Low-dimensional interference of mid-level sound statistics predicts human speech recognition in natural environmental noise

Alex C. Clonan[1,2,4], Xiu Zhai[5], Ian H. Stevenson[2-4], Monty A. Escabí[1-4],

[1]Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269
[2]Biomedical Engineering, University of Connecticut, Storrs, CT 06269
[3]Psychological Sciences, University of Connecticut, Storrs, CT 06269
[4]Institute of Brain and Cognitive Sciences, University of Connecticut, Storrs, CT 06269
[5]Biomedical Engineering, Wentworth Institute of Technology, Boston, MA 02115

Correspondence: Monty A. Escabí
Electrical and Computer Engineering
371 Fairfield Way U4157
Storrs, CT 06269
escabi@engr.uconn.edu

Keywords: speech in noise, cocktail party , energetic masking, modulation masking, psychoacoustics, texture statistics, sound statistics, neural model, behavioral model, natural sounds

Figure: 5 (+14 Supplementary)
Multimedia: 77 Supplementary Sound Files

Funding: This work was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under award R01DC020097. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Recognizing speech in noise, such as in a busy street or restaurant, is an essential listening task where the task difficulty varies across acoustic environments and noise levels. Yet, current cognitive models are unable to account for changing real-world hearing sensitivity. Here, using natural and perturbed background sounds we demonstrate that spectrum and modulations statistics of environmental backgrounds drastically impact human word recognition accuracy and they do so independently of the noise level. These sound statistics can facilitate or hinder recognition – at the same noise level accuracy can range from 0% to 100%, depending on the background. To explain this perceptual variability, we optimized a biologically grounded hierarchical model, consisting of frequency-tuned cochlear filters and subsequent mid-level modulation-tuned filters that account for central auditory tuning. Low-dimensional summary statistics from the mid-level model accurately predict single trial perceptual judgments, accounting for more than 90% of the perceptual variance across backgrounds and noise levels, and substantially outperforming a cochlear model. Furthermore, perceptual transfer functions in the mid-level auditory space identify multi-dimensional natural sound features that impact recognition. Thus speech recognition in natural backgrounds involves interference of multiple summary statistics that are well described by an interpretable, low-dimensional auditory model. Since this framework relates salient natural sound cues to single trial perceptual judgements, it may improve outcomes for auditory prosthetics and clinical measurements of real-world hearing sensitivity.

Main

Listening in competing noise is a central problem for both humans and animals given that isolated sounds rarely occur during real-world listening. The *cocktail party problem*, where conversational speech is relatively easily understood in a noisy social setting [1], is a classic example of the resilience of hearing in noisy scenarios. Humans often identify speech even when the power of environmental background noise exceeds the power of the speech signal (with signal-to-noise ratio, SNR, as low as ~ -10 dB). However, noisy acoustic environments create substantial challenges for individuals with hearing loss[2,3] and even the most accurate automated speech recognition systems often fail in moderate levels of environmental noise. In such scenarios, sound pressure waveforms from the foreground and background sources mix and create complex interference patterns that the auditory system must somehow segregate into multiple streams and interpret.

Spectral and temporal sound cues can impact sound segregation and listeners ability to recognize foreground target sounds, including speech. When the spectrum of a background sound overlaps the spectrum of the foreground, this can energetically mask speech making it difficult to recognize [2,4]. Temporal fluctuations in power, otherwise known as amplitude modulations, can also create interference which can either obscure [5-8] or, in certain instances, facilitate the perception of a target sound [9-11]. Additional interference can come from high-order sound structure, such as spectro-temporal modulation statistics [12,13], which convey information about phoneme- and word-level structure [14] as well as non-acoustic informational factors such as stimulus uncertainty and attention [2,3,15,16]. Each of these sound cues can vary extensively across real-world background sounds[17,18], and here we aim to develop a comprehensive computational framework that accounts for these results and predicts speech recognition accuracy for arbitrary natural background sounds.

The computations for speech recognition in noise involve multiple sound cues, and these cues are represented at distinct stages along the auditory pathway. The cochlea is highly sensitive to spectral cues[19], while auditory brainstem, midbrain and early auditory cortex [20-22], are sensitive

to spectro-temporal modulation cues, and secondary cortical areas to linguistic structure[23,24]. Invariance to background noise increases along this hierarchy[25-27] suggesting that sequential processing of multiple structured cues is necessary to segregate speech from competing noise. Here we measure speech recognition ability in human listeners under a wide range of natural background conditions. We find that differences between spectrum and modulation statistics in real world environmental sounds can produce extensive differences in interference either improving or reducing speech recognition accuracy. We then then show that a multi-stage biologically inspired auditory model accounts for the complex patterns of acoustic interference observed for real world natural sounds.

Spectral and modulation statistics can mask and unmask speech

Human participants (n=18) listened to a series of randomly chosen digit triplets (e.g., five-nine-three; zero-seven-two, etc.) in the presence of natural and perturbed environmental noises and were asked to identify the digit sequence heard (Fig. 1A; see **Methods**). Here, ten natural environmental noises (e.g., crackling fire, speech babble, machine noise etc.) and white noise (as a control) were selected to encompass a wide range of natural sound spectra and modulation statistics (Fig. 1B, shown for 8-speaker babble; other sounds in **Supplementary Fig. 1**). In the first experiment, we examined how the spectrum and modulation statistics of natural background sounds separately influenced digit recognition, either beneficially or detrimentally. In addition to presenting original backgrounds (OR, Fig. 1B), we also presented perturbed environmental background noises where the contribution of the spectrum or modulation statistics towards listening in noise was isolated. In a first perturbation, the background waveforms were phase randomized (PR) which preserves the spectrum of each individual sound but randomizes and whitens the modulation statistics (PR, Fig. 1B). For these background sounds, only the background sound spectrum influenced digit recognition. These variants preserve the original sound spectrum and, thus, the unique timbre of the original backgrounds and are each perceived as spectrum shaped noise with no clear identity (**Supplementary Audio 1**). In a second perturbation, the sound spectra of the backgrounds were equalized to follow a 1/f spectral profile[28] (spectrum equalization, SE). Since these SE backgrounds have identical spectra, only their individual modulation statistics influenced digit recognition across backgrounds (SE, Fig. 1B). Qualitatively, when listening to each of the SE variants, the original background can be easily identified (e.g., water, fire, street noise etc.), although all the SE sounds have similar timbre. That is, in contrast the PR variants, spectrum equalization preserves the spectrotemporal modulations that allow for the identification of sound category [28], so that fire sounds like fire and water sounds like water (**Supplementary Audio 1**).

Digit recognition accuracy varied substantially for different background sounds and was strongly influenced by the unique spectrum and/or modulation statistics of each background (at a fixed Signal-to-Noise ratio, SNR, -9 dB). For the OR backgrounds, accuracy varied from near 0 (e.g., 4.4% for babble-8) to near 100% (e.g., 96.2% for white noise, fire, Fig. 1C, black) and perturbing the spectrum (SE, yellow) or modulation (PR, red) statistics could drastically modulate recognition accuracy. For instance, although the OR jackhammer noise showed moderately high accuracy (65.3%), the PR jackhammer noise (Fig. 1C, red) had substantially lower accuracy (28.9%). Since these two sounds have identical spectra, the modulation statistics in the OR jackhammer noise appear to have a beneficial unmasking effect on digit recognition. The spectrum-equalized jackhammer background (Fig. 1C; SE, yellow), which contains the original modulation statistics but has a whitened sound spectrum, had a high accuracy (73.1%), similar to

the OR condition. In stark contrast, the modulation statistics of eight speaker babble had a detrimental effect on digit recognition. Here, the OR eight speaker babble resulted in the lowest digit recognition accuracy of all backgrounds tested (4.4%). However, accuracy was substantially higher (47.1%) after whitening the modulation spectrum of the eight-speaker babble (PR condition). On the other hand, the SE variant, which retained the modulation statistics, resulted in a low accuracy (5.8%) similar to OR implying that the modulation statistics strongly drove the perceptual interference. Across background sounds, both the PR and SE manipulations produced changes in accuracy that were highly dependent upon the specific background sound. These trends were conserved and highly consistent across participants (across participant $r = 0.91 \pm 0.04$, mean $\pm$ sd) suggesting a common pattern of spectral and/or modulation interference.

Differences in the spectrum and modulation statistics of individual background sounds led to either masking or unmasking of speech digits (Fig. 1D). Here, changes in accuracy between OR and PR were used to assess modulation masking effects whereas changes in accuracy between OR and SE allowed us to isolate spectral masking effects. Backgrounds such as the freeway, bird babble, fire, and white noise had greater accuracy in the OR compared to the SE configuration, indicating that the original spectrum of these sounds had a beneficial effect. By comparison, for city and factory noise accuracy was higher for the SE compared to the OR condition, suggesting that the original sound spectra had a detrimental effect on digit recognition accuracy. Yet for other sounds, such as eight-speaker babble and the jackhammer noise, OR and SE background had similar accuracy, but OR and PR backgrounds differed, indicating effects due to modulation statistics. As with the spectral effect, the presence of the original modulation statistics could either be beneficial (e.g., jackhammer, PR<OR) or detrimental (e.g., babble-8, PR>OR) to accuracy. In each case, these effects were highly clustered across participants (Fig. 1D) suggesting that the unique spectrum or modulation statistics had a consistent impact on digit recognition.

In a second experiment, we further explored how background sound statistics contribute to recognition in noise by using texture synthesis [29] to systematically manipulate the background statistics. For these tests, we selected eight-speaker babble (n=10 participants) and jackhammer noise (n=6 participants) because they exhibit opposing accuracy trends driven by the modulation statistics (Fig. 1C, D). We generated synthetic variants of the babble and jackhammer backgrounds by sequentially incorporating summary statistics and repeated the digit recognition task with the synthetic backgrounds at SNRs -12 to -3 dB. Starting with a synthetic variant that strictly incorporated the spectrum statistics (Spec; similar to PR), backgrounds were sequentially manipulated to include marginal (+Mar), modulation power (+Mod), and correlation statistics (+Corr) (Fig. 2; A2-A4, B2-B4). Each of these summary statistics has unique structure that produces increasingly more realistic percepts as they are incorporated in the synthetic variants [27-31] (**Supplemental Audio 2**). For the spectrum and marginal statistics (envelope mean and variance) both background sounds have similar structure (Fig. 2, A2 vs. B2)[27]. For the jackhammer, the modulation statistic (Fig. 2, A3; the sound power at specific frequency and modulation frequency combinations) is consistent with several audible components, particularly a high frequency (>~2 kHz) dominant jackhammer sound with ~20-30 Hz modulation frequency and a weaker low frequency engine noise with modulations > 10 Hz. By comparison, speech babble has greater energy for modulation frequencies < 10Hz (Fig. 2, B3) due to the slow rhythmic fluctuations present in speech [30,31] and also contains strong high frequency modulations (>~100 Hz) that are generated by voice periodicity[18]. Finally, the correlation statistics (between envelopes of each frequency channel) differ, with jackhammer noise having somewhat broader correlations across frequency channels.

Digit recognition accuracy deteriorated upon adding high-order summary statistics to the synthetic eight-speaker babble backgrounds; however, and surprisingly, it improved upon adding summary statistics to the synthetic jackhammer (Fig. 2, C1-C4). For instance, at an SNR of -9 dB adding summary statistics to the synthetic eight-speaker babble reduced the digit recognition accuracy from 45.6% to 14.8%, while for jackhammer noise, adding statistics improved accuracy from 39.3% to 66.0% (Fig 2, C2). These effects appear to be largely independent of SNR, and the accuracy trends were well described by independent functions of SNR and summary statistics (separability index=0.98 $\pm$0.01, mean$\pm$sd; see **Methods**). The improvement in recognition when adding statistics to the jackhammer noise may result from spectrally correlated fluctuations seen at ~20 Hz (see Supplementary Fig. 1), since comodulation can in some instances unmask the perception of a target sound [9,11]. The deterioration in recognition when adding statistics to the eight-speaker babble may reflect the fact that the spectrotemporal modulations in babble overlap closely with those of the speech digits. Once again, the perceptual trends were highly consistent across participants (across participant r=0.95 $\pm$0.02, mean$\pm$sd) suggesting that a common set of perceptual strategies and cues influence recognition. Collectively, these results show that layering different levels of high-order acoustic structure in synthetic background variants can impact digit recognition, either beneficially or detrimentally.

Low-dimensional interference of mid-level sound statistics predicts perceptual sensitivity

Using the perceptual observations above, we next asked whether summary statistics derived from a biologically inspired auditory model can predict human recognition behavior. Recent studies support the view that hierarchical computations performed by peripheral and mid-level auditory structures are critical for the perception of environmental sounds [28,32]. Here we propose a model based on the transformations that occur, first, at the cochlea and, subsequently, in mid-level auditory structures (inferior colliculus) [18] and use this model to predict speech in noise behavior with natural background sounds. Sounds are first decomposed by frequency, generating a cochleogram representation (Fig. 3A). Subsequently, sounds are decomposed through a set of modulation selective filters that model the decomposition of sounds into spectro-temporal components in auditory midbrain. While the time-average power spectrum distribution of the cochelogram (the cochlear power spectrum, cPS; Fig. 3, B2-D2) accounts for the distribution of sound power across frequencies, the midbrain modulation power spectrum (mMPS; Fig. 3, B3-D3) generates a multi-dimensional distribution of the sound power in the midbrain feature space. Here, sounds are represented as a function of frequency, and temporal and spectral modulation components. Speech digits and natural backgrounds not only differ in the spectral content, as seen in their cPS (Fig. 3, B2-D2), but they also have unique mMPS that can be similar (2-speaker babble, Fig. D3) or quite different (bird babble and jackhammer; Fig. 3, C3 and B3) from speech. Such differences in the power distributions between the foreground and background sounds has the potential to produce distinct patterns of acoustic interference. For instance, whereas the 2-speaker babble sound has high power that interferes with speech digits primarily at low frequencies (D2), the bird babble background sound interferes predominantly at high frequencies (B2), and the jackhammer sound interferes across the entire cochlear spectrum range (C2). While the cPS only accounts for frequency-based interference, the mMPS offers insight into how modulation components might interfere. For example, 2-speaker babble contains substantial power for temporal modulations within the vicinity of 4 Hz (D3, broadband 4 Hz peak visible at 0 cycles/octave spectral modulation) that have the potential to interfere with similar temporal modulations in the speech digit sequences. The jackhammer sound, by comparison, contains

repetitive temporal fluctuations that produce a high-power band within the vicinity of ~20-30 Hz temporal modulation and which are not present in the speech sequences (C3). These foregrounds and backgrounds also contain strong fine-structure fluctuations (visible for high temporal and low spectral modulations) that have potential for interfering at low frequencies (<1 kHz). The unique spectro-temporal modulations of each background have the potential to interfere with the modulations in speech to varying degrees and, thus, generate differences in recognition accuracy.

We model listener accuracy on single trials by fitting a logistic regression model (see **Methods** and **Supplementary Fig. 3**). By considering either the cPS or mMPS of both the foreground and background as the model input (Fig. 4 A1 and A2, respectively), we optimize the model parameters to predict the probability of "correct" (1) and "incorrect" (0) responses. The model coefficients identify background and foreground sound features that improve (positive coefficients) or reduce (negative coefficients) digit recognition accuracy. This approach is analogous regression methods that are widely used to estimate neural receptive fields and which map sensory features to neural spikes [33,34] (0s or 1s indicating presence or absence of a neural response). Here, our approach uses perceptual judgments in lieu of neural spike trains. Based on our perceptual data, we expect that the cPS representation may be able to describe masking and unmasking related to spectral differences (e.g., OR vs SE), while the mMPS representation may be necessary to account for modulation effects (e.g., OR vs PR and differences based on high-order statistics).

