Supplementary Information for: "Frequency following responses to tone glides: effects of age and hearing loss," M.R. Molis, W.J. Bologna, B.M. Madsen, R. Muralimanohar, and C.J. Billings, *Journal of the Association for Research in Otolaryngology*

Table of Contents

Defining the problem and our approach

Given the dataset containing one average waveform per subject per condition per montage, our goal was to be able to state, for each of these waveforms, whether it contains a detectable FFR signal or not. The purpose of this was to check whether any trends in the data across stimulus conditions would change qualitatively if one were to exclude from the dataset those waveforms without detectable FFR.

Posed in this way, the signal detection problem essentially boils down to a question of binary classification: "is this waveform in the *Present* or *Absent* category?" (with the assumption that all waveforms must fall into one or the other).

The FFR was quantified in two ways: (1) response strength, measured as the signal-to-noise-ratio (SNR) of the response, and (2) temporal coherence between the stimulus and response, measured as the stimulus-to-response correlation coefficient (SRCC).

The steps involved in our method can be summarized as follows:

(1) Model each FFR measure (SNR, SRCC) as a random variable that follows a certain family of distributions.

(2) Estimate distribution parameters separately for each montage (horizontal, vertical) and response category (*Present*, *Absent*) using a maximum likelihood estimation (MLE) fitting algorithm.

(3) Model the relationships between the variables by linking them with a copula, separately for each montage.

(4) Use the model to simulate a large data sample.

(5) Use the simulated sample observations to train a binary classifier using a logistic regression algorithm. The classifier assigns each simulated observation a score that rates the relative likelihood that such an observation came from the *Present* rather than the *Absent* distribution.

(6) Select the minimum required score for detection by referencing the ROC curve and finding the point with the minimum linear distance from (0,1).

(7) Apply the classifier to the empirical (non-simulated) data, and note the effect (or lack therefore) of excluding observations that fall below the threshold score.

Each of these steps is explained in greater detail in the sections below.

Establishing a "gold standard"

First, however, there is the question of how to define the *Present* and *Absent* categories. That is, to specify a "gold standard" on which to base our models and estimates in the steps above.

For each of the two categories, we identified multiple options to use as the gold standard:

Absent

(a) Flatline (zero) response

i. Advantages: Simple; reflects ideal, intuitive concept of a non-response.

ii. Disadvantages: Not reflective of what non-FFR EEG recordings actually look like; will overestimate separability between conditions; does not address how to estimate dispersion.

(b) Artifact-only blocks (stimulus on, but inaudible because earphone isolated from ear)

i. Advantages: Actual non-FFR EEG recording; accounts for possible presence of electromagnetic transducer artifact.

ii. Disadvantages: Only acquired for 1/3-octave rising stimulus condition; separated in time from FFR recordings; was always recorded last, so may overestimate noise levels.

(c) Prestimulus interval

i. Advantages: Actual non-FFR EEG recording; interleaved in time with the FFR recordings; corresponding data exists for every FFR stimulus condition; consistent with noise floor calculation used in manuscript.

ii. Disadvantages: Would not account for any transducer artifact, if present.

Present

(a) Scaled version of stimulus waveform

i. Advantages: Simple; reflects ideal, intuitive conception of what a response should look like.

ii. Disadvantages: Cleaner than even the best-case empirical observations, which will increase likelihood of excluding recordings with valid FFRs; SRCC will be at ceiling of 1.0, and this paradigm does not address how to estimate dispersion from that.

(b) All observed data from YNH group only

i. Advantages: Actual data from recordings intended to elicit FFRs; FFR obviously visible in majority of cases; risk of circularity mitigated by modeling based on only a subset of the full dataset.

ii. Disadvantages: Still some risk of circularity/overfitting since data used to fit models overlaps with data which models will be used to sort; to the extent that true FFRs look different in non-YNH populations, will increase exclusions of recordings with valid FFRs; partially predetermines the pattern of group differences by ensuring YNH has the least exclusions; part of what is detected may not in fact be FFR.

(c) All observed data from all groups

i. Advantages: Models will be maximally reflective of sampled population; corresponds to one intuitive concept of "detectability" (detection = more likely to observe in stimulus-on condition than stimulus-off condition).

ii. Disadvantages: Greatest risk of circularity; some data points used to fit *Present* model may not contain any response, making the algorithm more likely to err on the side of not excluding enough data.

(d) None (do not attempt to model FFR-present condition; model FFR detectability as complement of probability of non-response)

i. Advantages: Acknowledges that we have no true gold standard; does not make any assumptions about the distribution of the FFR signal.

ii. Disadvantages: Does not allow for estimation of detection sensitivity, precision, negative predictive value, or accuracy, only specificity and false positive rate; assumes anything deviating from baseline is signal of interest.

In both cases, we went with option (c). In the *Absent* case, using the prestimulus interval was an easy decision. The advantages overwhelmingly outweigh the disadvantages, and the disadvantages of this method are the least significant of any of the three options. This is especially true given that the data from our artifact-only runs suggest that there was no significant transducer artifact to account for (which is not surprising, given the shielding on the transducer and the distance from the EEG sensors).

One complication of this approach, however, is that we could not use the SNR measure defined in the main paper, because the prestimulus baseline is the denominator of that measure. Therefore, in place of the SNR measure, we simply used the average spectral amplitude in a 50-Hz bin centered on the stimulus frequency (the same as the SNR denominator in the main paper).

The *Present* case is more complicated. Ultimately, we chose to use the observed data in the response interval because we considered it important for the methods of quantifying the *Present* and *Absent* conditions to "match" as closely as possible. By using the same measure, with the same number of observations, for both parts of the model—with the only difference being *when* and *under what conditions* the measures were taken—we believed we could minimize the bias in the separation process.

This led to the same complication as with the *Absent* condition, i.e., taking this approach does not allow us to employ the SNR measure from the main study, because we are already using the prestimulus baseline response as the *Absent* gold standard. Accordingly, we used the spectral amplitude from the response interval rather than the SNR measure.

We used average spectral amplitude in a 50-Hz bin centered on the stimulus frequency (as we did with the prestimulus baseline) instead of using the peak spectral amplitude in that range (as we did in the main study). We made that choice because we didn't want to artificially increase the ease of separation

between the *Absent* and *Present* conditions by using separate measures for each of them. We considered it important to use measures for each that were as equivalent to one another as possible.

In retrospect, we think it may have been better (i.e., would have been more sensitive in detecting FFRs) if we had used peak values for both the *Present* and the *Absent* conditions. But we did not think to try that at the time.

