Behavioral/Cognitive

# Optimal Combination of Neural Temporal Envelope and Fine Structure Cues to Explain Speech Identification in Background Noise

**Il Joon Moon,[1] Jong Ho Won,[2] Min-Hyun Park,[3] D. Timothy Ives,[4] Kaibao Nie,[5] Michael G. Heinz,[6,7] Christian Lorenzi,[4] and Jay T. Rubinstein[5]**

[1]Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University, School of Medicine, Seoul 135-710, Korea, [2]Department of Audiology and Speech Pathology, University of Tennessee Health Science Center, Knoxville, Tennessee 37996, [3]Department of Otorhinolaryngology, Seoul Metropolitan Government Boramae Medical Center, Seoul National University, Seoul 156-707, Korea, [4]Equipe Audition (UMR 8248 CNRS LSP), Institut d'Etude de la Cognition, Ecole Normale Superieure, Paris Sciences et Lettres, Paris 75005, France, [5]Virginia Merrill Bloedel Hearing Research Center, Department of Otolaryngology-Head and Neck Surgery, University of Washington, Seattle, Washington 98195, and [6]Department of Speech, Language, and Hearing Sciences and [7]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana 47907

**The dichotomy between acoustic temporal envelope (ENV) and fine structure (TFS) cues has stimulated numerous studies over the past decade to understand the relative role of acoustic ENV and TFS in human speech perception. Such acoustic temporal speech cues produce distinct neural discharge patterns at the level of the auditory nerve, yet little is known about the central neural mechanisms underlying the dichotomy in speech perception between neural ENV and TFS cues. We explored the question of how the peripheral auditory system encodes neural ENV and TFS cues in steady or fluctuating background noise, and how the central auditory system combines these forms of neural information for speech identification. We sought to address this question by (1) measuring sentence identification in background noise for human subjects as a function of the degree of available acoustic TFS information and (2) examining the optimal combination of neural ENV and TFS cues to explain human speech perception performance using computational models of the peripheral auditory system and central neural observers. Speech-identification performance by human subjects decreased as the acoustic TFS information was degraded in the speech signals. The model predictions best matched human performance when a greater emphasis was placed on neural ENV coding rather than neural TFS. However, neural TFS cues were necessary to account for the full effect of background-noise modulations on human speech-identification performance.**

*Key words:* **computational model; neural mechanism; speech perception; temporal cues**

## Introduction

Rate-place and temporal codes are the two primary neural codes for sound perception (Plack and Oxenham, 2005). In terms of frequency coding, the rate-place code utilizes the basilar-membrane place corresponding to the maximum discharge rate of auditory-nerve (AN) responses over the stimulus duration. However, most natural sounds fluctuate over time, therefore it is necessary to consider temporal codes that rely on shorter timescales. Any acoustic signal can be decomposed into the mathematical product of a slowly varying temporal envelope ($ENV_{acoust}$) and a rapidly varying temporal fine structure ($TFS_{acoust}$) (Hilbert, 1912). At the level of AN, $ENV_{acoust}$ is represented as variations in the discharge rate over several milliseconds ($ENV_{neural}$), whereas $TFS_{acoust}$ is represented as phase-locking information to individual cycles of the stimulus waveform ($TFS_{neural}$) (Johnson, 1980; Joris and Yin, 1992). It should be noted that complex mappings exist between the acoustic signal and its neural representation in the peripheral auditory system. For example, when acoustic signals are passed through cochlear filters, there is no one-to-one mapping between acoustic and neural ENV or TFS information (Heinz and Swaminathan, 2009; Shamma and Lorenzi, 2013).

There has been a long-standing debate about the contribution of these two acoustic features to speech intelligibility. $ENV_{acoust}$ information over a few spectral channels is often thought to provide sufficient information for speech intelligibility in quiet (Shannon et al., 1995), whereas $TFS_{acoust}$ information may play

Received March 14, 2014; revised July 24, 2014; accepted July 30, 2014.

Author contributions: I.J.M., J.H.W., M.-H.P., K.N., M.G.H., C.L., and J.T.R. designed research; I.J.M., J.H.W., M.-H.P., and D.T.I. performed research; I.J.M., J.H.W., M.-H.P., M.G.H., C.L., and J.T.R. analyzed data; I.J.M., J.H.W., M.-H.P., M.G.H., C.L., and J.T.R. wrote the paper.

Research subject compensation was supported by an educational fellowship from Advanced Bionics. I.J.M. was supported by the Samsung Medical Center. J.H.W. was supported by National Institutes of Health (NIH) F31-DC009755, the University of Tennessee Health Science Center, the NeuroNet seed grant, the Hearing Health Foundation Grant, and the Todd M. Bader Grant of the Barbara Epstein Foundation, Inc. M.-H.P. was supported by the Boramae Medical Center, Seoul National University. C.L. was supported by a grant from Agence Nationale pour la Recherche (ANR) (HEARFIN project) and ANR-11-0001-02 PSL and ANR-10-LABX-0087. M.G.H. was supported by NIH R01-DC009838.