We first tested whether interference in the cochlear auditory space can account for differences in digit recognition. Here the cPS of both the foreground and background serve as independent inputs to a logistic regression model designed to predict recognition accuracy (Fig. 4, A1). Conceptually, the background and foreground sound power at each frequency are independently weighted by the model so that the foreground and background confer an opponent-weighted contribution to the predicted perceptual accuracy. For instance, when the background power at a critical frequency exceeds that of the foreground power, the predicted accuracy should be reduced. By comparison, increasing the foreground power at the same frequency should produce an increase in predicted accuracy. That is, the foreground and background cPS should produce an interference pattern with antagonistic influence on the predicted accuracy. To minimize the complexity of the model and possibility of overfitting, we used a low-dimensional representation of the cPS (20 dimensions; see **Methods**) and the model weights were optimized and predictions cross-validated (25-fold) for each participant (**Supplementary Fig. 2**). Results for a single participant demonstrate that the cochlear model has some predictive power and accounts for coarse sound dependent trends. The average accuracy predicted by the model for each background condition has a strong correlation with the observed accuracy of each condition (Fig. 4B1, C1; single participant $r^2 = 0.6 \pm 0.1$, mean $\pm$ sem). The model captures coarse trends observed for the OR and PR conditions (Fig. 4, B1; left and middle panels; dots are the participant accuracies) and, compared to the behavioral data, it has a relatively narrow output dynamic range (12-85 % versus 0-100 % for the behavior). Furthermore, it fails to capture detailed changes in accuracy when the sound spectra are equalized (SE condition; Fig. 4B1, right panel) and it does not differentiate OR and PR responses, since it produces similar predictions for both (Fig. 4B1, left and middle; since OR and PR sounds have identical cPS).

We next tested whether multi-dimensional interference in a mid-level auditory space better explains the perceptual findings (Fig. 4 A2). Here, a low-dimensional representation of the mMPS is again used as input to the logistic regression model, which limits the model complexity and minimizes overfitting (optimal solution, 24 dimensions; see **Methods**). For the same participant,

there is a marked improvement in the predicted accuracy for the mMPS input (Fig. 4, B2, C2; single participant $r^2 = 0.87 \pm 0.05$, mean $\pm$ sem) compared to the cPS. In addition to accurately predicting the OR and PR conditions, the mMPS input also captures modulation dependent changes in accuracy for the SE condition (Fig. 4B2, right panel). Across the sample of participants (n=18), the mid-level model outperforms the cochlear model and explains nearly 90% of the behavioral variance (Fig. 4D-F). The predictive power of the mid-level model is significantly higher than the cochlear model (Fig. 4E; average $r^2 = 0.89 \pm 0.01$ vs. $r^2 = 0.63 \pm 0.01$, mean $\pm$ sem; Fisher z $-$ transform paired t $-$ test, $t(17) = 20, p < 10^{-12}$) and this was true even when the mid-level model was constrained to the same dimensionality as the cochlear model (20 dimensions, average $r^2 = 0.85 \pm 0.01$, mean $\pm$ sem; Fisher z $-$ transform paired t $-$ test, $t(17) = 15.7, p < 10^{-10}$).

The logistic models also provides accuracy predictions for individual trials, and the mid-level representation can account for accuracy differences, not just across conditions (e.g., white noise OR vs babble-8 OR), but also within each condition (e.g. two different trials with babble-8 OR noise). Subtle differences between specific foregrounds and differences between multiple instances of the same background sound may have predictable effects on accuracy that are not apparent when all the trials within a condition are averaged. As a reference, we fit a *category-specific Bernoulli* model which uses the average accuracy of each sound condition on training data to predict test trials (33 parameters for OR, PR, and SE of each of the 11 background sounds). While this category-specific model lacks trial-specific information (null hypothesis), the mid-level model uses acoustic information from each trial, which could potentially account for trial specific changes in accuracy. For a single participant, the mid-level model receiver operating characteristic (ROC) curve is above those of both the category-specific and cochlear models (Fig. 4D), suggesting an improvement in single trial prediction quality across all trials. Across participants, the area under the ROC curve (AUC) for the mid-level model (Fig. 4F; 0.79 ± 0.01, mean $\pm$ sd) was significantly larger than the AUC of the category-specific and cochlear models (Fig. 4F; 0.75 ± 0.01 and 0.73 ± 0.02, mean $\pm$ sd; t(17)=12, $p<10^{-9}$ and t(17)=16.5, $p<10^{-11}$, paired t-test). Furthermore, while the categorical model has a sensitivity index (*d'*) of 0 for single-trial predictions within background conditions, the mid-level model *d'* was significantly larger than the categorical model *d'* when averaged across sound condition (participant average *d'*=0.31 ± 0.03, mean$\pm$sem; t-test, t(17)=11.6, $p<10^{-8}$; **Supplementary Fig. 3**). Thus, the mid-level model not only outperforms the cochlear model in predicting differences between background conditions, but it also accounts for trial-by-trial changes in perceptual accuracy within background conditions.

To explore the source of acoustic interference between speech and natural background sounds further, we inspected the model weights and projected them onto frequency and temporal modulation dimensions of the auditory model. The foreground (Fig. 4G1) and background (Fig. 4G2) weights have an antagonistic relationship that depends both on the sound frequency and modulation content. For the background, model weights show strong broadband negativity centered at ~1 kHz and very slow temporal modulation frequencies (~4-8 Hz; Fig. 4G2). Backgrounds that contain substantial power within this modulation band, as seen for 2-speaker babble (Fig. 3 D3, ~4-8 Hz), interfere more strongly with speech leading to reduced recognition accuracy. By comparison, at high frequencies (>0.8 kHz) and for temporal modulations between ~16-64 Hz, the background model weights are positive, suggesting that background energy for these components enhances recognition accuracy. This range likely reflects a beneficial interference consistent with comodulation masking release [9-11,35], as observed for the jackhammer which contains strong comodulated fluctuations that overlap this acoustic range (Fig. 3 C3) and

which has reduce accuracy for the PR manipulation (Fig. 2; accuracy enhancement in Fig. 3 when statistics are included). For the foreground sound, the model weights show a positive component within the vicinity of 4 Hz temporal modulation that opposes the negative component in the background and increases accuracy for foreground speech sequences with strong temporal modulations in this range. Collectively, the pattern of mid-level model weights suggests that there is multi-dimensional interference between foreground and background sound cues and that background statistics can either facilitate or hinder word recognition.

SNR contributes to speech recognition independently of spectro-temporal features

Beyond the statistical structure of natural sounds, background sound power can vary extensively in real world listening environments, which influences the listening SNR and ultimately the perceptual difficulty. Although lower SNRs clearly make it more difficult to hear and identify sounds, how the SNR interacts with speech perception for specific natural background sounds and their related statistics is unclear. Here we explored how SNR impacts speech recognition across natural background sound conditions.

Nine participants listened to digit triplets embedded in the non-perturbed, natural backgrounds (OR) at variable SNRs (-18 to 0 dB) (see **Methods**). The paradigm is challenging (average accuracy 55%), but covers a broad range of perceptual accuracies across both SNR and backgrounds, extending from very easy (e.g., 100 % for white noise at SNR=0 dB) to extremely difficult (e.g., 0 % for babble-8 at SNR=-18). As shown for a single participant, digit recognition accuracy increased with increasing SNR for all background sounds tested in a sigmoidal fashion with SNR (Fig. 5A). However, the intercept (inflection point of the sigmoid) varied across backgrounds, consistent with the fact that the spectrum and modulation statistics of each background drastically impact listening difficulty even at a constant SNR (as in Fig. 1, -9 dB). To quantify these differences, we estimated the SNR where digit recognition accuracy is 50% for each background (the logistic midpoint). Across participants (n=9, Fig. 5E), backgrounds that interfered strongly with foreground speech, such as the eight-speaker babble, had a high logistic midpoint SNR (~ –3 dB for 50% accuracy). By comparison, backgrounds that do not extensively overlap with the foreground speech in either the spectrum and/or modulation statistics, such as bird babble, interfered less with speech perception and thus showed a much lower logistic midpoint SNR (~ –16 dB at 50% accuracy). Thus, to maintain 50% accuracy across background sounds, the SNR needs to be adjusted by as much as a 13 dBs. These general trends were highly consistent across listeners as noted by the relatively narrow standard deviation of midpoint distributions (Fig. 5E) and high inter-participant correlation ($r = 0.95 \pm 0.005$, mean$\pm$sem).

To determine whether the summary statistics of the background sound and SNR contribute independently to perceptual accuracy, we computed a separability index for the SNR across background trends (see **Methods**). The separability index was near 1 (0.93 for participant of Fig. 5A; population average=0.93 ± 0.01, mean $\pm$ sd) consistent with the hypothesis that SNR makes an independent contribution to masking. We additionally modified the logistic models described above by incorporating SNR as an independent input variable. The mid-level model incorporates the same spectrotemporal features used previously, but adds a single independent regression weight to account for the SNR dependency of all the backgrounds (see **Methods**). This model accurately reconstructs digit recognition accuracy trends across backgrounds and SNRs for a single participant (Fig. 5B, C; $r^2 = 0.95 \pm 0.01$, mean $\pm$ sem). ROC analysis further confirms that, for this participant, the mid-level model is comparable to a category-specific Bernoulli model (Fig. 5D; AUC=0.92 vs. 0.91) in predicting single trial outcomes across all conditions. Across

participants, the model predicts more than 90% of the perceptual variance ($r^2 = 0.93 \pm 0.005$, mean $\pm$ sem), accurately replicates the logistic-midpoint distributions (Fig. 5 E vs. F), and slightly outperforms the categorical model (Fig. 5G; AUC=0.90 ± 0.005 vs. 0.89 ± 0.005; mean $\pm$ sd; paired t-test, t(8)=8.3, $p<5x10^{-5}$). Together, these results suggest that spectrotemporal sound features and SNR make independent contributions to digit recognition accuracy and demonstrate that a modified mid-level acoustic model can accurately explain human performance at across noise levels.

Discussion

Our results suggest that speech recognition in competing natural background sounds involves the interplay of multiple summary statistics. For natural sounds, these summary statistics can be modeled through a biologically inspired auditory model, which accounts for both the spectral decomposition of the cochlea and the modulation decomposition in mid-level auditory centers[18]. Here we find that a low-dimensional representation of these natural sound statistics accounts for perceptual findings. Importantly, our behavioral and modeling results suggest that sounds statistics influence recognition independently of the noise level (SNR) and that complex masking and unmasking phenomena can be described through interpretable perceptual transfer functions.

The mid-level auditory model is a simplified representation of the computations carried out by the auditory system and it likely does not provide an exact representation of the mid-level neural statistics contributing to recognition in noise. Neural summary statistics from auditory midbrain are only partially correlated with the sound statistics that can be measured from our model[28], yet they can accurately predict human perception of sound textures. Our primary conclusion and claim is that average statistics from features homologous to those encoded in mid-level auditory system can explain the multi-dimensional interference between speech and natural background sounds. As seen for the human data, spectro-temporal modulation statistics of the competing background sound strongly impact recognition accuracy and the comparison between the cochlear and mid-level model predictions demonstrate a clear improvement for the mid-level representation. Whereas the cochlear model relies strictly on frequency-based interference to predict behavioral outcomes, the modulation-based interference in the mid-level model more accurately predicts perceptual outcomes. The mid-level model transfer functions suggest that slow fluctuations in background sounds in the vicinity of 4 Hz are a critical source of interference that competes with comparable fluctuations in speech. This is consistent with prior observations indicating that cortical activity for slowly varying non-stationary sounds is stronger than for stationary backgrounds[25] and may thus interfere more strongly with speech. The mid-level transfer functions also identify frequency and modulation specific enhancements in recognition accuracy (observed for frequencies >0.8 kHz and temporal modulations ~16-64 Hz) that is consistent with comodulation masking release phenomena [9-11,35].

The source of interference captured by the model is strictly acoustic and other factors beyond summary statistics also influence speech recognition in noise, particularly during conversational speech perception. For instance, lexical and semantic top-down signaling influence speech perception during continuous speech[36-38]. Our perceptual paradigm intentionally avoids such influence by using randomly ordered digits from randomly selected talkers, thus focusing exclusively on the acoustic features underlying masking. Our model also does not account for attention and informational factors such as varying degrees of stimulus uncertainty [15,16] which also influence speech recognition in noise. By adding temporal structure and additional linguistic

variables to the logistic model introduced here, future studies may be able to account for continuous speech recognition in natural environmental noise under a wider range of scenarios.

There is a long history of cognitive models that attempt to explain human psychometric performance yet, traditionally, these have been applied to the perception of relatively simple synthetic stimuli (e.g., visual or auditory) and their parameters. Recently, several task optimized deep neural network models have been developed that can recognize or categorize natural sensory stimuli and, interestingly, these networks can exhibit human like performance in both auditory and visual tasks [39,40]. Attempts to train neural networks to directly predict human perceptual data for image categorization can also yield networks that approach human perception [41]. For listening tasks, such neural networks models can recognize individual words in competing noise with human-like accuracy despite the fact that they are not trained directly on human perceptual data[39], yet, they often underperform in real environments or when exposed to novel datasets (e.g., **Supplementary Fig. 4 A-B**). Furthermore, such neural network models appear to exhibit task-specific invariances that diverge from those of human listeners[42], suggesting that their computational strategies can be quite different from humans (e.g., **Supplementary Fig. 4 C-D**). Thus, despite their success, interpreting these complex neural network models and their possible relationships to cognitive and brain processing mechanisms remains challenging since neural networks are extremely nonlinear systems that typically contain tens-of-thousands to millions of parameters. An alternative modeling strategy is to consider biologically inspired models of perception which have fewer parameters and greater interpretability. These have yielded quality metrics that correlate with perceptual preferences[43-45] and are widely used to benchmark audio and video processing technologies. Our approach differs from and extends these approaches since it combines a biologically inspired model feature space with a relatively simple and optimal quasi-linear mapping to predict human single trial behavioral judgments for arbitrary sounds. Here, unlike neural networks, which are trained to predict individual sound or visual categories, the model predicts correct and incorrect judgements which greatly reduces the output dimensionality. Following optimization and dimensionality reduction, the model accurately explains trial-by-trial behavioral responses with a low-dimensional representation of the auditory model space, thus requiring few model parameters. Furthermore, the resulting model weights are highly interpretable and can be viewed as perceptual transfer functions that map the model-based acoustic features to behavioral accuracy.

Overall, the findings demonstrate that recognition accuracy in a complex real world behavioral task can be mapped onto a low-dimensional and quasi-linear representation of a biologically inspired model feature space. This generalized model framework yields an interpretable and plausible neural inspired representation that provides insights into the features and neural computations that underlie perceptual judgments. Thus, it can theoretically be applied to a variety of behavioral tasks across various sensory modalities and has the potential to be extended to animal behavior and model performance evaluation. For the speech in noise task, the model reveals multi-dimensional features and transformations underlying interference between foreground and background sounds. Since this model can make predictions for arbitrary background and foregrounds sounds, it may be useful for studying a wide range of complex real world listening scenarios and may allow for rapid clinical diagnostics that can identify real world listening deficits.