Step 1: Modeling the FFR measures as random variables

We started from the hypothesis that we could effectively model both cross-correlation variables with beta distributions (because the measure is bounded between 0 and 1 and not uniformly distributed) and both spectral amplitude variables with lognormal distributions (because the measure is an absolute physical quantity). We then used visualizations in R such as PDFs (probability density functions), CDFs (cumulative distribution functions), P-P plots, and Q-Q plots (see Figures 1-4 below) to confirm that these distributions appeared to be appropriate fits to the trends in the observed data. We concluded that they were indeed appropriate. The code used to generate the figures in R is included as Appendix A below. The code also shows which packages must be installed and how to install them. (Note: 'RTools' is required to properly install some of the packages.)

Step 2: Estimate distribution parameters

For the model used in the final analysis, population distribution parameters were estimated separately for each montage, with the reasoning that each montage emphasizes activity at a different level of the auditory system and has a different mean value for each measure. For the Matlab code used to estimate the distribution parameters, see Appendix B (requires Statistics & Machine Learning Toolbox to run properly). The parameter values resulting from this analysis are given in Table 1.

Because the baseline and response distributions for the mean spectral amplitude measure had such a large degree of overlap with one another, in all subsequent steps we removed this measure from the detection model. All following steps use only SRCC. In retrospect, it is possible that using peak spectral amplitudes would have yielded better results, but we did not attempt this at the time.

Step 3: Model the relationships between variables

To capture the relationship between the two remaining variables (*Present* SRCC, *Absent* SRCC), we fit a Student's-*t* copula. We chose this particular copula type because it not only has a parameter quantifying the *strength* of the correlation but also (unlike many other copulas) a second parameter that varies the *shape* of the correlation, granting additional flexibility in the fitting process.

A requirement for fitting a copula is that the marginal distribution of each variable in the multivariate relationship to be modeled must be *uniform*. Therefore, prior to fitting the copula, the observed data for

each variable were subjected to a probability-integral transform, using the CDFs from the fits in Step 2. This had the effect of rescaling the data as if they were drawn from standard uniform distributions.

In parallel, we also fit two alternative copulas: one that used only the two spectral amplitude variables, and another that used only the two cross-correlation variables so that we could perform the subsequent analyses on each of the three models (SRCC-only, SNR-only, SRCC+SNR) in parallel, to determine if either of the measures on its own was able to perform similarly to the model that included both measures—or if, on the other hand, each measure contributes meaningful and unique information to the process of separating *Present* from *Absent*.

The copula fittings were performed in Matlab; for the code, see Appendix B. The copula parameters are given in Table 2.

Step 4: Simulate a large data sample from the model

We generated a large sample of simulated observations (100,000 baseline-and-response pairs) from the fitted copula. Then we used the marginal CDFs from Step 2 to perform inverse probability-integral transforms, bringing the simulated observations back to the natural scale for each measure. For the code used to generate the simulated observations in Matlab, see Appendix B. Note that, because this analysis involves random sampling from the copulas and we did not use a fixed seed for the random number generator, attempting to replicate this analysis will likely produce slightly different results, even with access to the original data set. That said, the expected random variation should not have any practical impact on the conclusions with an *n* this large.

Again, all this was repeated separately for each of the two montages.

Step 5: Train a binary logistic-regression classifier

We trained three separate binary logistic-regression classifiers using the simulated observations from each of the three copula models. Each classifier assigns each observation a score which estimates the relative likelihood that such an observation would have come from the *Present* rather than the *Absent* distribution. A score of 1 would indicate that an observation should definitely be sorted into the *Present* category, and a score of 0 that an observation should definitely be sorted into the *Absent* category. Any value in between would indicate that the observation could be sorted into either category, depending on the placement of the threshold.

For the Matlab code used to train the classifiers, see Appendix B. The parameters for each classifier are shown in Table 3.

Again, all this was repeated separately for each of the two montages.

Step 6: Use ROC curves to set the threshold

To select the threshold value that would serve as the criterion for FFR detection, we generated ROC curves for each of the three models to quantify and visualize the tradeoff between sensitivity (portion of FFRs detected) and specificity (1 – [portion of FFR-absent observations where an FFR was falsely "detected"]) at various thresholds. The threshold value was chosen to correspond with the point on the ROC curve with the shortest distance to the upper-left corner (0,1)—the corner that represents an ideal separation between conditions (i.e., all FFRs detected, with no false detections).

Again, this was performed separately for each of the two montages.

The ROC curves are shown in Figure 5. The optimal threshold values and their SRCC equivalents are given in the figure caption. The resulting estimates of sensitivity and specificity are shown in the figure itself.

Step 7: Apply the classifier to the actual observations

In applying the classifier to the data, we chose to use the one from the SRCC-only model. We believe this was the most parsimonious choice, given that the estimated sensitivity and specificity achieved by this model are essentially the same as in the model that includes both cross-correlation and spectral-amplitude measures—that is, it achieves equally good results while using fewer variables.

The non-detection counts by group and montage are detailed in Table II in the main manuscript. We generated new versions of all the data visualizations from the paper, this time with the observations which produced non-detections excluded from the dataset. The resulting plots are included here as Figures 6-9.