The authors declare no competing financial interests.

Correspondence should be addressed to Min-Hyun Park, MD, PhD, Department of Otorhinolaryngology, Seoul Metropolitan Government Boramae Medical Center, Seoul National University, Seoul 156-707, Korea. E-mail address: drpark@snu.ac.kr.

DOI:10.1523/JNEUROSCI.1025-14.2014

 0270-6474/14/3412145-10$15.00/0

an important role when speech is presented against a complex background noise (Gnansia et al., 2009; Hopkins and Moore, 2009). However, caution should be taken to interpret these previous conclusions because $TFS_{acoust}$ and $ENV_{acoust}$ do not factor in the neural representations of these acoustic features. Furthermore it is still unclear how the central auditory system utilizes the neural information that is processed and conveyed from the peripheral auditory system.

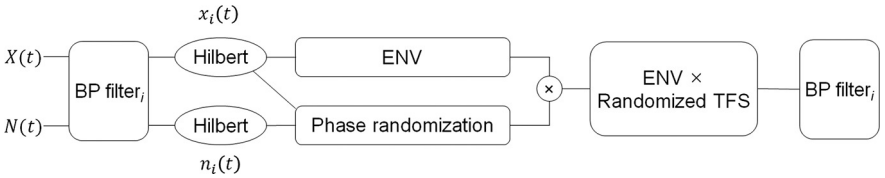


**Figure 1.** Schematic diagram of signal processing to randomize Hilbert phase is shown. See text for details. BP, bandpass.

The current study approaches the question of the neural coding of speech using psychoacoustic experiments and computational models. A recent study by Swaminathan and Heinz (2012) evaluated the contribution of peripheral $ENV_{neural}$ and $TFS_{neural}$ to speech perception in noise using a simple regression model, and showed that $ENV_{neural}$ was the primary contributor to speech perception. In their study, $TFS_{neural}$ contributed mainly in the presence of $ENV_{neural}$, but rarely as the primary cue itself. However, the question of how the central auditory system utilizes such $ENV_{neural}$ and $TFS_{neural}$ information is largely unknown. To address this question, we gradually jittered the acoustic phase cues to degrade $TFS_{acoust}$ over 32 frequency channels with an expectation that the jittered acoustic phase cues may affect neural synchrony in AN fibers. We simulated the AN response using a computational model of peripheral auditory processing (Zilany and Bruce, 2006, 2007) to evaluate how degraded $TFS_{acoust}$ cues affect the encoding of $TFS_{neural}$ and $ENV_{neural}$ information. Finally, we processed the peripheral neural information using a computational neural-observer model and compared human speech identification to model predictions. In doing so, we determined the optimal combination of $ENV_{neural}$ and $TFS_{neural}$ information needed to account for speech intelligibility in steady and fluctuating noise.

## Materials and Methods

*Subjects.* Six native speakers of American English participated (four females and two males; five subjects between the ages of 28 and 30 and one 23 year old; mean age 27.8 years). All subjects had audiometric thresholds of 20 dB HL or less at octave frequencies between 250 and 8000 Hz in both ears. The current study was approved by the University of Washington Institutional Review Board.

*Phase vocoder processing.* Figure 1 shows a schematic diagram of the Hilbert phase randomization procedure. Input waveforms, $X(t)$, were filtered into 32 channels using an array of finite impulse-response analysis filters. They were evenly spaced on an $ERB_N$ scale between 100 and 10,000 Hz, where $ERB_N$ refers to the average value of the equivalent rectangular bandwidth (ERB) of the auditory filter for young, normal-hearing listeners at moderate sound levels (Glasberg and Moore, 1990). The bandwidth for each channel was set to 1-$ERB_N$ wide with the assumption that if vocoding is performed with analysis channels that have a similar bandwidth to auditory filters (i.e., 1-$ERB_N$ wide), the effect of the phase randomization will be primarily restricted to the temporal information encoded via neural phase locking, while the spectral information available to the central auditory system (encoded spatially on the basilar membrane in the cochlea) will be only minimally affected, if at all. In other words, the present approach explored only the role of within-channel (neural) temporal cues for speech identification. The Hilbert transform (Hilbert, 1912) was used to decompose each sub-band signal into its $ENV_{acoust}$ and $TFS_{acoust}$. The absolute value of each sub-band analytic signal, $x_i(t)$, was taken as the sub-band $ENV_{acoust}$. No additional processing was applied to the sub-band $ENV_{acoust}$. The $TFS_{acoust}$ for each sub-band was computed as the cosine value of the angle of the analytic signal. The same analysis filters were used to filter a wideband noise, $N(t)$, to generate a band-limited noise carrier for each sub-band, $n_i(t)$. The root mean square (RMS) value of $n_i(t)$ was set to the same RMS value as $x_i(t)$. The following equation was used to vary the extent of phase randomization:

$$y_i(t) = abs(x_i(t)) \times \cos(angle([(1 - NF) \times x_i(t)] + [NF \times n_i(t)])), \quad (1)$$

in which, $y_i(t)$ is the output stimulus, $x_i(t)$ is the analytic signal of the $i$th analysis channel, $n_i(t)$ is the filtered noise of the $i$th analysis channel in an analytic form, and $NF$ is a "noise factor" from 0 to 1. As shown in Equation 1, the weighted random noise component, analytic signal $[NF \times n_i(t)]$, was added to the weighted original sub-band analytic signal $[(1 - NF) \times x_i(t)]$. Then, the randomized $TFS_{acoust}$ was obtained by taking the cosine value of the angle of these mixed signals. The randomized $TFS_{acoust}$ was then modulated with the $ENV_{acoust}$ of the sub-band signal. Each modulated sub-band signal was subsequently filtered with the initial analysis filters. The real values of the filtered signals were then summed over all channels. For a pilot study, different types of equations were tested for the phase randomization process such as $\cos(angle(x_i(t)) + angle(NF \times n_i(t)))$, or $\cos(angle(x_i(t)) \times [n_i(t) \times e^{-jw_0(t)}]^{NF}))$, where $w_0$ is the center frequency of each analysis channel. Equation 1 was chosen because when correlations between the Hilbert $TFS_{acoust}$ of a vocoded signal and the corresponding original signal were computed, Equation 1 shows that, as a function of NF, the correlation coefficients decrease in a more linear fashion (Fig. 4*A*) than other functions tested.

*Human speech-identification test procedure.* The target sentences were taken from Institute of Electrical and Electronics Engineers (IEEE) sentences (Rothauser et al., 1969), sampled at 20,000 Hz, spoken by a female speaker. The IEEE sentences were presented in either steady or amplitude-modulated, speech-shaped noise. The steady noise maskers were spectrally shaped to have the same long-term power spectrum as the IEEE sentences. The modulated noise maskers were generated by amplitude modulating the steady noise with an 8 Hz sinusoid on a logarithmic scale with a peak-to-valley ratio of 30 dB. In all trials, maskers were gated on and off with 50 ms linear ramps 500 ms before and 50 ms after the target sentences. The mixture of the target sentence and masker stimuli were then vocoded and presented monaurally to the right ear via an ER3-A insert phone at an overall target speech level of 65 dB SPL. Before actual testing, subjects listened to vocoded sentences that were processed with NF of 1 in the absence of background noise while they were presented with the sentences on the computer screen.

Speech reception thresholds (SRTs) corresponding to the 50% intelligibility level were measured using a one-up, one-down adaptive procedure. For each testing condition, subjects identified the IEEE sentences in the presence of either steady or modulated noise. Each test run started with a signal-to-noise ratio (SNR) of 6 dB, for which subjects were easily able to identify the words in the target sentence correctly. If subjects correctly identified three or more of the five keywords, the response was counted as correct and the SNR for the next sentence was decreased. If two or fewer keywords were correctly identified, the response was counted as incorrect and the SNR for the next sentence was increased. The level of the target sentence was always fixed, but the level of background noise was varied in an adaptive manner. An initial step size of 4 dB was used for the first two reversals in the adaptive track, after which the step size was fixed at 2 dB for the next six reversals. The SRT for a given track was based on the average of the SNRs corresponding to each of the last six reversals in the adaptive track. No target sentence was repeated to any subject. NF values of 0, 0.1, 0.2, 0.3, 0.5, 0.7, and 1.0 were tested in random order. For each NF value, two adaptive tracks were

completed and a further adaptive track was obtained if the difference of the first two tracks exceeded 3 dB. Only about 7% of conditions across six subjects required the third adaptive track. The final threshold for each NF value was the mean of these two (or three) adaptive tracks.

*Computational model of peripheral auditory processing.* A phenomenological model of the AN (Zilany and Bruce, 2006, 2007) was used to simulate spike-train responses in AN fibers to the same stimuli used in the human sentence-identification test. This model is an extension of a previously established model that has been tested extensively against neurophysiological data obtained from animals in response to both simple and complex stimuli, such as tones, broadband noise, and speech-like signals (Carney, 1993; Heinz et al., 2001; Zhang et al., 2001; Bruce et al., 2003; Tan and Carney, 2003; Zilany and Bruce, 2006, 2007). The model incorporates diverse nonlinear physiological properties of the cochlea, including compression, suppression, broadened tuning, and best-frequency shifts with increases in sound level. Inputs to the AN model were the IEEE sentence waveforms, and the output of the model was a set of spike times for four high spontaneous rate AN fibers with characteristic frequencies (CFs) of 200, 464, 1077, and 2500 Hz (CF is the frequency at which the fiber responds at the lowest sound level). These four CFs were chosen because phase locking considerably decreases above ~2.5 kHz (Johnson, 1980).