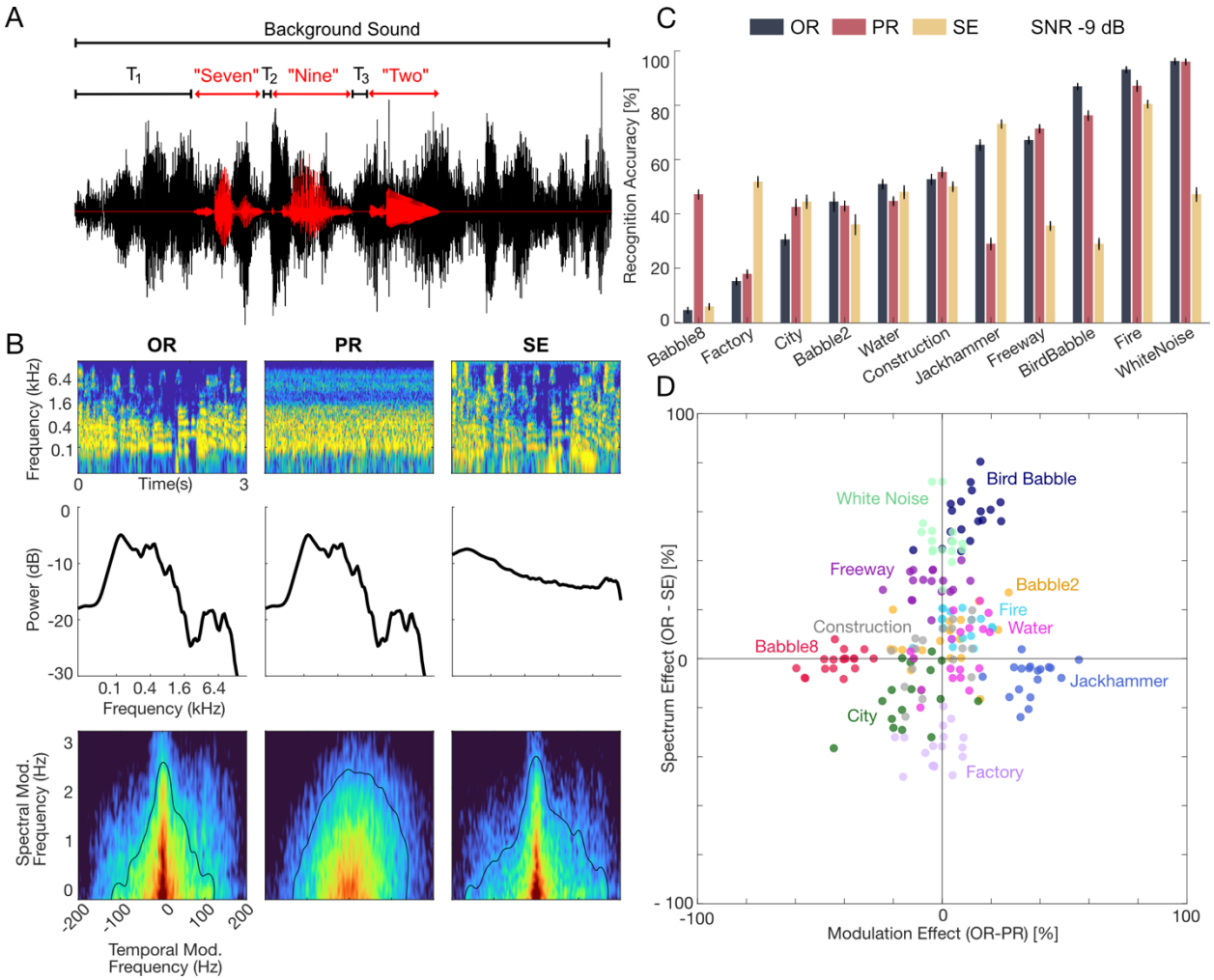


Figure 1. Experiment 1, The contribution of spectrum and modulation statistics to digit recognition in natural environmental noise. (A) Speech digits were delivered to human participants in competing natural background noises at -9 dB SNR (11 backgrounds in three configurations). The intervals between digits (T_1, T_2, T_3) were randomly jittered to minimize learning and leveraging the word timing. Participants were required to identify the digit sequence (e.g., Seven-Nine-Two). (B) Each of the backgrounds was delivered in the original (OR), phase randomized (PR), and spectrum equalize (SE) configurations (illustrated for the speech babble; other sounds shown in **Supplementary Fig. 1**). Cochleograms, cochlear power spectrum (cPS), and modulation power spectrum are shown in the 1st, 2nd and 3rd rows. The PR condition preserves the original spectrum of the sound (middle panel, 2nd row) but whitens the modulation power (middle panel, 3rd row). Alternately, the SE condition whitens the sound cPS, but retains much of the original modulation content. (C) Average results across n=18 participants shows digit recognition accuracy in the different background sound conditions (rank ordered according to OR accuracy; OR=black, PR=red, SE=yellow). Removing the modulation (PR) or spectral (SE) cues can drastically change recognition accuracy, indicating that both cues contribute to masking. (D) Change in accuracy between OR and PR condition are indicative of a modulation masking effect while change between OR and SE are indicative of a spectral masking effect. Recognition accuracy shows clustering across participants for different backgrounds (color coded) indicating a predominant and sound dependent modulation (ordinate; e.g., babble-8, jackhammer) or spectral (abscissa; e.g., bird babble, white noise, or factory) masking effect. The spectral and/or modulation statistics in OR could either improve (e.g., jackhammer, bird babble) or reduce accuracy (e.g., factory, babble-8).

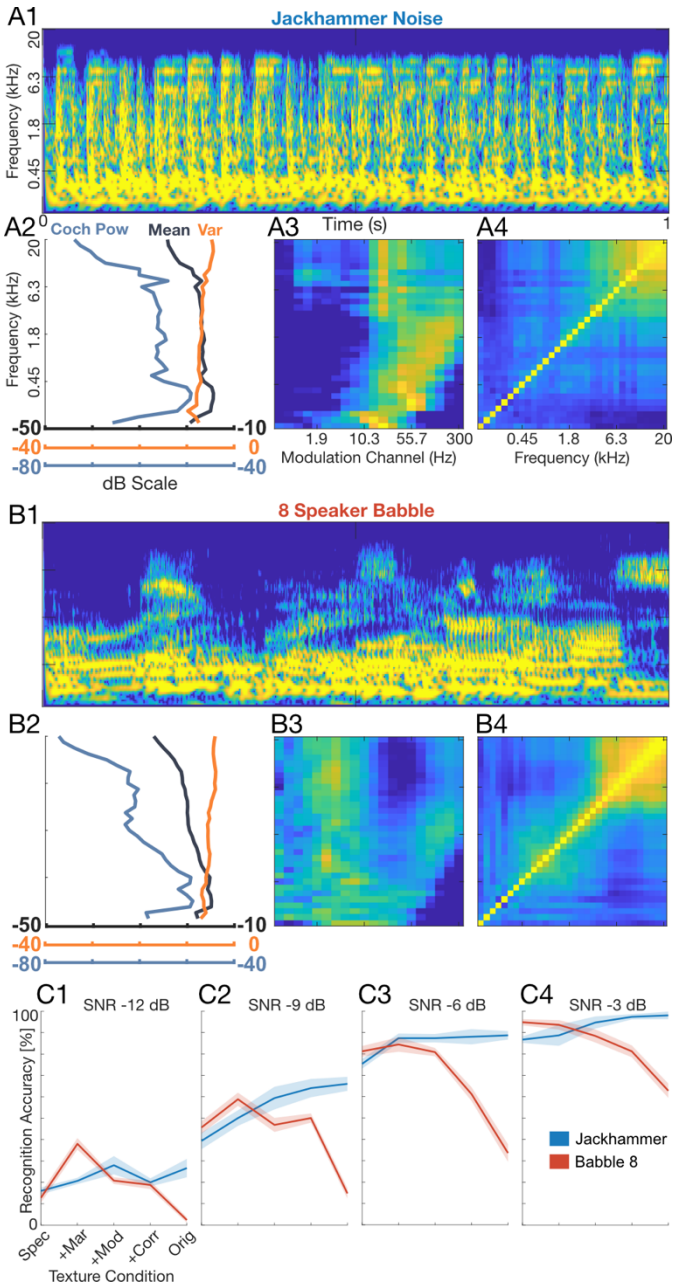


Figure 2. Experiment 2, Modulation summary statistics influence digit recognition accuracy. Sounds were delivered as in Experiment 1, but here, texture synthesis[29] was used to manipulate individual summary statistics for the jackhammer (A1-A4) and babble-8 (B1-B4) backgrounds. The cochleograms are shown in A1 and B1, along with the marginal (Mar; A2, B2), modulation power (Mod; A3, B3) and correlation (Corr; A4, B4) statistics which show unique structure for both sounds. Although Mar statistics are similar for both sounds (A2, B2), the Mod statistics show that speech contains extensive low frequency modulations (<10 Hz) along with fast temporal modulation arising from vocal fold vibration (B3; >55 Hz). Speech babble shows correlations across frequency channels (B4) but these are somewhat more compact than the jackhammer sound (A4). The jackhammer contains limited Mod power for low frequencies (<10 Hz) and contains a strong comodulated component at ~20 Hz for frequencies > ~2kHz and extensive correlations across frequencies (A4). (C1-C4) Digit recognition accuracy increases systematically with SNR, but also varies with the included summary statistics. For the speech babble (n=10 participants), accuracy decreases upon adding high-order summary statistics indicating that modulation structure in speech babble interferes with digit recognition. By comparison, for the jackhammer (n=6), including high-order summary statistics improves digit recognition accuracy indicating a beneficial unmasking. Error bars indicate SEM.

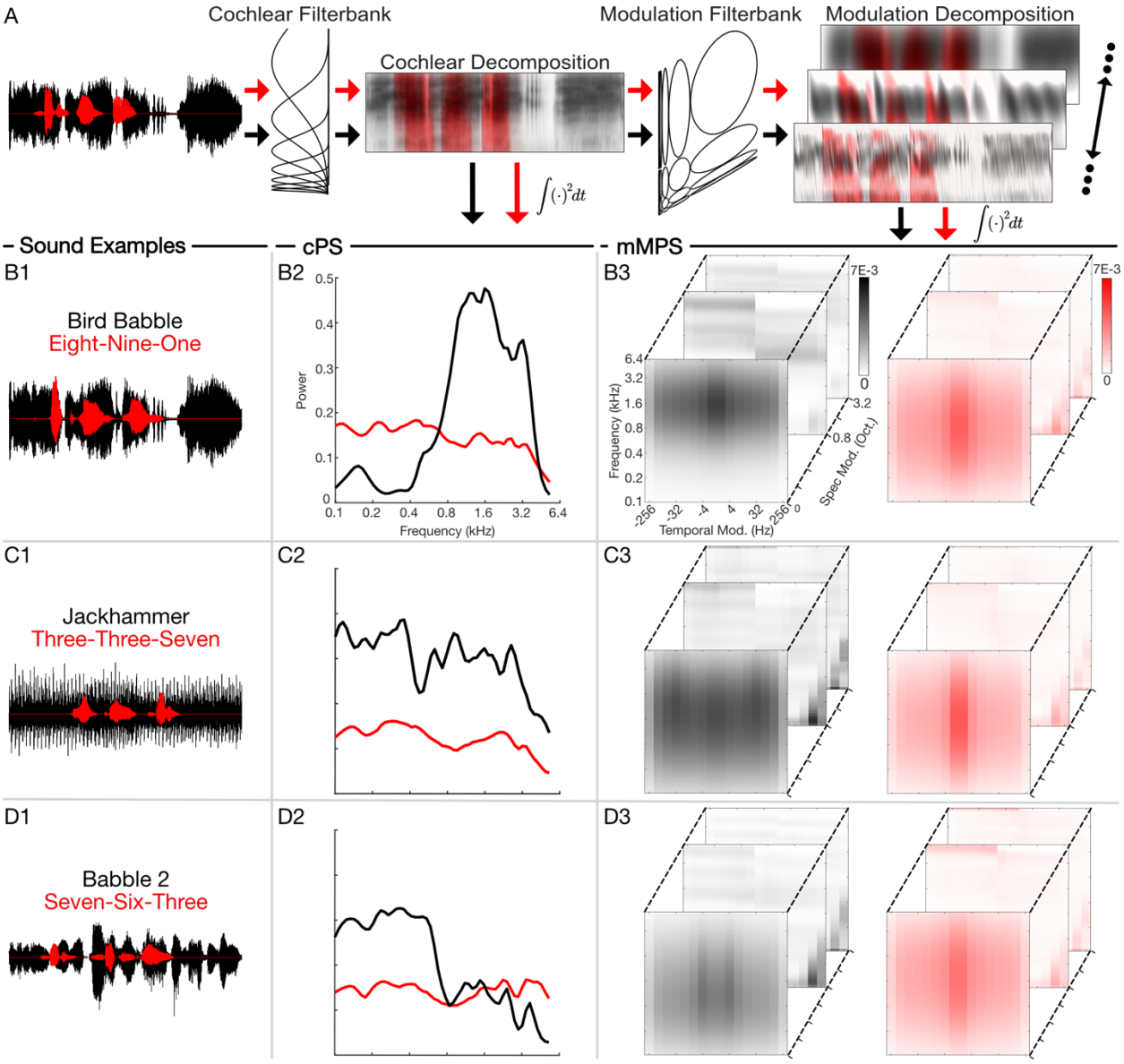


Figure 3: Cochlear and mid-level auditory feature representations. (A) Both, the foreground (red) and background (black) sounds are first decomposed into frequency components through a cochlear auditory model. These cochlear model outputs are then passed through a mid-level filterbank which further decomposes the sounds into time-varying spectral and temporal modulation components. (B-D) The time-averaged output power yields power spectrum distributions in either the cochlear or mid-level auditory space. The resulting cochlear power spectrum (cPS; B2-D2) and mid-level power spectrum (mMPS, B3-D3) are illustrated for three representative sound mixtures (B1-D1). Whereas the cPS only captures foreground and background sound power distributions as a function of frequency, the multi-dimensional mMPS distribution represents the sound as a function of frequency, spectral, and temporal modulation frequencies.

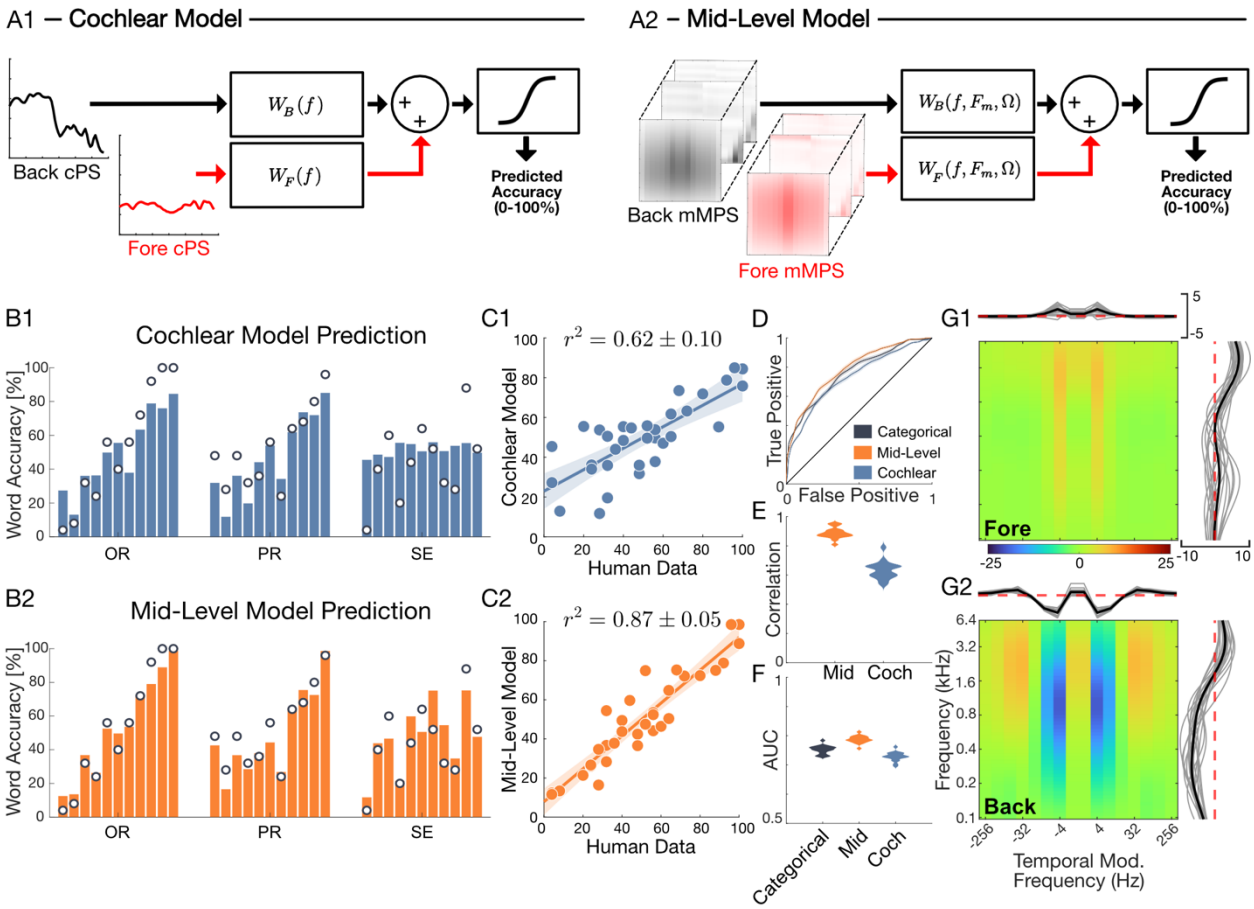


Figure 4: Mid-level auditory model predicts digit recognition accuracy in natural environmental noise. (A,B) Comparing digit recognition accuracy predictions for the cochlear (A1, B1, C1) and mid-level (A2, B2, C2) auditory models for a single participant show a vast improvement for the mid-level auditory representation. Bar plots (B1) show the cochlear model predicted accuracy for each of the three conditions of Experiment 1 (OR, PR and SE; background sounds are ordered as in Fig. 1C, from Babble-8 to White Noise). The predicted accuracies differ substantially from the actual measured accuracy for this participant (B1, superimposed dots; C1, $r^2 = 0.6 \pm 0.1$), especially for SE condition (B1, far right) where masking differences are attributed solely to the modulation content and the cochlear model produces a near constant prediction. (B2, C2) The mid-level model accounts for the modulation content of the sound and produces vastly improved predictions (C2, $r^2 = 0.87 \pm 0.05$). (D) The mid-level model ROC curve (participant from B1, B2) is higher than both the cochlear and categorical Bernoulli models. (E and F) Population results (n=18) confirm that the mid-level model predicts a higher proportion of the perceptual variance than the cochlear model (C; $r^2 = 0.89 \pm 0.01$ vs. $r^2 = 0.63 \pm 0.01$) and it outperforms both the cochlear model and categorical Bernoulli model on single-trial predictions as indicated by the higher AUC (F). (G1, G2) The mid-level model regression weights can be interpreted as perceptual transfer functions. Here, the foreground (G1) and background (G2) model weights show antagonistic structure whereby specific frequencies and modulation components in the foreground and background can compete to improve or reduce digit recognition accuracy. Notably, mid frequency (~0.4-3.2 kHz) and slow modulation background components (~4 Hz) tend to reduce accuracy (negative weights), while broadband speech driven components around ~4 Hz show positive weights that can improve digit recognition accuracy.