Figures

Figure 1. Pre-stimulus interval SNR: lognormal distribution

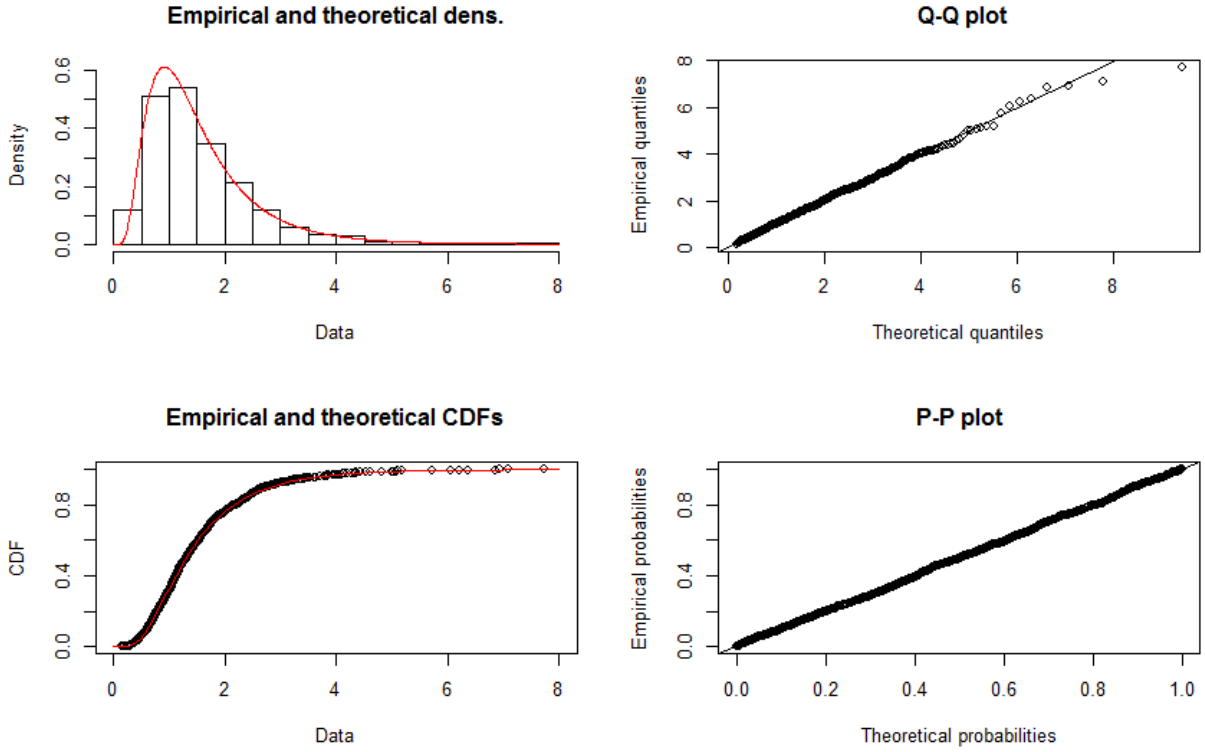

Figure 1. Goodness-of-fit evaluation using lognormal distribution to model pre-stimulus-interval average spectral amplitude in a 50-Hz bin centered on each response window's stimulus frequency. Plots show correspondence between theoretical lognormal population distribution (lines) and empirical distribution of observed data sample (circles and bars).

Figure 2. Response interval SNR: lognormal distribution

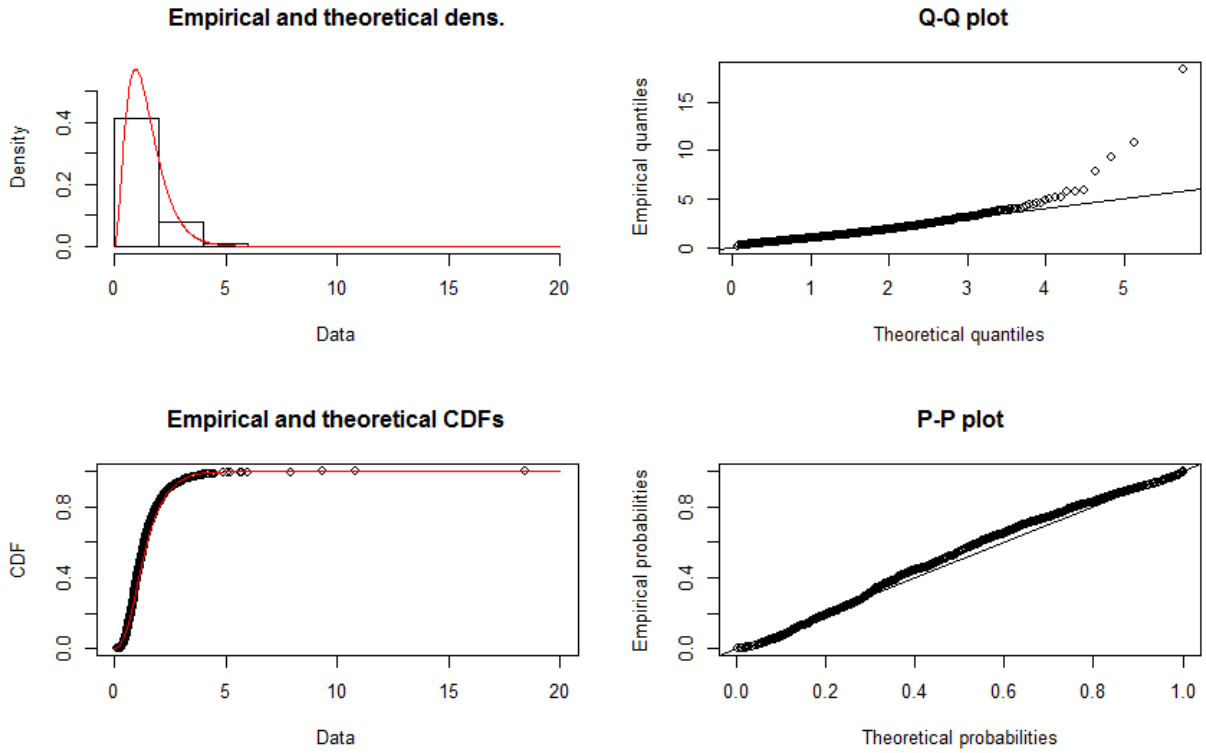

Figure 2. Goodness-of-fit evaluation using lognormal distribution to model response-interval average spectral amplitude in a 50-Hz bin centered on each response window's stimulus frequency. Plots show correspondence between theoretical lognormal population distribution (lines) and empirical distribution of observed data sample (circles and bars).