To quantify the similarity between neural ENV or TFS responses to two different stimuli, the two sets of predicted AN spike trains were compared by computing neural cross-correlation coefficients ($\rho_{ENV}$ and $\rho_{TFS}$) using established techniques (Heinz and Swaminathan, 2009). Briefly, for two different stimuli passed through the AN model, one serves as a reference stimulus (A) and the other serves as a test stimulus (B). To create test stimuli, two different types of speech degradation were used: phase vocoding and adding background noise. These stimuli were passed through the AN model (Zilany and Bruce, 2006, 2007) to simulate AN spike-train responses. Shuffled autocorrelogram and shuffled cross-correlogram (SAC and SCC) analyses (Joris, 2003; Louage et al., 2004; Heinz and Swaminathan, 2009) of these spike-train responses were performed to compute neural cross-correlation coefficients ($\rho_{ENV}$ and $\rho_{TFS}$). These correlation coefficients describe the similarity in neural representations of ENV or TFS between the reference and test stimuli. Simulated AN spike-train responses were obtained to each of the reference (A) and test (B) stimuli, and their polarity-inverted waveforms (e.g., A+ with A−, and B+ with B−). For a given stimulus, AN responses were generated for 20 repetitions at each CF. Here, polarity inversion results in inverting the TFS but keeping ENV the same.

To quantify the strength of TFS or ENV coding to a given stimulus (A), simulated spike trains from A+ and A− were compared by creating histograms of the intervals between spikes across all pairs of repetitions (i.e., via shuffling). The shuffled autocorrelogram [SAC(A+)] was computed from spike intervals in the spike-train responses to stimulus A+, whereas the shuffled cross-polarity correlogram [SCC(A+, A−)] was computed from spike intervals between the A+ and A− spike trains. Following normalization of the histograms, TFS and ENV coding was characterized by computing "difcor" and "sumcor " functions, respectively, as follows:

$$\mathrm{difcor_A} = \mathrm{SAC(A+)} - \mathrm{SCC(A+, A-)}, \quad (2)$$

$$\mathrm{sumcor_A} = [\mathrm{SAC(A+)} + \mathrm{SCC(A+, A-)}]/2. \quad (3)$$

The peak height of the difcor or sumcor at a delay (spike interval) of zero represents the strength of TFS and ENV, respectively. To quantify the overall strength of common TFS or ENV coding between responses to stimuli A and B, a difcor or sumcor function was computed based on shuffled cross-stimulus correlograms as follows:

$$\mathrm{difcor_{AB}} = \mathrm{SCC(A+, B+)} - \mathrm{SCC(A+, B-)}, \quad (4)$$

$$\mathrm{sumcor_{AB}} = [\mathrm{SCC(A+, B+)} + \mathrm{SCC(A+, B-)}]/2. \quad (5)$$

Using these difcor and sumcor functions, neural cross-correlation coefficients ranging between 0 and 1 were computed by comparing the degree of common TFS or ENV coding between A and B relative to the degree of TFS or ENV coding to A and B individually (Heinz and Swaminathan, 2009) as follows:

$$\rho_{TFS} = \frac{\mathrm{difcor_{AB}}}{\sqrt{\mathrm{difcor_A} \times \mathrm{difcor_B}}}, \quad (6)$$

$$\rho_{ENV} = \frac{\mathrm{sumcor_{AB}} - 1}{\sqrt{(\mathrm{sumcor_A} - 1) \times (\mathrm{sumcor_B} - 1)}}. \quad (7)$$

Values of $\rho_{ENV}$ and $\rho_{TFS}$ range from near 0 for completely uncorrelated reference/test stimuli (noise floor for $\rho_{ENV}$ and $\rho_{TFS} \cong 0.1$ and 0.01, respectively) to near 1 if the reference and test stimuli were perfectly correlated. For a complete illustration of the computation of $\rho_{ENV}$ and $\rho_{TFS}$ metrics, see Heinz and Swaminathan (2009).

*Computational model of the central neural observer.* To gain further insight on how the central auditory system utilizes neural TFS and ENV information for speech perception, the sentence-identification test for human subjects was implemented using neural observers. The neural-observer sentence-identification test was conducted using the same testing paradigm as with human subjects. Human subjects were tested on an "open-set" IEEE sentence test; however, to make the computational speech-identification test more tractable, the central neural observer was provided with a set of "exemplar" sentences. Figure 2 shows the block diagram of the neural-observer computation. In each trial, the testing program randomly selected a "target" sentence out of the 50 exemplar sentences, and the peripheral model outputs were then compared between the target sentence and each of the 50 exemplar sentences. Both the target and 50 exemplars were degraded by both noise and phase vocoding at the corresponding SNR and NF. For example, for the 0 dB SNR and NF = 1 condition, when the testing program selected a target sentence randomly from the pool of 50 exemplar sentences, the background noise was added to that sentence at 0 dB SNR and the mixture of the target sentence and noise was passed through the phase vocoder with NF = 1. Here, fresh noise was generated and used for each sentence. The resulting vocoded signal was then provided to the peripheral model as a target stimulus. The 50 sentences were also degraded at the corresponding SNR and NF values; for this example, 0 dB and NF = 1. The degraded 50 exemplar sentences were provided to the peripheral model one by one. Therefore, to make a single decision, 50 different mean (across 4 CFs) $\rho_{ENV}$ and $\rho_{TFS}$ values were computed.