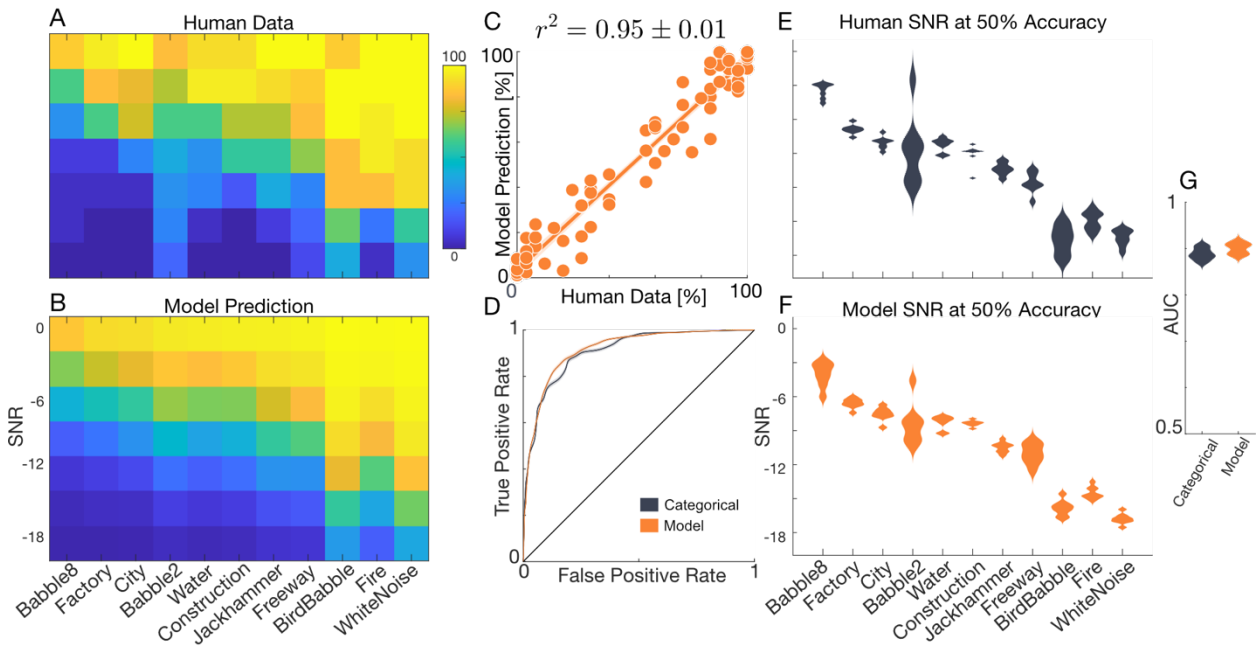


Figure 5. Experiment 3, SNR influences recognition independently of summary statistics. (A) Average accuracy for one participant is shown as a function of SNR and each of the eleven OR backgrounds. (B) Predictions using a modified mid-level model with SNR as an independent input variable accurately predict the participants digit recognition accuracy. (C) Scatter plot comparing the actual (A) versus the modified mid-level model predicted digit recognition accuracy (B). Across participants, the modified mid-level model explains more than 90% of the behavioral variance ($r^2 = 0.93 \pm 0.005$, mean $\pm$ sem). (D) ROC curve for the modified mid-level and a 77-parameter categorical Bernoulli model for the same participant as panels A and B. The distribution of logistic midpoints across nine participants and background sound conditions (E) is accurately predicted by the modified mid-level model (F). The AUC for the modified mid-level and categorical Bernoulli models (G; AUC=0.90 $\pm$ 0.005 vs. 0.89 $\pm$ 0.005; mean $\pm$ sd).

Methods

Human Psychoacoustics

All procedures were approved by the University of Connecticut Institutional Review Board. Participants were provided verbal and written description of the experiment procedures and study rationale prior to consenting to participate in the experiments. We recruited male (n=9) and female (n=18) native English speaker participants ages 18-43 with normal hearing sensitivity (>20 dB HL for left and right ears; 0.5, 2, 8 kHz, tested with MA-42 audiometer (Maico Diagnostics Co.). Experiments were carried out in a sound shield room (Industrial Acoustics Co.) at the University of Connecticut, Storrs campus. Sounds were delivered at 44.1 kHz sampling rate using an Fireface UC digital audio interface (RME Audio Inc.), amplified through a BHA-1 headphone amplifier (Bryston Ltd.), and delivered through HD 820 headphones (Sennheiser Co.) at 70 dB SPL.

Experimental sounds consisted of speech digits from zero to nine vocalized by eight male talkers (TI46 LDC Corpus [46]). Digits were delivered in triplet sequences with random inter-digit intervals (Fig. 1A, T_1 - T_3), randomly chosen digits and talkers in the presence of a competing natural background noises (Fig. 1A). The silent intervals (T_1 - T_3) were not included in the estimation of the speech power when computing SNR. All sounds lasted 3.5 seconds duration and were gated with a 100 ms b-spline ramp to avoid broadband transients. Eleven environmental noises were selected to represent sounds widely encountered in nature (two- and eight-speaker babble, bird-babble, factory, city, construction, jackhammer, freeway, water, fire as well as white noise as a control) and which encompass a diverse range of spectrum and modulation statistics. For each of the experiments below, all the original and perturbed sound variants were delivered over 25 trials in block randomized order. Following each sound trial, participants were required to type the digit sequence heard on a keyboard.

Experiment 1: This experiment was carried out in n=18 participants. The experimental sounds are designed to assess how the spectrum and modulation statistics of natural environmental backgrounds interfere with speech recognition. In addition to delivering digit triplets in the original backgrounds (OR), which contains the original spectrum and modulation statistics, we delivered digits in phase randomized (PR) background variants which preserve the sound magnitude spectrum while randomizing the phase spectrum, thus distorting (whitening) the modulation statistics. These sounds are equivalent to spectrum shaped white noise. We also delivered digits in spectrum equalized (SE) background variants[28]. Here the background spectrum is whitened but the modulation statistics are preserved. All sounds were delivered at an SNR of -9 dB.

Experiment 2: This experiment explores how individual summary statistics of natural background sounds interfere with and influence speech recognition. Here, we used jackhammer (n=6 participants) and eight-speaker babble (n=10 participants) backgrounds since these sounds exhibited opposing recognition trends driven by the modulation and spectrum statistics of the background (Fig. 1C, D). Texture synthesis[29] was used to generate synthetic background variants that sequentially incorporate the 1) spectrum (Spec), 2) marginal (+Mar), 3) modulation power (+Mod), and 4) correlation (+Corr) summary statistics of each background. Sequentially adding summary statistics to the synthetic variants creates backgrounds that are perceptually more realistic and more closely match the statistics of the original background[29]. SNRs were varied between -12 to -3 dB (3 dB steps).

Experiment 3: In this final experiment, we tested whether summary statistics from the OR backgrounds influence speech recognition independently of the noise level (SNR). The original backgrounds used for Experiment 1, where delivered in the OR configuration multiple SNRs. For one group of participants (n=5 participants) sounds were delivered at -15, 12, -9, -6, -3 dB SNR. For n=4 participants, the range was extended to also include -18 and -3 dB.

In experiments 2 and 3, where the listening SNR was varied, we evaluated whether SNR influenced perceptual accuracy independently of the sound statistics. Here, singular value decomposition was first used to determine whether the response accuracy matrices are separable functions of the included background sound statistics and SNR (for experiment 2; background sound vs. SNR, for experiment 3). The separability index was defined as the fractional power explained by the first singular value, $SI = \lambda_1^2/\sum \lambda_i^2$. SI values near 1 indicate that the response accuracy matrix is separable, so that SNR influences accuracy independently of the background statistics. Secondly, in addition to considering spectro-temporal features of the background sound as variables that can predict perceptual accuracy (for the model-based approach below), we also included SNR as an independent explanatory variable into a logistic regression model. This approach provides a secondary model-based assessment of independence between SNR and spectro-temporal acoustic features.

Peripheral and Mid-Level Acoustic Features

We used a two-stage hierarchical auditory model representation of the auditory system to derive time-average summary statistics for both the foreground and background sounds and these time-averaged statistical features are subsequently used to predict human speech recognition in noise behavior. The auditory model has been described previously[18] and, thus, only necessary details are outlined below.

Both the foreground and background sounds for each trial in Experiments 1 and 2 were first decomposed through peripheral filterbank that models the frequency decomposition, envelope extraction, and nonlinear rectification of the cochlea. At this stage, sounds are represented by a cochlear spectrogram, $S(t, X_l)$, where X_l represents the octave frequency of the lth channel (57 filters between 0.1-5.3 kHz) and t is the time variable. The time-averaged cochlear power spectrum is then estimated as:

$$\text{cPS}(X_l) = \frac{1}{N}\sum_t S(t, X_l)^2$$

where N is the number of time-samples. The cPS summary statistics accounts for the spectral power distribution as represented through a cochlear model. It is subsequently used to assess the contribution of spectrum-based interference between the foreground speech and background.

Next, to represent the sounds in a mid-level auditory feature space, the foreground and background cochlear spectrograms are passed through a set of modulation selective filters that model the decomposition of sound into spectrotemporal modulation component in auditory midbrain. Here, each sound is represented as a multi-dimensional midbrainogram, $S(t, X_l, f_m, \Omega_n)$, where f_m is the temporal modulation frequency for the mth channel (filter modulation freq. of 0, ±2 to ±256 Hz in octave steps) and Ω_n is the spectral modulation frequency of the nth channel (0 and 0.1 to 3.2 cycles/oct in octave steps). The midbrain model derived modulation power spectrum is then computed as

$$\mathrm{mMPS}(X_l, f_m, \Omega_n) = \frac{1}{N}\sum_t S(t, X_l, f_m, \Omega_n)^2.$$

The mMPS summary statistics represents the sound power distribution jointly across frequency channels and combinations of spectral and temporal modulation channels, as represented through an auditory midbrain decomposition. This multidimensional sound representation is used below to model the interference between background and foreground speech sounds across spectral and modulation components in order to predict speech recognition in noise behavior.

Using Acoustic Interference Logistic Regression to Predict Speech-in-Noise Behavior

We tested whether the foreground and background summary statistics outlined above can serve as explanatory acoustic features to predict speech recognition capabilities in the presence of arbitrary real-world environmental background sounds. Conceptually, we expect that the foreground and background sound power in the cochlear or mid-level auditory space can account for acoustic interference underlying speech recognition in noise perceptual abilities. To test this, we used either the cPS or mMPS derived features as predictor variables and used logistic regression to predict digit recognition abilities in natural sources of environmental noise (Experiment 1 and 3).

The cPS and mMPS statistics described above, were first vectorized and concatenated for the background and foreground sounds yielding a composite feature vector, $\mathbf{X}$, for each of the acoustic representations. Both the cPS and mMPS representations were tested separately, which allows us to test how each of the feature spaces contributes to perceptual interference. For instance, cPS representation contains only spectral cues and thus should be able to account for spectral interference between the foreground and background. By comparison, the mMPS representation also decomposes the sound into modulation components, and thus should be able to account for higher-level modulation interference. The composite feature vector for the cPS representation was obtained by concatenating cPS feature vectors from the foreground and background:

$$\mathbf{X} = [\mathbf{cPS}_F \;\; \mathbf{cPS}_B]$$

where the individual features in $\mathbf{CPS}_F$ and $\mathbf{CPS}_B$ are whitened so that they have unit variance and zero mean across experimental trials. Alternately, we also considered a mid-level feature representation. Here the composite feature vector was defined as

$$\mathbf{X} = [\mathbf{mMPS}_F \;\; \mathbf{mMPS}_B]$$

where $\mathbf{mMPS}_F$ and m$\mathbf{MPS}_B$represent the vectorized foreground and background mid-level features. Given the high dimensionality of the cPS (57x2=114) and mMPS (57x7x17x2=13,566) spaces, principal component analysis is used to define dimensionality reduced feature vectors

$$\mathbf{Z}_L = \mathbf{X}\,\mathbf{W}_L$$

where $\mathbf{Z}_L$ are the principal component scores and $\mathbf{W}_L$ is an L-rank matrix containing the singular values of the covariance matrix, $\mathbf{X}^T\mathbf{X}$. This dimensionality reduced representation is then used to fit a low-dimensional logistic regression model, which minimizes the possibility of overfitting. For experiment three, where we also varied the noise level, SNR was appended to $\mathbf{Z}_L$ as an independent

explanatory variable to account for accuracy differences driven by the listening noise level. Here the probability of a correct behavioral decision is modeled by the logistic function

$$p(\mathbf{Z}_L) = \frac{1}{1+e^{-(\beta_0+\boldsymbol{\beta}\cdot\mathbf{Z}_L)}},$$

where β_0 is regression intercept and the vector $\boldsymbol{\beta} = [\beta_1 \cdots \beta_L]$ contains the logistic regression weight parameters. For a given feature representation (cPS or mMPS), the regression weights were derived using cross validated maximum likelihood estimation, where we optimized for the number of explanatory variables by adjusting L between 1 to 100. For each iteration of L, the optimal regression weights were derived using the dimensionality reduced feature vector, $\mathbf{Z}_{\boldsymbol{L}}$, as input to the model. The training and validation data samples were split using 25-fold cross validation and the cross-validated log-likelihood was estimated for each iteration of L. The optimal model dimensionality (L_{opt}) was then chosen by selecting the number of principal components that maximized the cross-validated log-likelihood (**Supplementary Figure 2**). Although the original stimulus dimensionality was substantially higher for the mid-level representation, the optimal cochlear and mid-level models had a similar dimensionality (cochlear L_{opt}=20; mid-level L_{opt}=24).

References

1 Cherry, E. C. Some Experiments on the Recognition of Speech, with One and with Two Ears. *J Acoust Soc Am* **5**, 975–979 (1953). https://doi.org:10.1121/1.1907229

2 Kidd, G., Jr. *et al.* Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss. *J Acoust Soc Am* **145**, 440 (2019). https://doi.org:10.1121/1.5087555

3 Shinn-Cunningham, B. G. & Best, V. Selective attention in normal and impaired hearing. *Trends Amplif* **12**, 283-299 (2008). https://doi.org:10.1177/1084713808325306

4 Miller, G. A. The masking of speech. *Psychol Bull* **44**, 105-129 (1947). https://doi.org:10.1037/h0055960

5 Bacon, S. P. & Grantham, D. W. Modulation masking: effects of modulation frequency, depth, and phase. *J Acoust Soc Am* **85**, 2575-2580 (1989). https://doi.org:10.1121/1.397751

6 Ewert, S. D. & Dau, T. Characterizing frequency selectivity for envelope fluctuations. *J Acoust Soc Am* **108**, 1181-1196 (2000).