Figure 3. Pre-stimulus interval SRCC: beta distribution

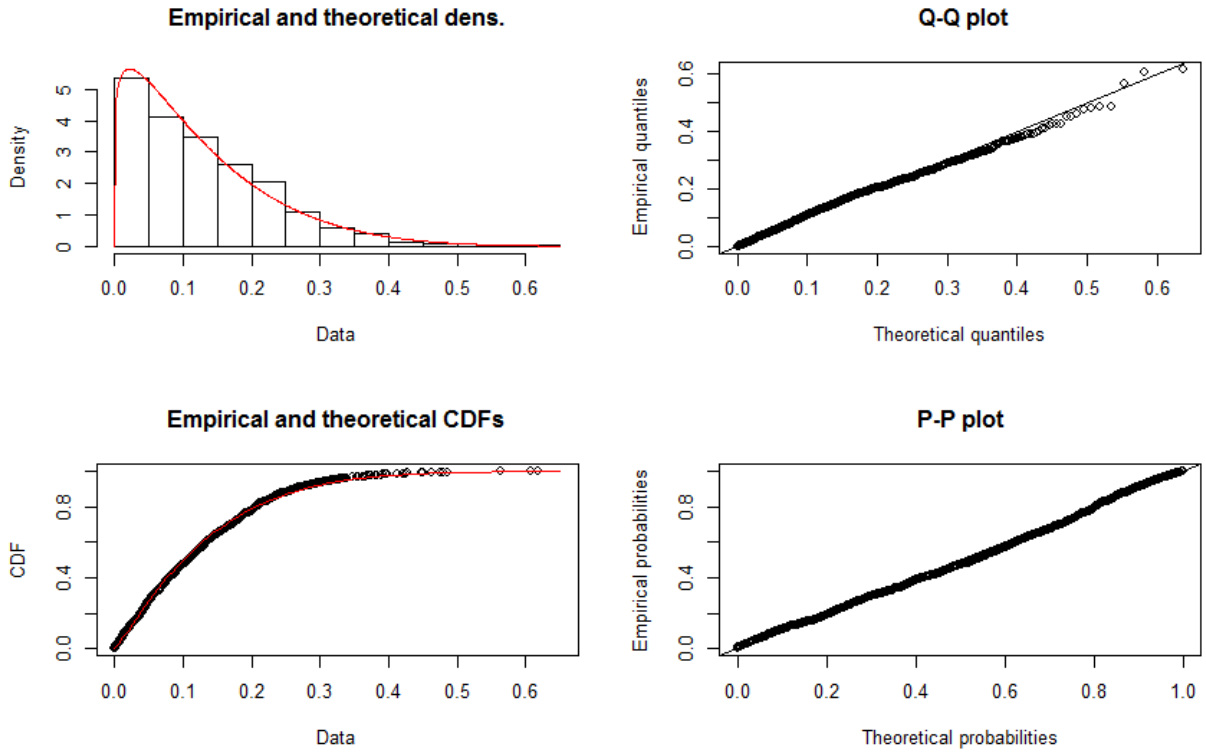

Figure 3. Goodness-of-fit evaluation using beta distribution to model the SRCC between the pre-stimulus baseline response and each of the three stimulus windows. Plots show correspondence between theoretical beta population distribution (lines) and empirical distribution of observed data sample (circles and bars).

Figure 4. Response interval SRCC: beta distribution

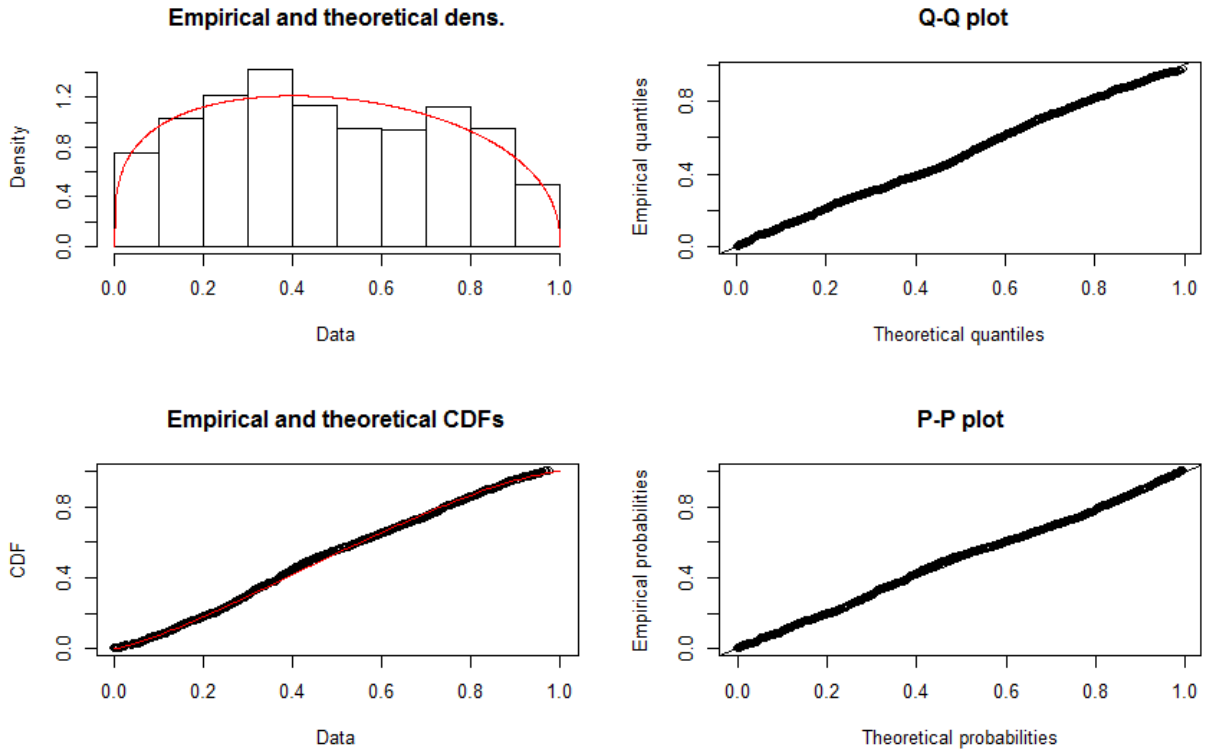

Figure 4. Goodness-of-fit evaluation using beta distribution to model the SRCC between each FFR window and the and the corresponding window in the stimulus. Plots show correspondence between theoretical beta population distribution (lines) and empirical distribution of observed data sample (circles and bars).

Figure 5. ROC curves: SRCC

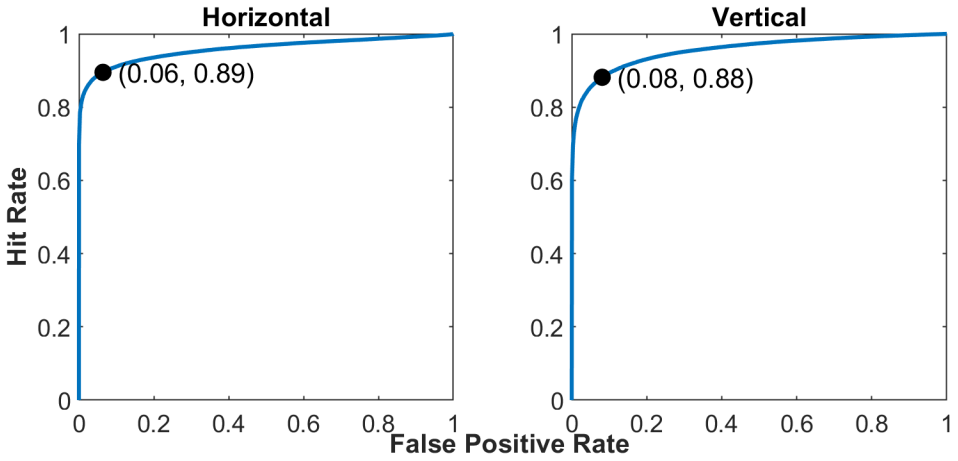

Figure 5. Receiver operating characteristic (ROC) curves, estimating the rate of true positives (hits) and false positives (false alarms) in terms of FFR detection, based on the simulated data. This analysis was performed separately for the horizontal montage (left) and the vertical montage (right). The dot on each curve represents the detection threshold criterion in terms of classifier score that optimizes sensitivity and specificity (i.e., the point closest to the upper left corner). For the horizontal montage, this corresponded to a classifier score of .4196, which translated to an SRCC of .2224. For the vertical montage, this corresponded to a classifier score of .4478, which translated to an SRCC of .2183.