The neural-observer model used in this study was based upon assumptions that the central auditory system may optimally use the sensory information provided by the peripheral auditory system for the best possible speech-identification performance, depending on the acoustic environment. To simulate such a decision process, different weightings were applied to the mean $\rho_{ENV}$ and $\rho_{TFS}$ values to compute a final decision metric for the neural observer. The following equation shows this computation:

$$\rho_{ENV\_TFS\_COMBINED} = (\alpha \times \rho_{ENV}) + ((1 - \alpha) \times \rho_{TFS}), \quad (8)$$

in which $\alpha$ is a weighting coefficient for $\rho_{ENV}$, and subsequently, $1 - \alpha$ is a weighting for $\rho_{TFS}$. Five different sets of weighting coefficients were tested ($\alpha = 1$, 0.75, 0.5, 0.25, and 0).

The above procedure was performed independently for all the weightings to compare the predicted speech-identification performance by the neural-observer model. The central model had a two-step decision process. First, 50 $\rho_{ENV\_TFS\_COMBINED}$ values (for 50 exemplar sentences) were scanned to determine the sentence that produced the highest $\rho_{ENV\_TFS\_COMBINED}$ value. This sentence was saved as the predicted sentence by the central model. Second, the testing program compared the predicted sentence with the originally selected target sentence, and recorded the central neural-observer's response as either correct or incorrect. As with the speech test for human subjects, the neural-observer testing program started with an SNR of 6 dB. If the neural-observer model was not able to predict sentence identification, such a condition is marked as "n/a" in the data plot. For each testing condition, the mean threshold from six model runs was obtained.