7 Fogerty, D., Xu, J. & Gibbs, B. E., 2nd. Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum. *J Acoust Soc Am* **140**, 1800 (2016). https://doi.org:10.1121/1.4962494

8 Dau, T., Kollmeier, B. & Kohlrausch, A. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J Acoust Soc Am* **102**, 2892-2905 (1997).

9 Ihlefeld, A., Shinn-Cunningham, B. G. & Carlyon, R. P. Comodulation masking release in speech identification with real and simulated cochlear-implant hearing. *J Acoust Soc Am* **131**, 1315-1324 (2012). https://doi.org:10.1121/1.3676701

10 Moore, B. C. Co-modulation masking release: spectro-temporal pattern analysis in hearing. *Br J Audiol* **24**, 131-137 (1990). https://doi.org:10.3109/03005369009077854

11 Hall, J. W., Haggard, M. P. & Fernandes, M. A. Detection in noise by spectro-temporal pattern analysis. *J Acoust Soc Am* **76**, 50-56 (1984).

12 Chi, T., Gao, Y., Guyton, M. C., Ru, P. & Shamma, S. Spectro-temporal modulation transfer functions and speech intelligibility. *J Acoust Soc Am* **106**, 2719-2732 (1999).

13 Elliott, T. M. & Theunissen, F. E. The modulation transfer function for speech intelligibility. *PLoS Comput Biol* **5**, e1000302 (2009). https://doi.org:10.1371/journal.pcbi.1000302

14 Mesgarani, N., David, S. V., Fritz, J. B. & Shamma, S. A. Phoneme representation and classification in primary auditory cortex. *J Acoust Soc Am* **123**, 899-909 (2008). https://doi.org:10.1121/1.2816572

15 Durlach, N. I. *et al.* Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *J Acoust Soc Am* **114**, 368-379 (2003).

16 Durlach, N. I. *et al.* Note on informational masking. *J Acoust Soc Am* **113**, 2984-2987 (2003).

17 Singh, N. C. & Theunissen, F. E. Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* **114**, 3394-3411 (2003).

18 He, F., Stevenson, I. H. & Escabi, M. A. Two stages of bandwidth scaling drives efficient neural coding of natural sounds. *PLoS computational biology* **19**, e1010862 (2023). https://doi.org:10.1371/journal.pcbi.1010862

19 von Békésy, G. & Peake, W. T. (Acoustical Society of America, 1990).

20 Depireux, D. A., Simon, J. Z., Klein, D. J. & Shamma, S. A. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J Neurophysiol* **85**, 1220-1234 (2001).
21 Miller, L. M., Escabi, M. A., Read, H. L. & Schreiner, C. E. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J Neurophysiol* **87**, 516-527 (2002).
22 Escabi, M. A. & Schreiner, C. E. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *J Neurosci* **22**, 4114-4131 (2002). https://doi.org:20026325
22/10/4114 [pii]
23 Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006-1010 (2014). https://doi.org:10.1126/science.1245994
24 Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Curr Biol* **28**, 3976-3983 e3975 (2018). https://doi.org:10.1016/j.cub.2018.10.042
25 Kell, A. J. E. & McDermott, J. H. Invariance to background noise as a signature of non-primary auditory cortex. *Nat Commun* **10**, 3958 (2019). https://doi.org:10.1038/s41467-019-11710-y
26 Schneider, D. M. & Woolley, S. M. Sparse and background-invariant coding of vocalizations in auditory scenes. *Neuron* **79**, 141-152 (2013). https://doi.org:10.1016/j.neuron.2013.04.038
27 Rabinowitz, N. C., Willmore, B. D., King, A. J. & Schnupp, J. W. Constructing noise-invariant representations of sound in the auditory pathway. *PLoS biology* **11**, e1001710 (2013). https://doi.org:10.1371/journal.pbio.1001710
28 Zhai, X. *et al.* Distinct neural ensemble response statistics are associated with recognition and discrimination of natural sound textures. *Proc Natl Acad Sci U S A* **117**, 31482-31493 (2020). https://doi.org:10.1073/pnas.2005644117
29 McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nat Neurosci* **16**, 493-498 (2013). https://doi.org:nn.3347 [pii]
10.1038/nn.3347
30 Khatami, F., Wohr, M., Read, H. L. & Escabi, M. A. Origins of scale invariance in vocalization sequences and speech. *PLoS Comput Biol* **14**, e1005996 (2018). https://doi.org:10.1371/journal.pcbi.1005996
31 Ding, N. *et al.* Temporal modulations in speech and music. *Neurosci Biobehav Rev* (2017). https://doi.org:10.1016/j.neubiorev.2017.02.011
32 McDermott, J. H. & Simoncelli, E. P. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron* **71**, 926-940 (2011). https://doi.org:S0896-6273(11)00562-9 [pii]
10.1016/j.neuron.2011.06.032
33 Holdgraf, C. R. *et al.* Encoding and Decoding Models in Cognitive Electrophysiology. *Front Syst Neurosci* **11**, 61 (2017). https://doi.org:10.3389/fnsys.2017.00061
34 Theunissen, F. E. *et al.* Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network* **12**, 289-316 (2001).
35 Dau, T., Ewert, S. & Oxenham, A. J. Auditory stream formation affects comodulation masking release retroactively. *J Acoust Soc Am* **125**, 2182-2188 (2009). https://doi.org:10.1121/1.3082121

36 Gauvin, H. S., McMahon, K. L. & de Zubicaray, G. I. Top-down resolution of lexico-semantic competition in speech production and the role of the left inferior frontal gyrus: an fMRI study. *Language, Cognition and Neuroscience* **36**, 1-12 (2021). https://doi.org:10.1080/23273798.2020.1762904

37 Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V. & Polley, D. B. Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing. *Elife* **9** (2020). https://doi.org:10.7554/eLife.51419

38 Zekveld, A. A., Heslenfeld, D. J., Festen, J. M. & Schoonhoven, R. Top-down and bottom-up processes in speech comprehension. *Neuroimage* **32**, 1826-1836 (2006). https://doi.org:10.1016/j.neuroimage.2006.04.199

39 Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98**, 630-644 e616 (2018). https://doi.org:10.1016/j.neuron.2018.03.044

40 Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 8619-8624 (2014). https://doi.org:10.1073/pnas.1403112111

41 Battleday, R. M., Peterson, J. C. & Griffiths, T. L. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat Commun* **11**, 5418 (2020). https://doi.org:10.1038/s41467-020-18946-z

42 Feather, J., Leclerc, G., Madry, A. & McDermott, J. H. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat Neurosci* **26**, 2017-2034 (2023). https://doi.org:10.1038/s41593-023-01442-0

43 Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* **13**, 600-612 (2004).

44 Grant, K. W. *et al.* An objective measure for selecting microphone modes in OMNI/DIR hearing aid circuits. *Ear Hear* **29**, 199-213 (2008).

45 Andreas, S., Ted, P. & Venkatraman, A. in *Audio Signal Processing and Coding* 383-403 (Wiley, 2007).

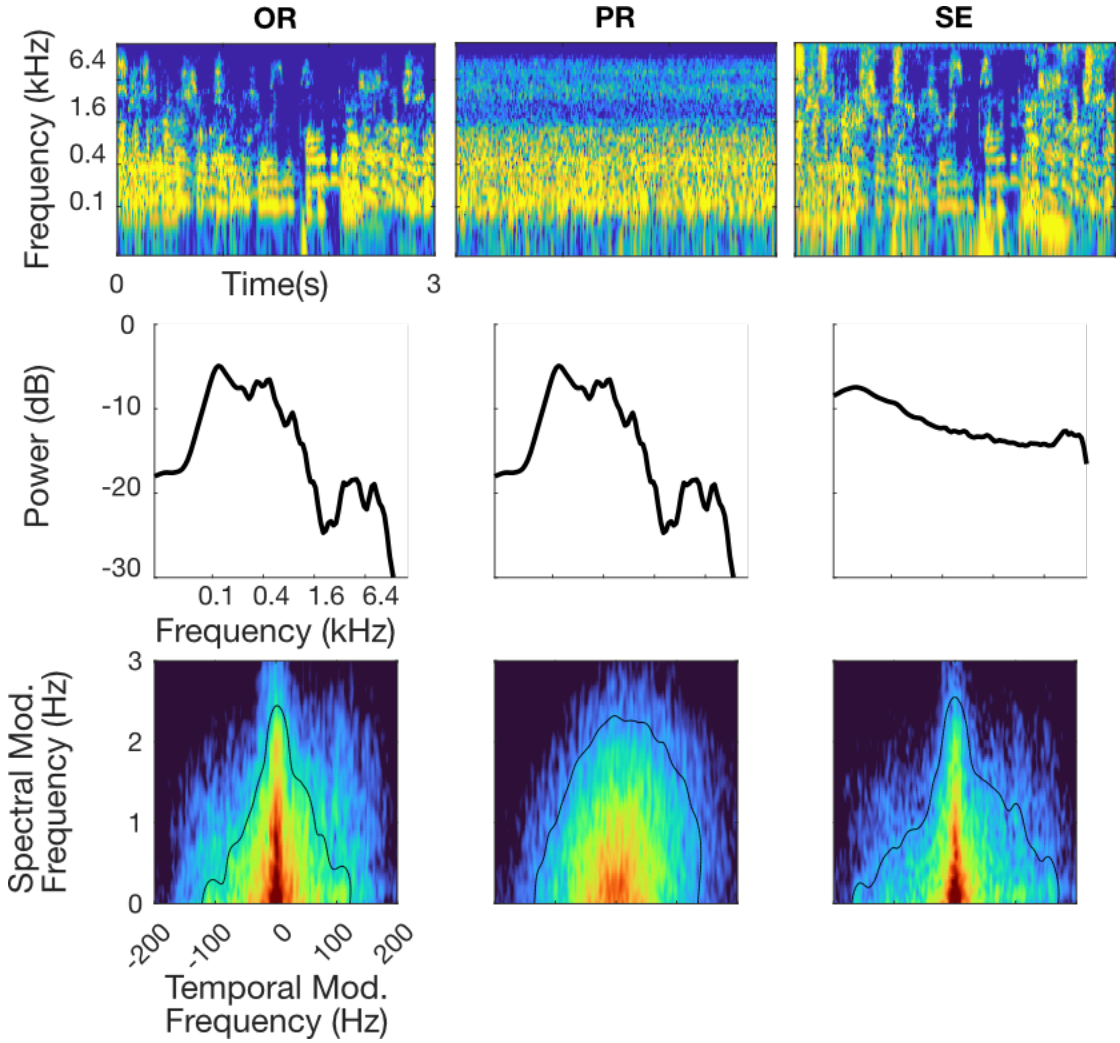
46 Liberman, M. e. a. (ed Linguistics Data Symposium) (NIST Speech Disc 7-1.1 (1 disc) 1991).

Supplementary Material: Low-dimensional interference of mid-level sound statistics predicts human speech recognition in natural environmental noise

Alex Clonan[1,2,4], Xiu Zhai[5], Ian H. Stevenson[2-4], Monty A. Escabí[1-4],

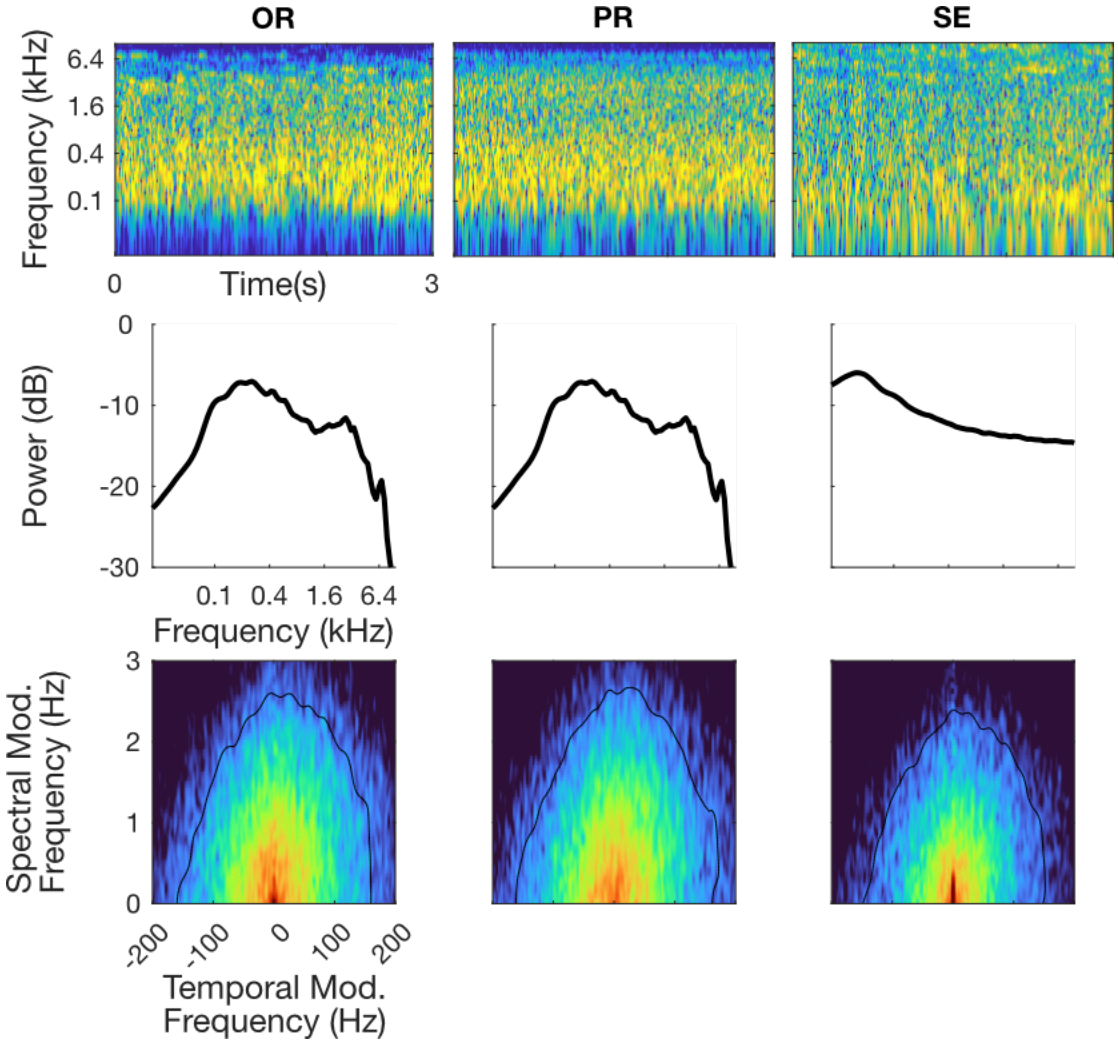
[1]Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269
[2]Biomedical Engineering, University of Connecticut, Storrs, CT 06269
[3]Psychological Sciences, University of Connecticut, Storrs, CT 06269
[4]Institute of Brain and Cognitive Sciences, University of Connecticut, Storrs, CT 06269
[5]Biomedical Engineering, Wentworth Institute of Technology, Boston, MA 02115