Figure 6. SNR and SRCC by frequency, montage, and direction: all groups w/o excluded data

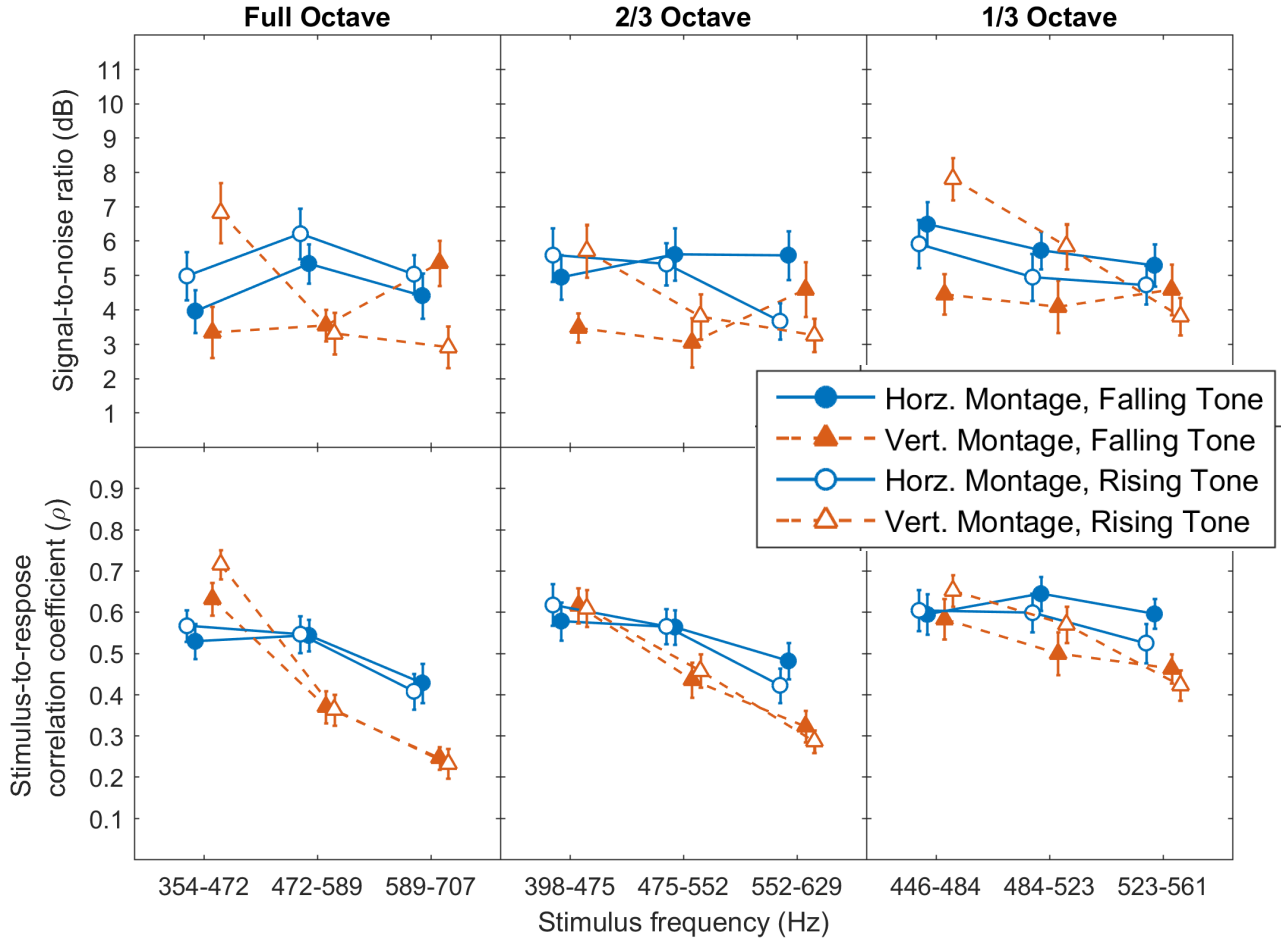

Figure 6. SNR and SRCC by frequency, montage, and direction. Includes participants from all groups, but with observations categorized as non-detections excluded. The pattern remains similar to Figure 3 in the main manuscript, which does not exclude any observations.

Figure 7. SNR and SRCC by montage, including participants: all groups, no data excluded

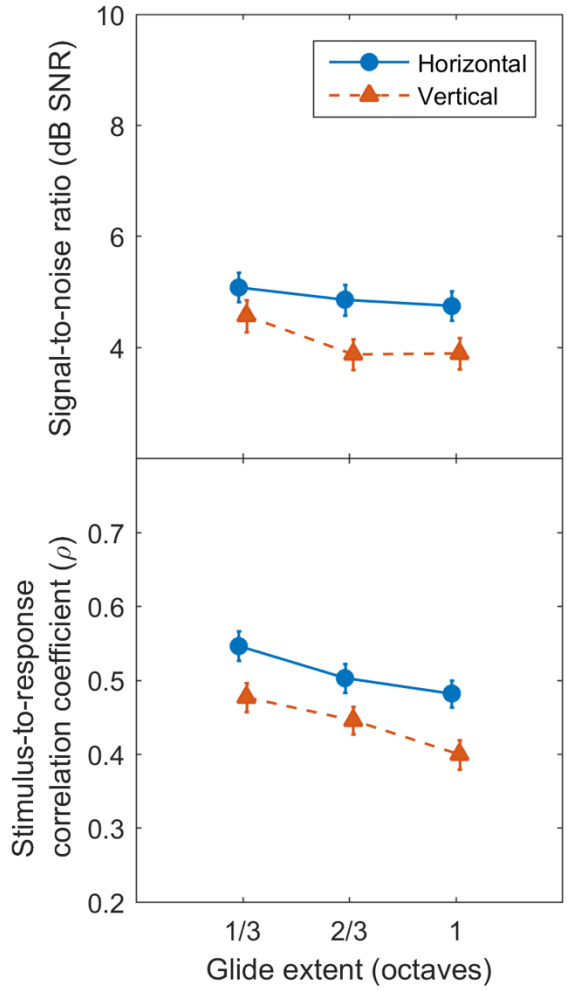

Figure 7. SNR and SRCC by montage, including participants from all groups, no data excluded. The pattern remains similar to the parallel figure in the main paper, which does not exclude any observations.