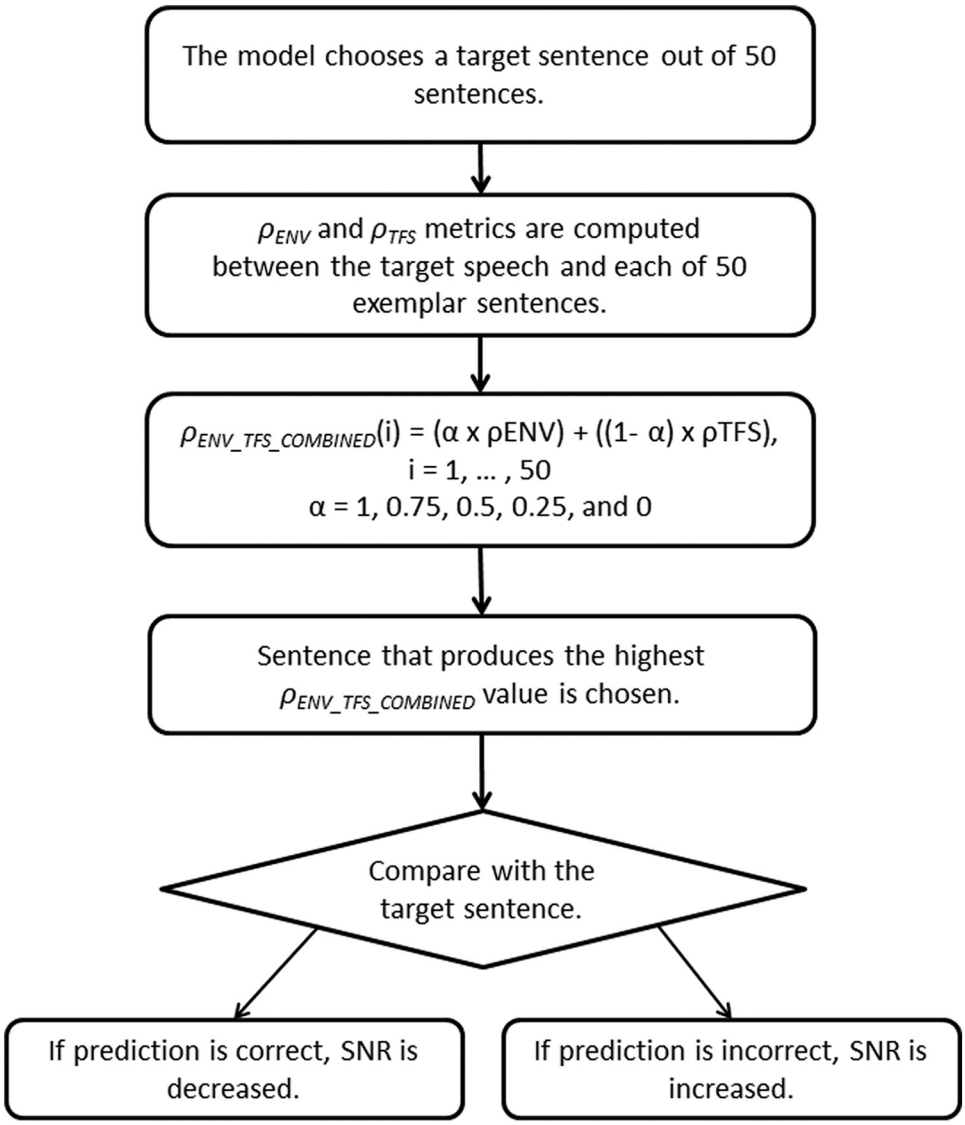


**Figure 2.** Block diagram for the neural-observer computation. See text for details.

## Results

### Sentence identification by human listeners

Figure 3*A* shows SRTs as a function of NF values. Here, a lower SRT value implies better speech-identification performance in background maskers. In general, performance was better in modulated noise (filled triangles) than in steady noise (open squares). A two-way repeated-measures ANOVA was performed to assess the effects of NF and masker type (steady or modulated noise). The main effects of NF ($F_{(6,30)} = 31.3, p < 0.001$) and masker type ($F_{(1,5)} = 142.0, p < 0.001$) were all highly significant. That is, as NF increased (i.e., more phase randomization), SRTs increased in both maskers, but the rate of increase for SRT values differed between the two masker types. The interaction between NF and masker type also reached significance ($F_{(6,30)} = 3.18, p = 0.015$), indicating that the contribution of $TFS_{acoust}$ is greater for SRTs in modulated noise.

Speech-masking release was calculated as the SRT for steady noise minus the SRT for modulated noise. Figure 3*B* shows the speech-masking release as a function of NF. The amount of masking release decreased from 9.4 to 6.8 dB as NF increased from 0 to 1, creating a difference of 2.6 dB. This size of speech-masking release effect is consistent with Hopkins and Moore (2009), where they tested normal-hearing listeners using tone vocoders that preserved either 0% or 100% TFS across 32 frequency bands. There was no difference in SRTs between the first and second adaptive tracks (paired *t* tests, $p > 0.05$ for both maskers) in the sentence-identification test (i.e., there was no training effect observed).

### Quantifying the effect of phase randomization on the $TFS_{acoust}$, $TFS_{neural}$, and $ENV_{neural}$ information

To estimate how the phase randomization varied the availability of $TFS_{acoust}$ information in the vocoded signal, correlations between the Hilbert $TFS_{acoust}$ (i.e., the cosine component of the analytic signal) of a vocoded signal and the corresponding original signal were computed. Ten sentences were used for this analysis. Figure 4*A* shows the mean correlation coefficients between the Hilbert $TFS_{acoust}$ of the original and vocoded signals. As NF increased from a value of 0 to 1, there was a gradual decrease in the correlation coefficients. When NF = 1 was used, the original $TFS_{acoust}$ information was completely degraded.

To estimate how the phase randomization varies the availability of $TFS_{neural}$ information in the AN, an IEEE sentence was

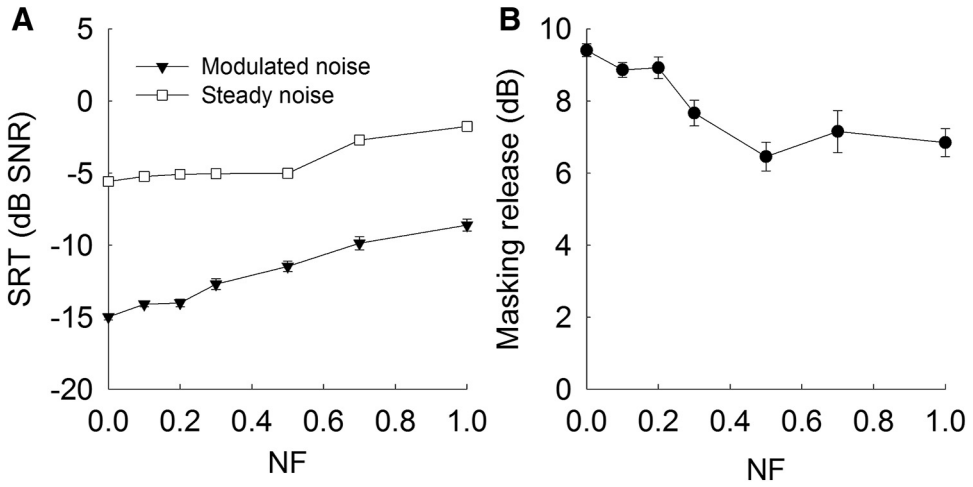


**Figure 3.** ***A***, SSRTs measured from six human listeners with normal hearing in steady and modulated noise maskers as a function of NF. Error bars show 1 SEM of the mean across six subjects. ***B***, Speech-masking release plotted as a function of NF. Error bars show 1 SEM across six subjects. Note that some error bars are invisible because the sizes of symbols are larger than the sizes of those error bars.