Supplementary Fig 1A. Eight-Speaker Babble

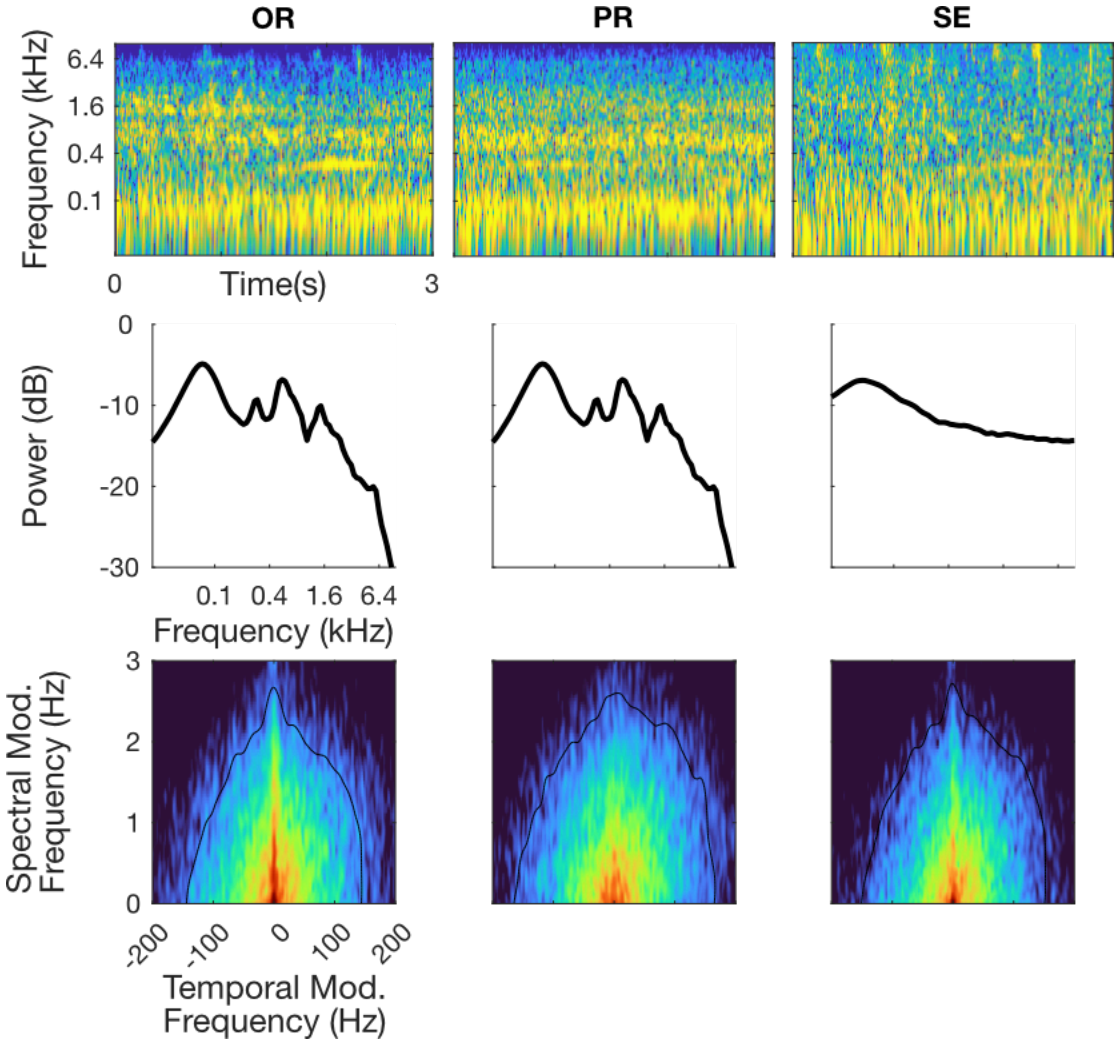


Supplementary Fig. 1: Cochlear and modulation statistics for 3-second exemplars of all of the backgrounds used in Experiment 1. Each panel (A-K) is formatted as for Fig. 1B. Each of the backgrounds is shown for the original (OR), phase randomized (PR), and spectrum equalize (SE) configurations (column 1-3, respectively). Cochleograms, cochlear power spectrum (cPS), and modulation power spectrum are shown in the 1st, 2nd and 3rd rows. Note that regardless of the background, the PR condition preserves the original (OR) spectrum of the sound (middle panel, 2nd row) but whitens the modulation power (middle panel, 3rd row). Alternately, the SE condition whitens the sound cPS, but retains much of the original (OR) sound modulation content.

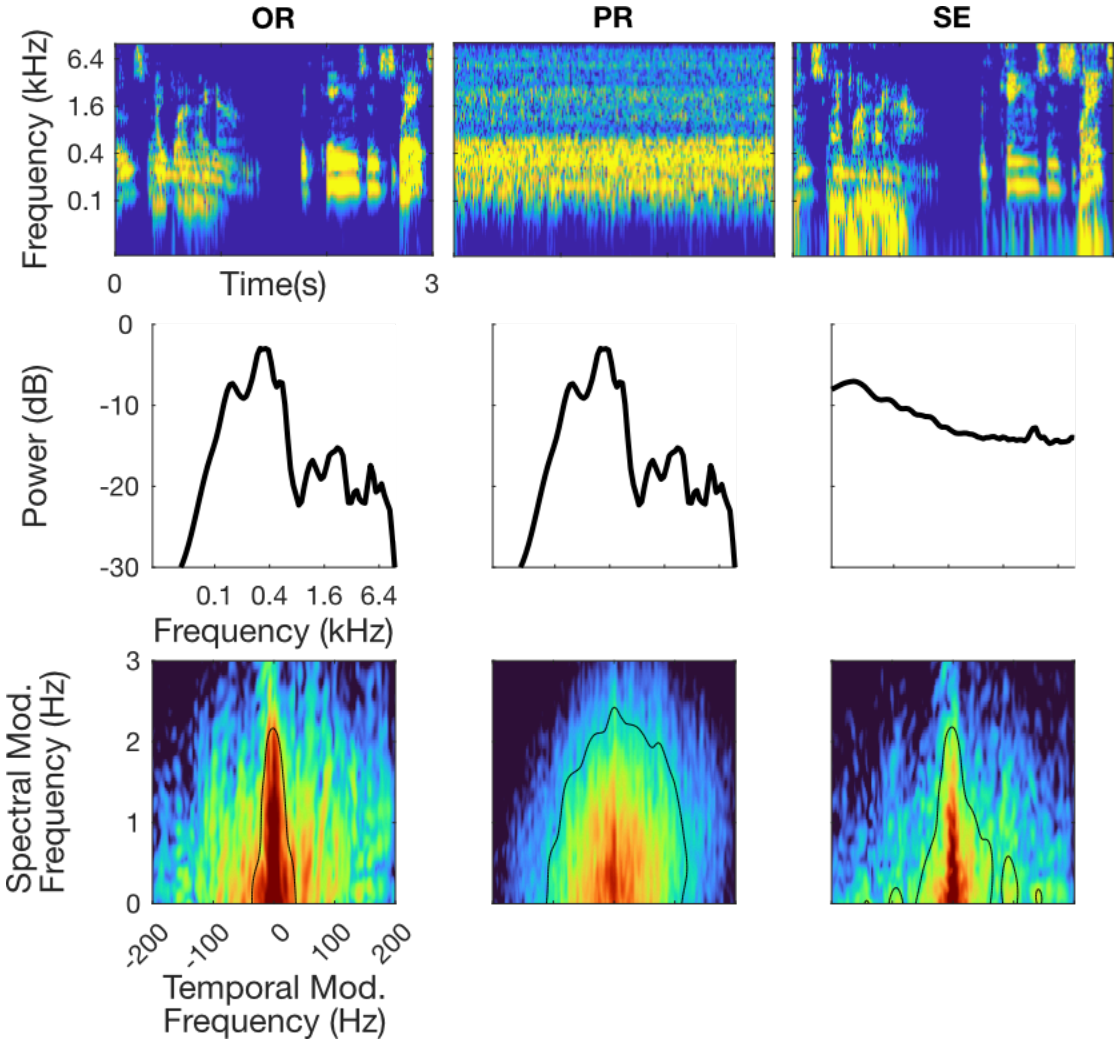
Supplementary Fig 1B. Factory



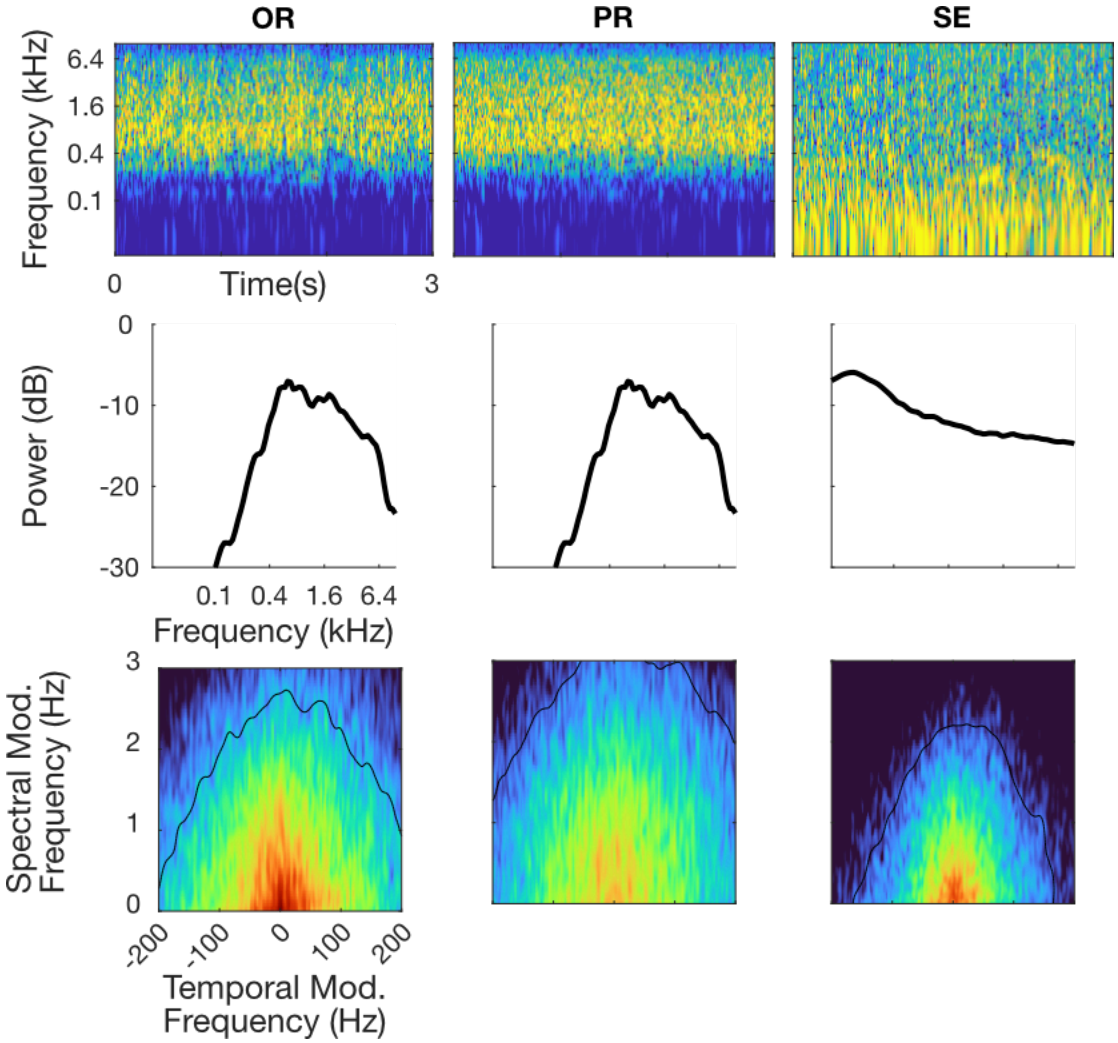
Supplementary Fig 1C. City



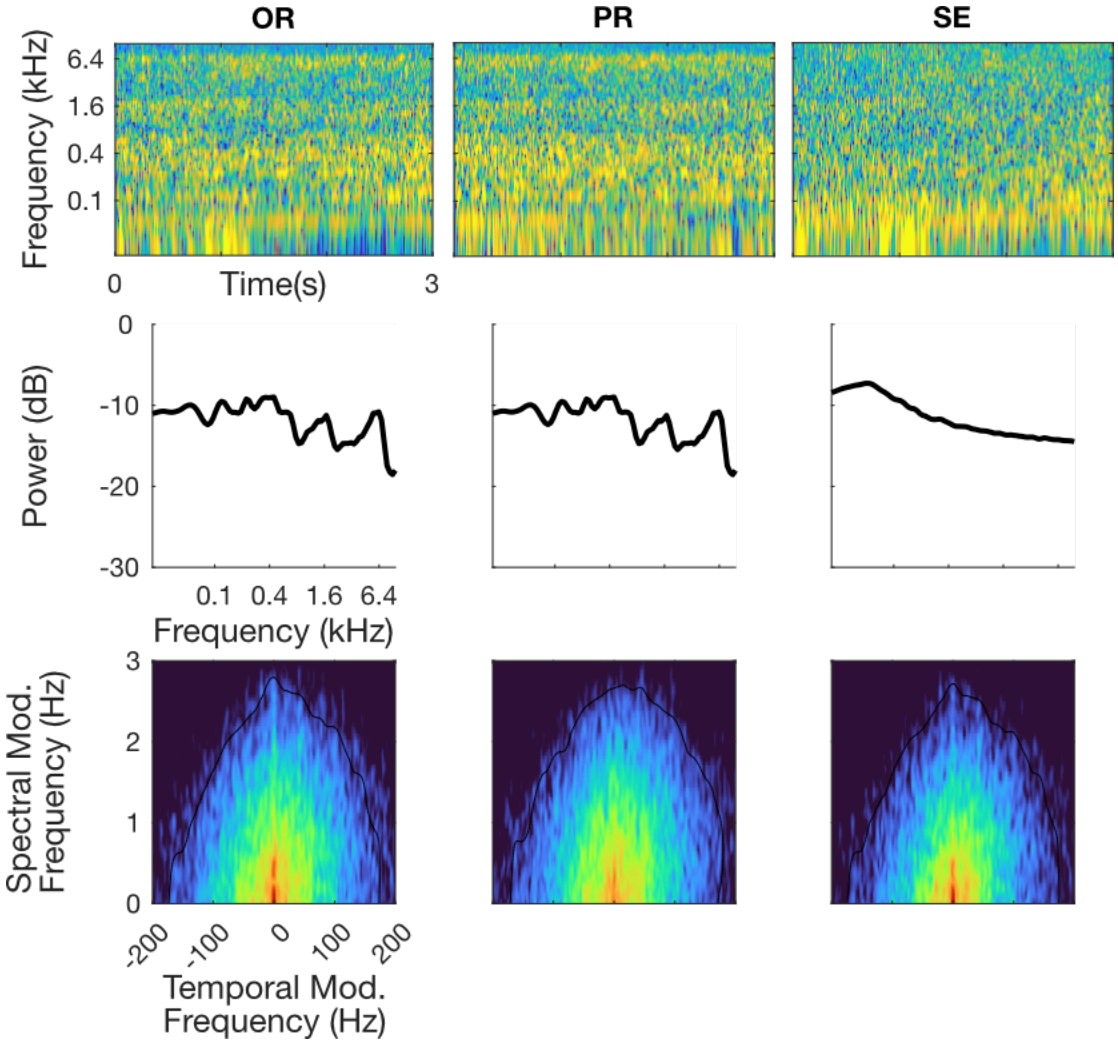
Supplementary Fig 1D. Two Speaker Babble