Figure 8. SNR and SRCC by montage, including participants: all groups w/o excluded data

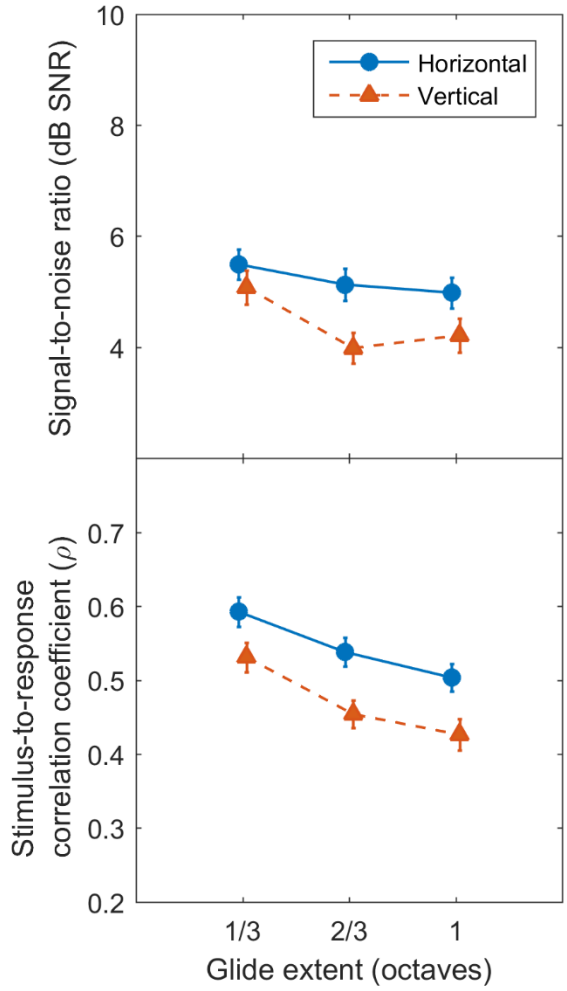

Figure 8. SNR and SRCC by montage, including participants from all groups, but with observations categorized as non-detections excluded. The pattern remains similar to the parallel figure in the main paper, which does not exclude any observations.

Figure 9. SNR and SRCC by montage and time-window: all groups, no data excluded

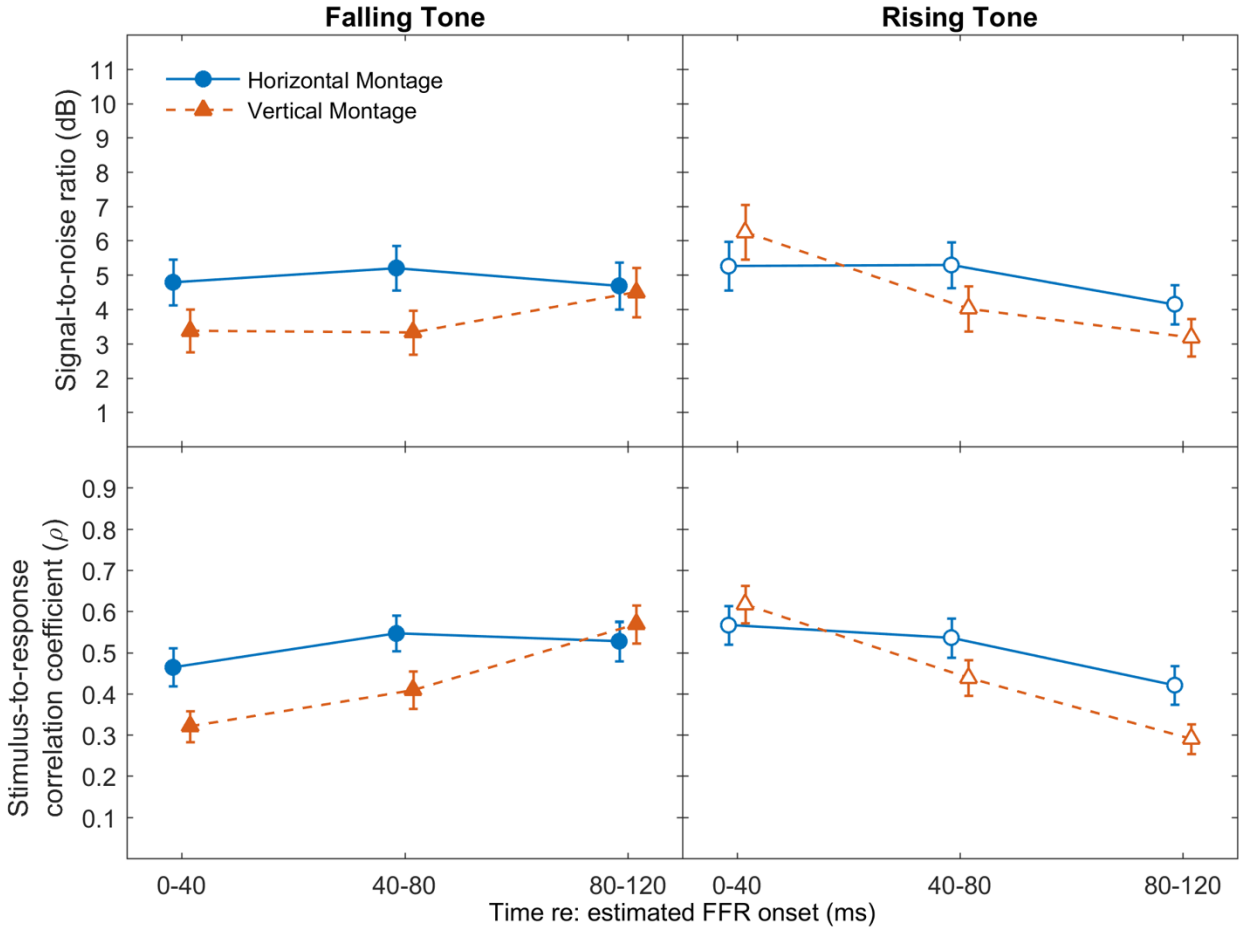

Figure 9. SNR and SRCC by montage and time-window, including participants from all groups, no data excluded. The pattern remains similar to the parallel figure in the paper, which does not exclude any observations.