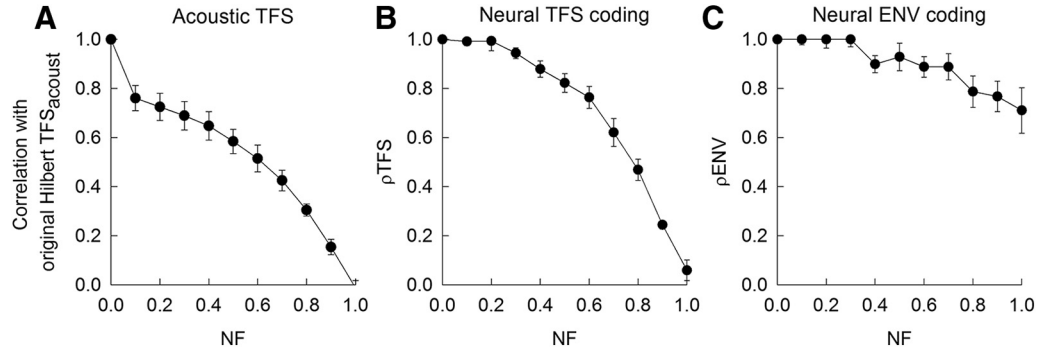


**Figure 4.** ***A***, Mean correlation between the Hilbert $TFS_{acoust}$ of a vocoded signal and the corresponding original signal. Error bars show 1 SD of the mean across 10 sentences. ***B***, ***C***, $TFS_{neural}$ and $ENV_{neural}$ values, where the neural cross-correlation coefficients were computed between model spike-train responses to the vocoded speech and a sentence in quiet processed with NF = 0 (as a reference stimulus). Mean $\rho_{TFS}$ and $\rho_{ENV}$ values across four AN fibers are plotted with SE bars. The IEEE sentence, the birch canoe slid on the smooth planks, was used.

processed with 11 NF values from 0 to 1 with a step size of 0.1. A sentence in quiet processed with NF = 0 served as a reference stimulus. Figure 4*B* shows mean values of $\rho_{TFS}$ averaged across four CFs between reference and vocoded stimuli as a function of NF. As expected, $\rho_{TFS}$ values decreased all the way to zero as NF increased from 0 to 1, which was consistent with the pattern of acoustic correlation coefficients presented in Figure 4*A*. This suggests that the phase randomization in the acoustic domain successfully varies the availability of $TFS_{neural}$ information in the AN.

It should be noted that the phase randomization in the acoustic domain may also degrade the encoding of $ENV_{neural}$ cues, because of the conversion of the frequency excursions of $TFS_{acoust}$ into dynamic variations of the output levels of the cochlear filters (i.e., $ENV_{neural}$; Ghitza, 2001). The same approach was thus used to estimate the extent to which the phase randomization also alters the availability of $ENV_{neural}$ information in the AN. Figure 4*C* shows mean values of $\rho_{ENV}$ averaged across four CFs for reference and vocoded speech as a function of NF. A sentence in quiet processed with NF = 0 served as a reference stimulus. As expected, $\rho_{ENV}$ values slightly decreased as NF increased from 0 to 1. Figure 4, *B* and *C*, also shows that the phase randomization in the acoustic domain had a differential effect on $TFS_{neural}$ and $ENV_{neural}$.

### Quantifying the effect of phase randomization on $TFS_{neural}$ and $ENV_{neural}$ information in the presence of background noise

The effects of noise on the neural coding of TFS and ENV for phase-vocoded speech are shown in Figure 5. Overall, a dynamic pattern of $\rho_{TFS}$ and $\rho_{ENV}$ values was observed. On the left side of Figure 5, mean $\rho_{TFS}$ values averaged across four CFs in steady and modulated noise are shown on the top and bottom, respectively. Symbols represent different SNR conditions. Error bars show 1 SEM across four CFs. At each NF value, $\rho_{TFS}$ values decreased as SNR decreased. At NF = 1, $\rho_{TFS}$ became zero, indicating that the vocoded sentence produced completely different neural TFS coding from the original sentence. At positive SNRs, $\rho_{TFS}$ values were relatively high and changed little until NF reached 0.5, but beyond NF = 0.5, there was a rapid decrease in $\rho_{TFS}$ values. At negative SNRs, $\rho_{TFS}$ values were already low, even at NF = 0. This observation illustrates the deleterious effect of adding background noise on the speech-related TFS coding in the AN. The right side of Figure 5 shows mean values of $\rho_{ENV}$ averaged across four CFs in steady and modulated noise. $\rho_{ENV}$ values measured for speech masked by noise tended to decrease as NF increased from 0 to 1, but the slopes of the functions were shallower than those for $\rho_{TFS}$. These data are consistent with the simulation data obtained in quiet (Fig. 4).