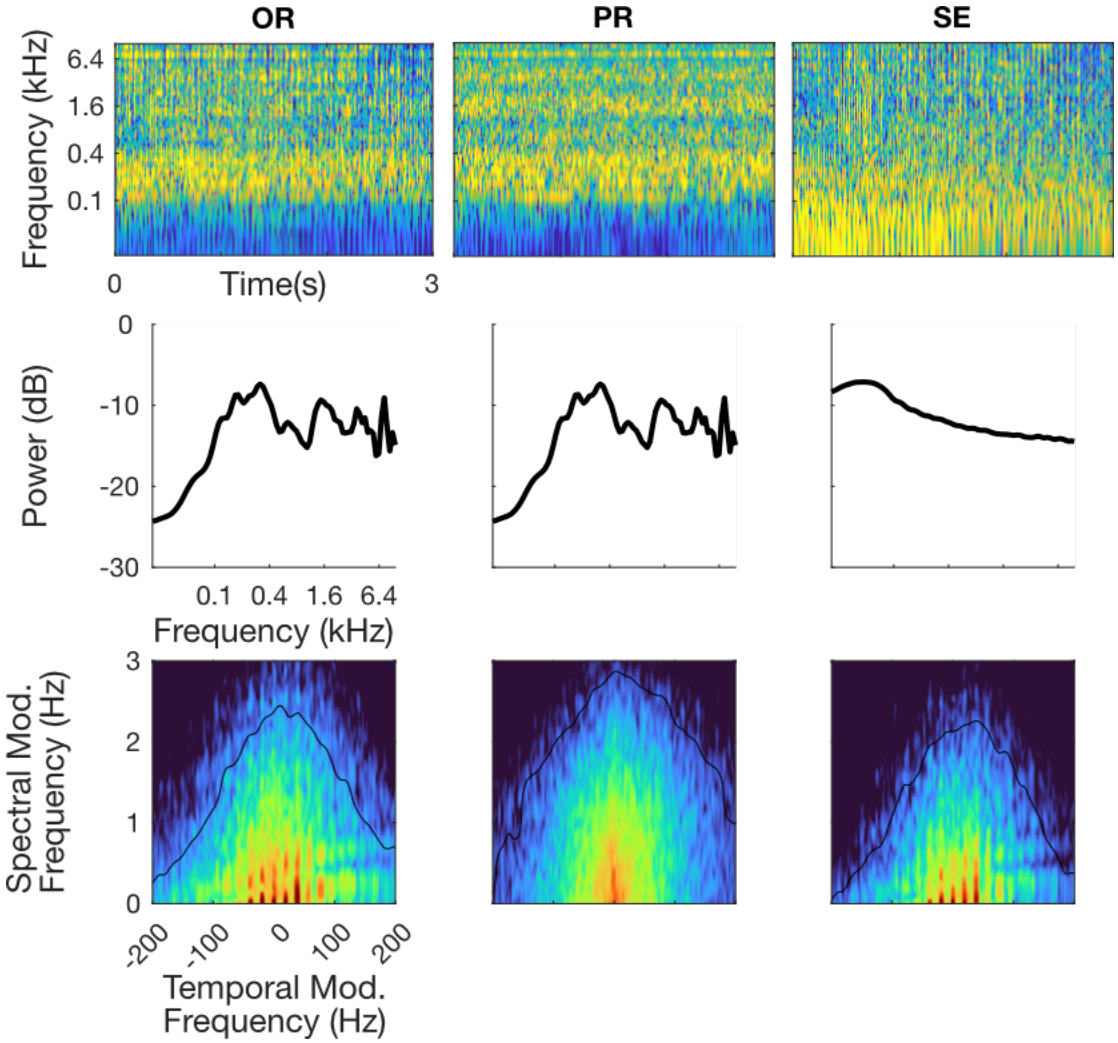
Supplementary Fig 1E. Water



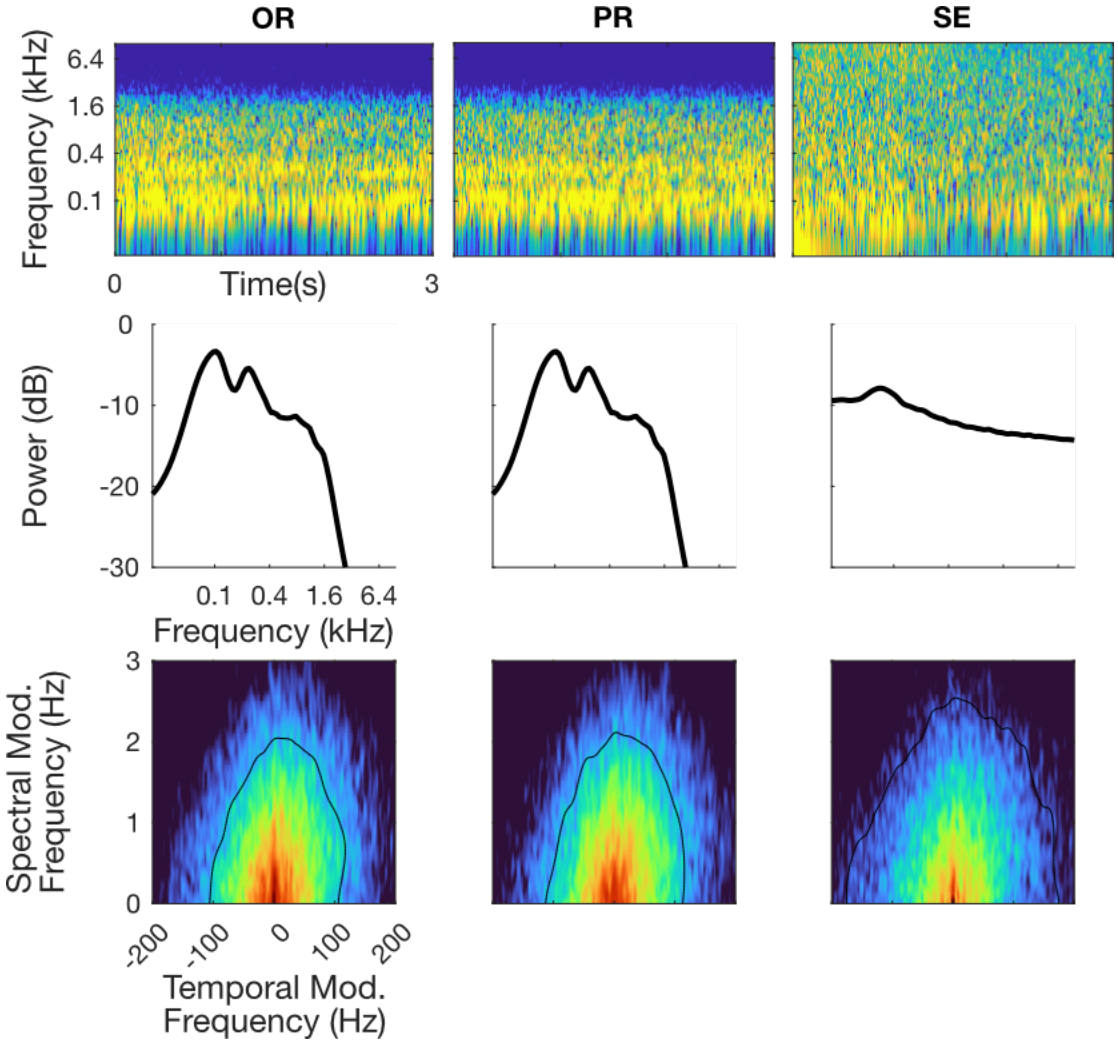
Supplementary Fig 1F. Construction



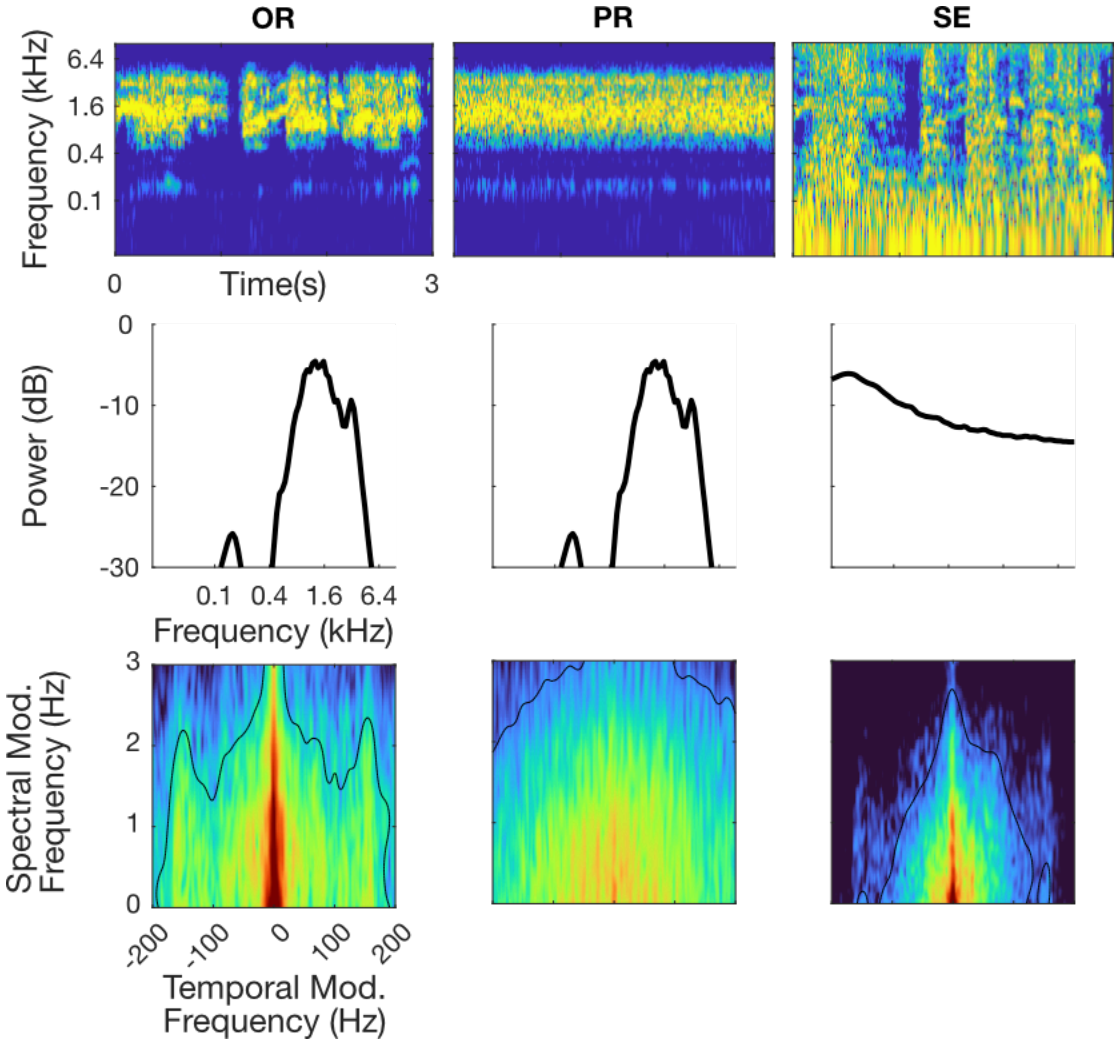
Supplementary Fig 1G. Jackhammer



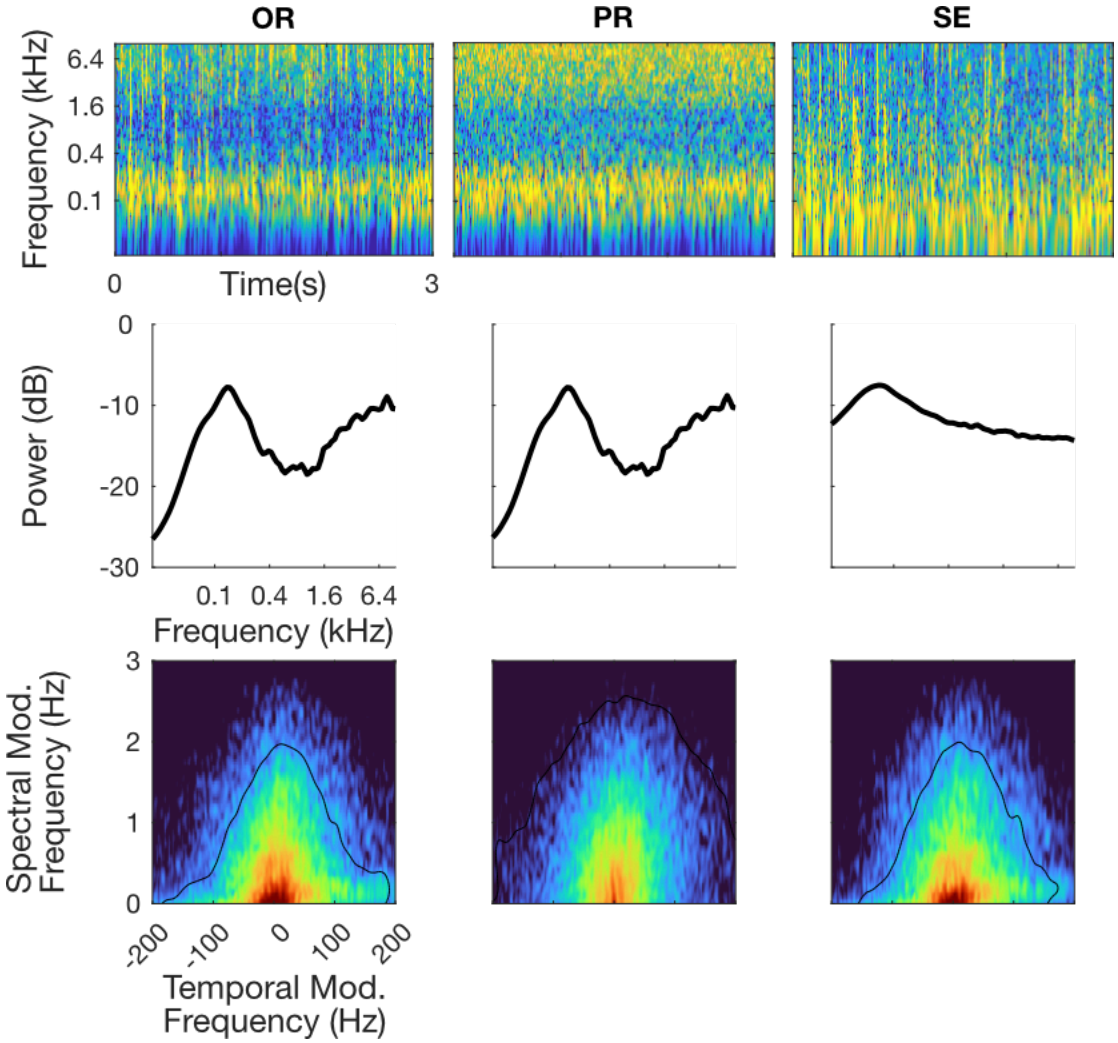
Supplementary Fig 1H. Freeway



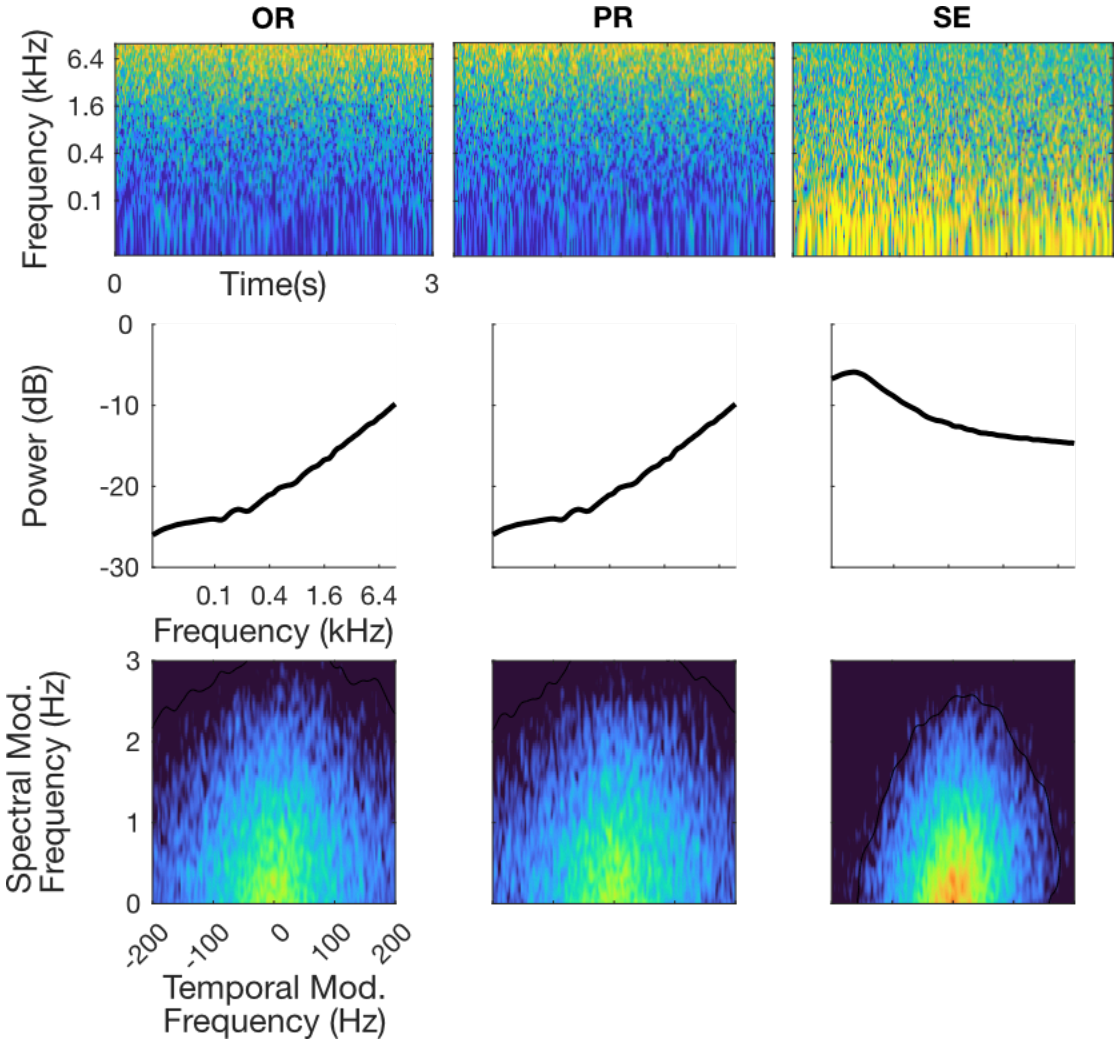
Supplementary Fig 1I. Bird Babble

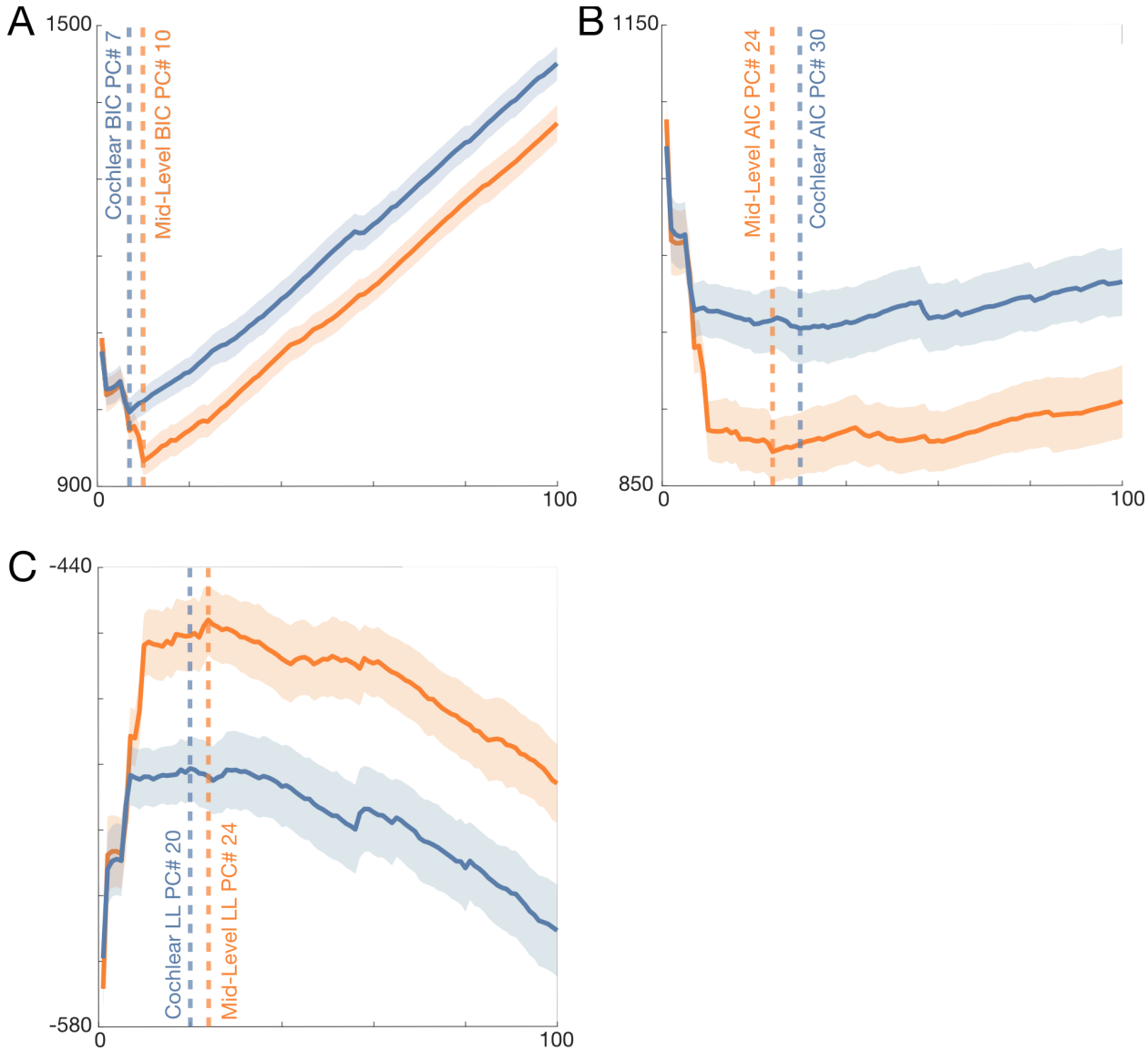


Supplementary Fig 1J. Fire

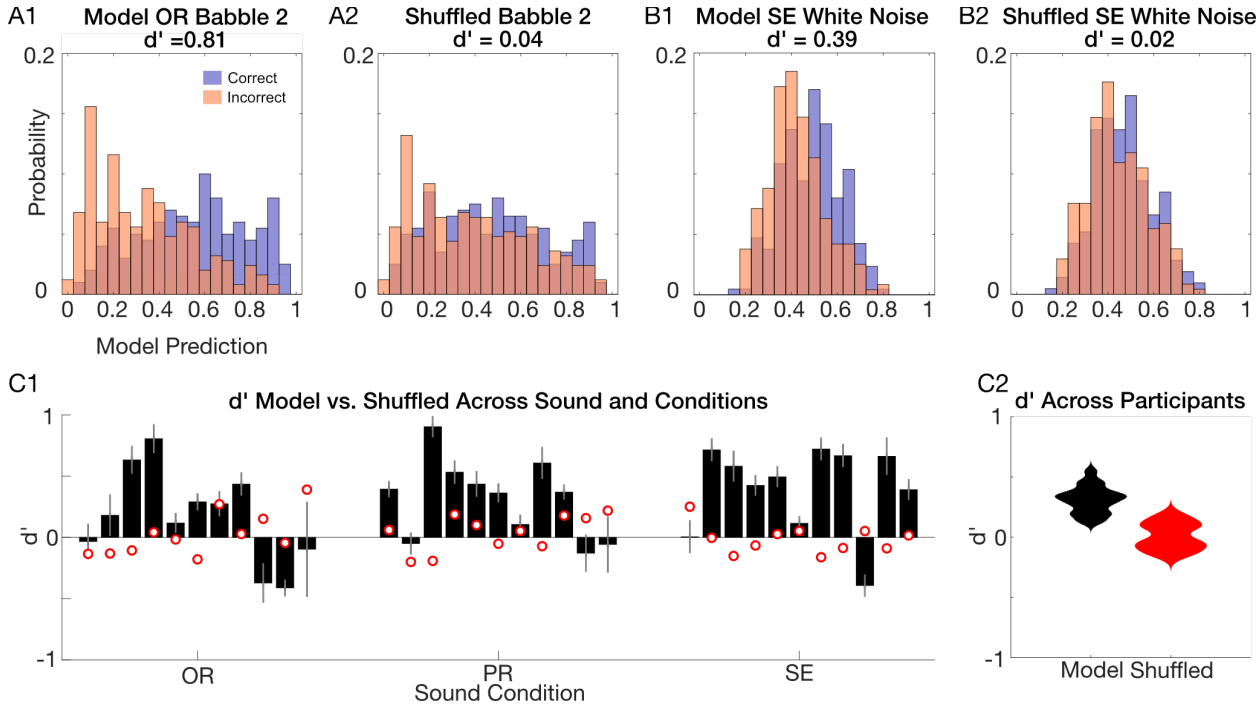


Supplementary Fig 1K. White Noise

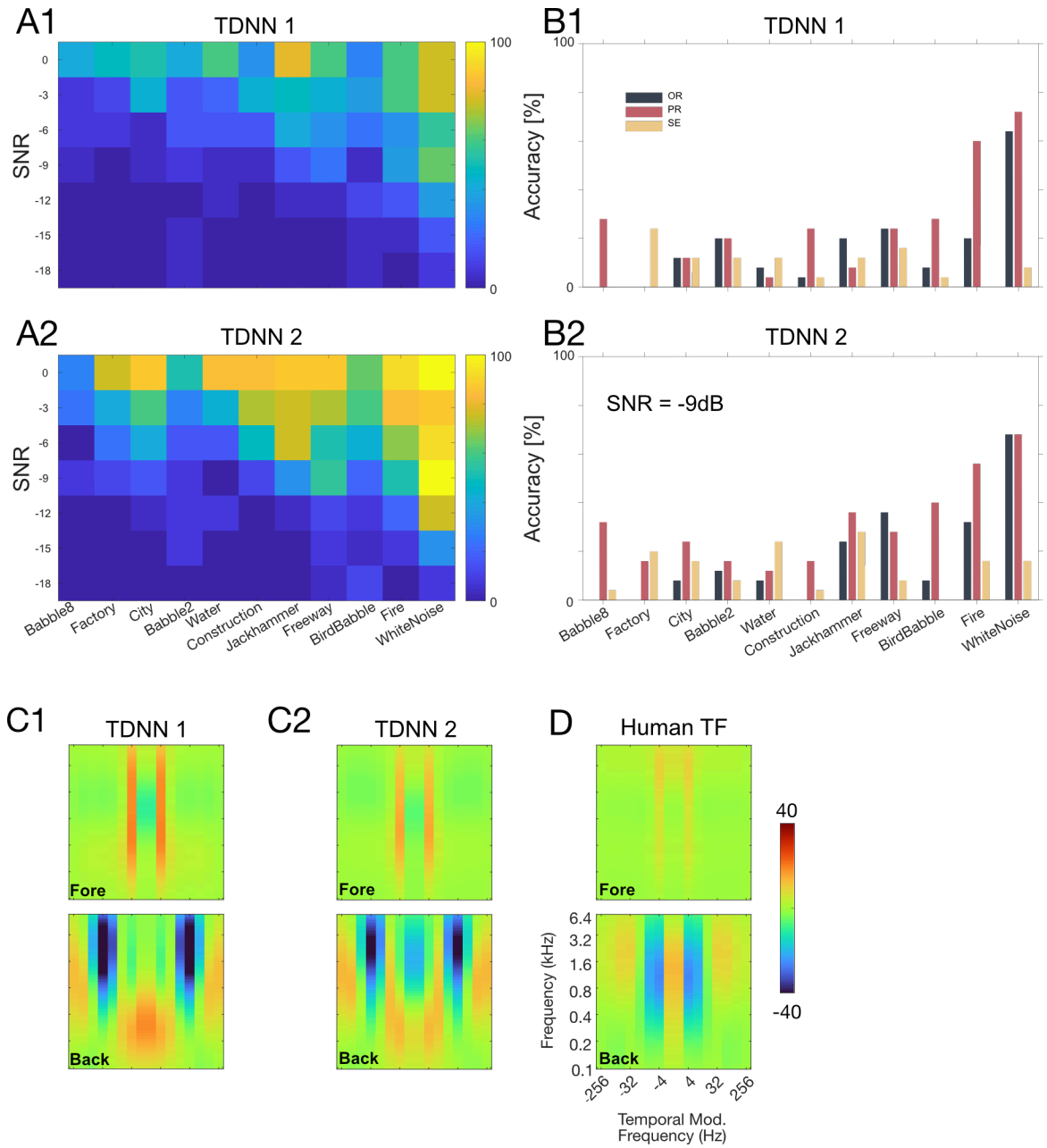




Supplementary Fig. 2: Deriving the optimal number of principal components (PCs) for both the cochlear and mid-level models. For both the cochlear (blue) and mid-level model (orange), we measured the (A) Bayesian information criterion (BIC), (B) Akaike information criterion (AIC), and (C) log-likelihood (LL) while varying the number of principal components used for each model. The curves indicate the average value of each metric and the shaded region indicates SD (across participants). Vertical dashed lines indicate the number of PCs needed to minimize the AIC or BIC or maximize the LL.



Supplementary Fig. 3: Mid-Level logistic-model accounts for single-trial changes in accuracy and outperforms a category-specific Bernoulli model for single trial predictions. (A) Distribution of predicted accuracy for the mid-level model and category-specific Bernoulli model simulation (trial shuffled predictions) for two different backgrounds (A1-A2, Bable-2; B1-B2, SE White Noise). For both backgrounds, model predictions for correct trials (blue) were biased towards higher predicted accuracy while incorrect trials (pink) showed lower predicted accuracy, which for both cases produced a positive sensitivity index (A1, babble-2, d'=0.81; B1, SE White Noise, d'=0.39). When the prediction trials are shuffled (A2, B2), any association between the spectrotemporal sound cues, participant responses, and predicted performance is lost as expected for a Bernoulli process where correct and incorrect trials would occur independently of the acoustic cues for each background category. After trial shuffling, the model produces overlapping prediction distributions for correct and incorrect decisions and a low sensitivity index near zero (A2, babble-2, d'=0.04; B1, SE White Noise, d'=0.02). (C1) Histogram showing the sensitivity index for all the background sounds and different background conditions (OR, PR and SE). The background sounds are ordered in ascending OR accuracy (as in Fig. 1C). Bar plots show the average sensitivity index (*d*') for each background and condition for the Mid-Level logistic-model whereas the red dots show the sensitivity index (*d*') for the category-specific Bernoulli model (simulated using trial shuffling). Error bars designate the SEM (across participants). (C) Distribution of average sensitivity index (n=18 participants; averaged across all backgrounds and conditions, OR, SE, PR) for the mid-level model (black) and the category-specific Bernoulli model (trial shuffled simulation, red). * indicates a significant difference (t-test, $t(17)=7.9$, $p<10^{-6}$).



