Hz]: temporal modulation frequency

ΔM
Modulation Depth [%] (stimulus strength psychometric curves)

c/o
cycle per octave

DATA AVAILABILITY
Data will be made available upon reasonable request.

ACKNOWLEDGEMENTS
We thank D. Heeren, S. Martens and H. Kleijnen for valuable technical assistance. We are grateful to A.M.M. Fransen for her help performing the monkey psychophysical experiments. We thank B.A. Wandell (Stanford University, CA) for supplying us with the data needed to construct the spatiotemporal MTF (top-right, Fig. 14).

GRANTS
This research was supported by the Dutch Organization for Scientific Research (NWO) ALW/VICI grant 865.05.003 and by the program “AI & Ethiek” Rotterdam University of Applied Sciences (R.W.), by ALW grant 809.37.002 (HV), and by a EU Horizon 2020 ERC Advanced Grant, 2016 Orient (grant 693400, AJVO).

DISCLOSURES
No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS
R.W., H.V. and J.O. conceived and designed research; R.W. performed the human (h1-h5) and monkey (m1-m3) psychophysical S-T MTF experiments. H.V. performed the monkey (m4-m5) psychophysical S-T MTF experiments R.W. performed the monkey hearing threshold experiments. R.W. analyzed data; R.W., H.V. and J.O. interpreted results of experiments; R.W. prepared figures and drafted the manuscript; R.W., H.V. and J.O. edited, revised, and approved final version of manuscript.
REFERENCES


21. **Schönwiesner M, and Zatorre RJ.** Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-


34. **Miller LM, Escabí MA, Read HL, and Schreiner CE.** Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J Neurophysiol* 87: 516-527, 2002. doi:[10.1152/jn.00395.2001](https://doi.org/10.1152/jn.00395.2001)


37. **Woolley SM, Fremouw TE, Hsu A, and Theunissen FE.** Tuning for spectrotemporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat Neurosci* 8: 1371-1379, 2005. doi:[10.1038/nn1536](https://doi.org/10.1038/nn1536)


58. **Kuss M, Jakel F, and Wichmann FA.** Bayesian inference for psychometric functions. *J Vis* 5: 478-492, 2005. doi:[10.1167/5.5.8](10.1167/5.5.8)