Figure 10. SNR and SRCC by montage and time-window: all groups w/o excluded data

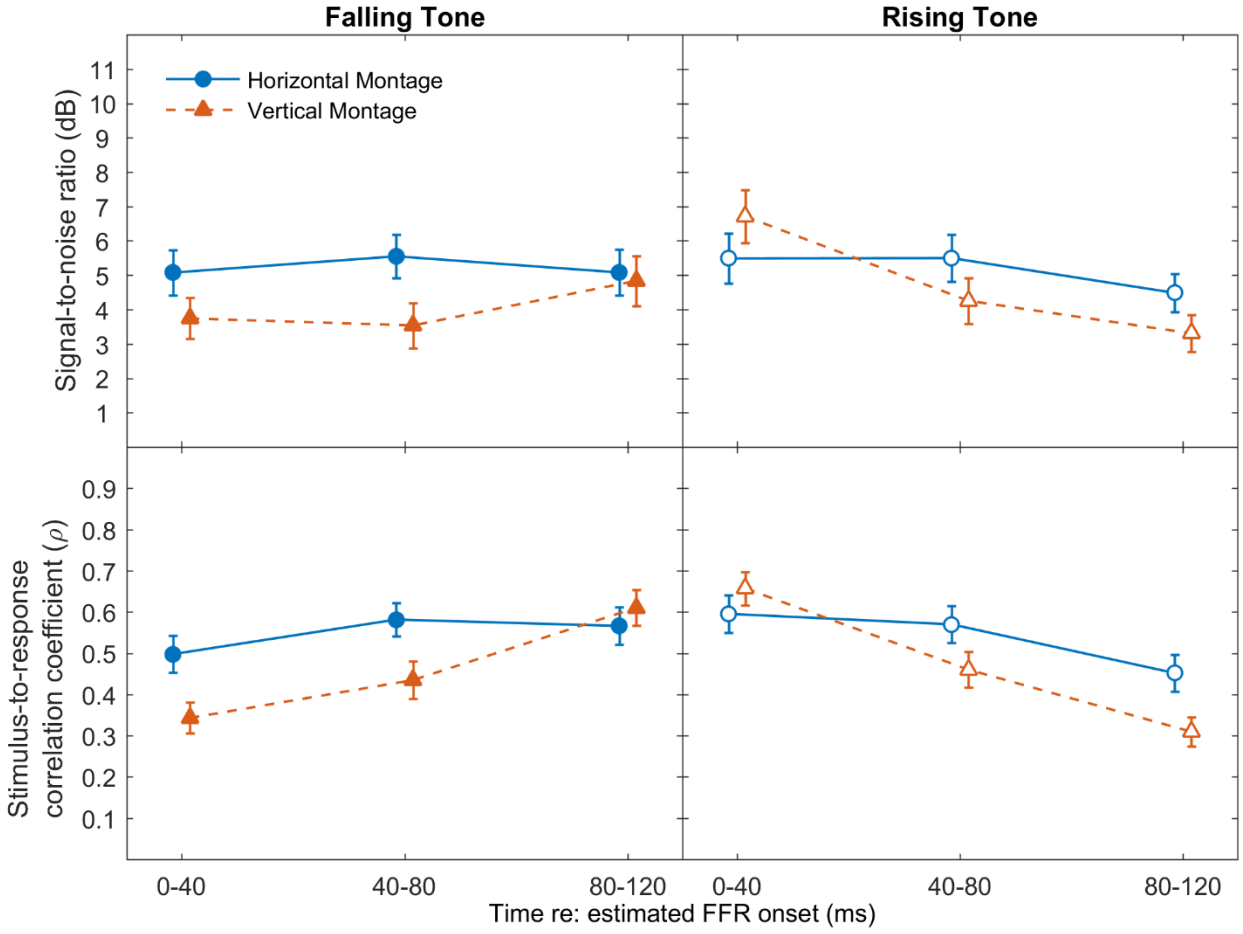

Figure 10. SNR and SRCC by montage and time-window, including participants from all groups, but with observations categorized as non-detections excluded. The pattern remains similar to the parallel figure in the paper, which does not exclude any observations.