Supplementary Fig. 4: Accuracy for two neural network models diverges from human results. Results are shown for the multiple SNR stimuli (as in Fig. 5; left, A1 and B1) and the OR/PR/SE stimuli at -9dB (as in Fig. 1; right A2 and B2). Here we tested task performance using two pretrained time delay neural network models (TDNN) with the pytorch-based SpeechBrain library [1]. One network was trained for word recognition using speech commands (Google Speech Commands dataset[2]) and the second was trained for speaker recognition (Voxceleb1+ Voxceleb2 datasets [3-5]). Both networks generate low-dimensional embeddings (512-d) of variable length sound input using X-Vectors[6], and here we trained an additional readout layer (L2-regularized logistic regression) to identify the single digits 0-9 using our digit-in-noise stimuli. To make the model output comparable to the human data, results here show cross-validated three-digit accuracy calculated by combining model outputs across single digits. Note that unlike the human participants, the models have information about the exact start and stop times of the foreground digits. As

for the human data, the outputs of these models were fitted to our mid-level logistic regression model (using 30 PCs) and the model weights were used for model evaluation. (D) For reference, the average background and foreground logistic model human weights are shown for all of our participants. (C1-C2) The foreground and background weights of both networks are quite different from those of human listeners indicating that they employ different invariances and use different spectrotemporal cues that impact their performance. For instance, there is a strong background suppressive component ~32 Hz in both networks not present for humans as well as positive hot-spots not visible for human listeners. Furthermore, the background weights of both networks are stronger than humans (shown on identical color scales) which is consistent with their lower digit recognition performance.

REFERENCES

1 Ravanelli, M. *et al.* SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624* (2021).

2 Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

3 Nagrani, A., Chung, J. S., Xie, W. & Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* **60**, 101027 (2020).

4 Chung, J. S., Nagrani, A. & Zisserman, A. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).

5 Nagrani, A., Chung, J. S. & Zisserman, A. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

6 Snyder, D. *et al.* in *The Speaker and Language Recognition Workshop.*