Figure 11. Grand average waveforms by slope: all groups w/o excluded data

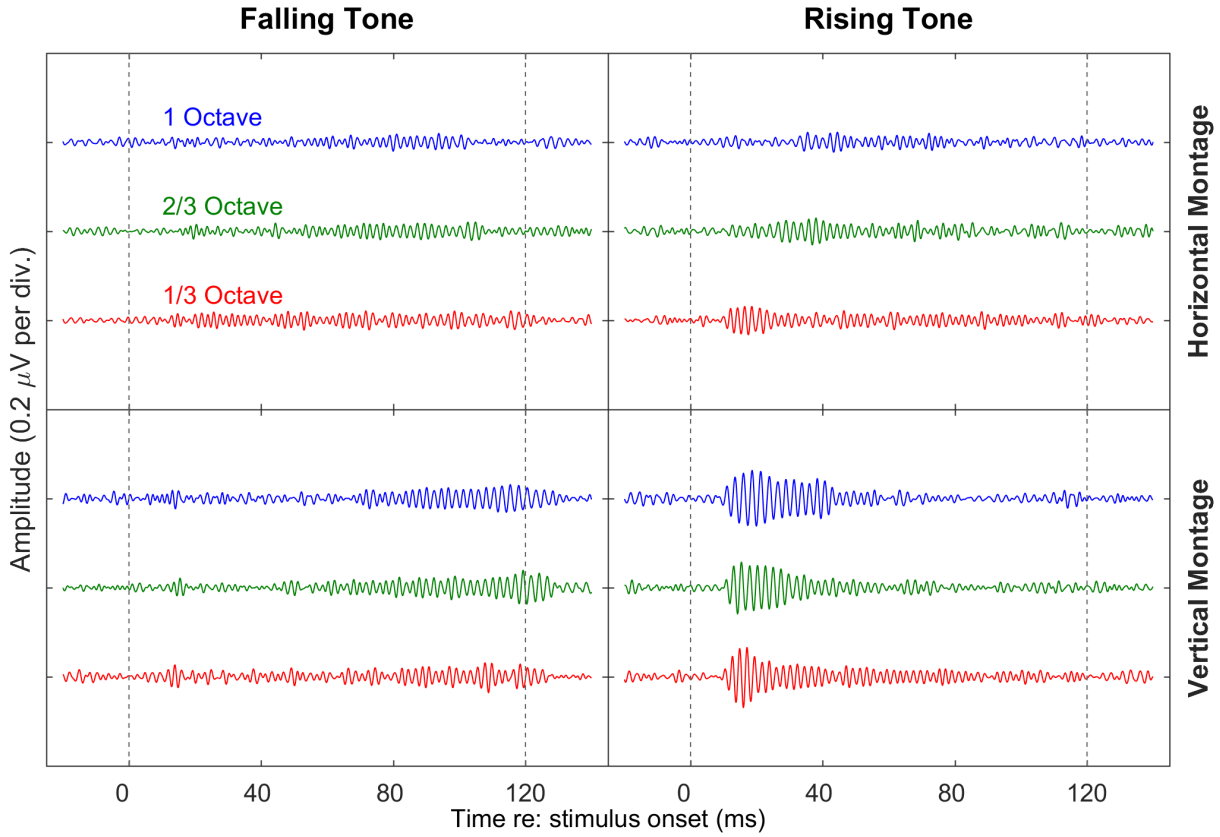

Figure 9. Grand average waveforms by slope, including participants from all groups, but with observations categorized as non-detections excluded. The pattern remains similar to the Figure 2 in the main manuscript, which does not exclude any observations.

Tables

Table 1. Fit parameters for marginal distributions of SRCC and SNR

		Baseline		Response	
Measure (Distribution)	**Param.**	**Horiz.**	**Vert.**	**Horiz.**	**Vert.**
SRCC (Beta)	**α**	1.089	1.067	1.518	1.143
	β	7.085	7.458	1.432	1.325
50-Hz-Mean Spectral Amplitude (Lognormal)	**μ**	0.245	0.442	0.240	0.284
	σ	0.420	0.437	0.401	0.395

Table 1. Fit parameters for marginal distributions of FFR measures across groups and stimulus conditions. SRCC has much better separation between its baseline and response distributions than mean spectral amplitude does.

Table 2. Fit parameters for Student's *t* copula: SRCC

		Montage	
Measure	**Param.**	**Horiz.**	**Vert.**
SRCC	$\hat{\boldsymbol{\rho}}$	-0.043	0.147
	$\hat{\boldsymbol{\nu}}$	42.4	4.67E+06
Amplitude	$\hat{\boldsymbol{\rho}}$	0.403	0.457
	$\hat{\boldsymbol{\nu}}$	11.8	7.52

Table 2. Fit parameters for Student's *t* copula, quantitatively estimating shape and magnitude of correlation between baseline (*Absent*) and response (*Present*) SRCC observations.

Table 3. Parameters for binary logistic regression classifier (SRCC only)

	Montage	
Coefficient	**Horizontal**	**Vertical**
A (amp)	(n/a)	(n/a)
B (SRCC)	22.316342	22.8069
C (intercept)	-5.287225	-5.1879

Table 3. Parameters for binary logistic regression classifier for the SRCC.

Table 4. Repeated-measures ANOVAs: Between-groups effects

Effect	Measure					
	SNR			SRCC		
	df	*F*	*p*	*df*	*F*	*p*
Group	2, 27	5.49	**.010**	2, 27	6.94	**.004**
Extent × Group	4, 54	2.16	.085	4, 54	2.41	.061
Direction × Group	2, 27	.15	.859	2, 27	.07	.936
Montage × Group	2, 27	1.32	.285	2, 27	3.38	**.049**
Window × Group	4, 54	1.79	.144	4, 54	1.65	.175
Extent × Direction × Group	4, 54	.84	.506	4, 54	1.17	.336
Extent × Montage × Group	4, 54	.74	.480	2.9, 38.6	1.78	*.170*
Extent × Window × Group	8, 108	1.11	.362	5.6, 75.1	1.24	*.297*
Direction × Montage × Group	2, 27	.19	.825	2, 27	.34	.718
Direction × Window × Group	2.9, 38.6	.63	*.593*	2.4, 32.4	.42	*.694*
Montage × Window × Group	4, 54	3.59	**.011**	3.3, 44.7	1.13	*.350*
Extent × Direction × Montage × Group	4, 54	.97	.433	4, 54	1.28	.291
Extent × Direction × Window × Group	6.3, 84.9	.81	*.602*	8, 108	1.13	.351
Extent × Montage × Window × Group	8, 108	.57	.805	8, 108	1.12	.356
Direction × Montage × Window × Group	4, 54	.06	.992	3.1, 42.0	.56	*.654*
Extent × Direction × Montage × Window × Group	8, 108	1.08	.383	8, 108	1.24	.286

Table 4. Results of a four-way repeated-measures ANOVAs for the two response measures. Main effects and interactions are listed in the left-hand column, with the associated statistical values for the SNR and SRCC in the middle and right columns, respectively. Bolded *p*-values indicate statistical significance at an alpha level of .05. Italics indicate instances where Greenhouse-Geisser corrections were applied to adjust for nonsphericity. Compare this table with Table I in the main manuscript

Appendix A: R Code for Verification of Model Distributions

```
# INSTALL PACKAGES (only needs to be done once)
install.packages("fitdistrplus")
install.packages("logspline")

# LOAD PACKAGES AND DATA
library(fitdistrplus)
library(logspline)
d <- read.csv([path to data]) # insert actual data path

# CALCULATE FITS
fit.spectAmp.lognormal <- fitdist(d$spectAmp, "lnorm", method="mle")
fit.pre_spectAmp.lognormal <- fitdist(d$pre_spectAmp, "lnorm", method="mle")
fit.srcc.beta <- fitdist(d$srcc, "beta", method="mle")
fit.pre_srcc.beta <- fitdist(d$pre_srcc, "beta", method="mle")

# MAKE PLOTS
plot(fit.pre_spectAmp.lognormal) # Figure 1
plot(fit.spectAmp.lognormal) # Figure 2
plot(fit.pre_srcc.beta) # Figure 3
plot(fit.srcc.beta) # Figure 4
```

Appendix B: Matlab Code for Remaining Analyses

The code used in all the Matlab analyses can be found on the Molis Lab Github: http://www.github.com/mrmolis.