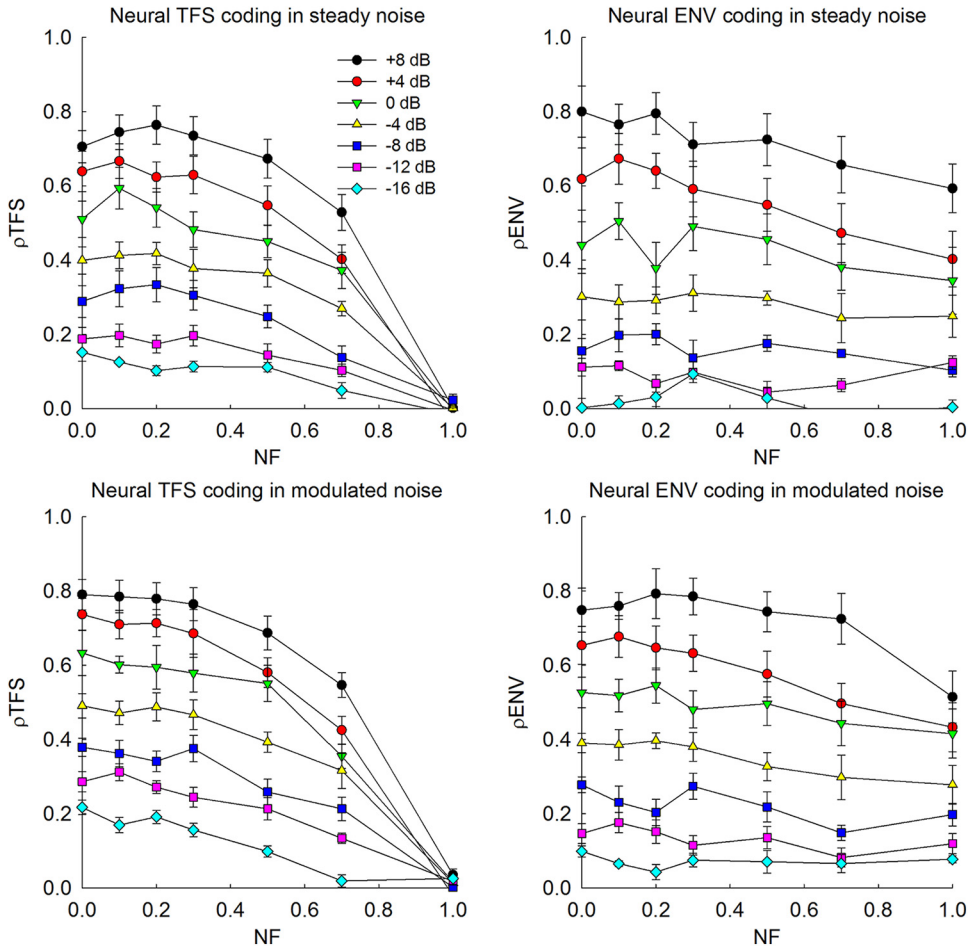


**Figure 5.** Neural coding of TFS and ENV in background noise, where the neural cross-correlation coefficients were computed between model spike-train responses to the noisy vocoded speech and a sentence in quiet processed with NF = 0 (as a reference stimulus). Symbols show different SNR conditions. Mean values and SEs across four CFs are shown. The IEEE sentence, the birch canoe slid on the smooth planks, was used.

### Sentence identification predicted by the central neural observer

Figure 6 shows SRTs predicted by the central neural-observer model and obtained from six human subjects (black circles). For clarity, error bars are not plotted in Figure 5, but SEs across six model runs were generally below 1 dB. SEs across human subjects are shown in Figure 3*A* and they are <0.5 dB. For steady noise, the central model generally outperformed human subjects, but for modulated noise, the range of model predictions overlapped with the human data. Some patterns of the human data were well depicted by the central model. For example, the predicted SRTs of the central model consistently increased as a function of NF. When comparing steady and modulated noise, the model predicted substantially lower SNRs for modulated noise than for steady noise, consistent with human data.

The central neural-observer data were generated using five different sets of weightings for $\rho_{ENV}$ and $\rho_{TFS}$ to examine the respective effects of $ENV_{neural}$ and $TFS_{neural}$ on the intelligibility of speech (see Materials and Methods for details). When the $ENV_{neural}$ weighting coefficient $\alpha$ was set to 1, the central model used 100% of $\rho_{ENV}$ and 0% of $\rho_{TFS}$ to conduct the sentence-identification task. Likewise, if the $ENV_{neural}$ weighting coefficient $\alpha$ was set to 0, the central model used $\rho_{TFS}$ only without using $\rho_{ENV}$. To quantify the difference between the human and central model predicted SRTs, the mean squared error (MSE) was computed for each model condition as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Observed\ SRT_i - Predicted\ SRT_i)^2, \quad (9)$$

in which $N$ indicates the number of NF values considered for computation. Except for the $\rho_{ENV}$ weighting coefficient of 0, MSE values were computed over seven NF values. For the $\rho_{ENV}$ coefficient of 0, MSE values were computed over six NF values including 0, 0.1, 0.2, 0.3, 0.5, and 0.7. Here, smaller MSE values indicate a better fit between the observed and predicted SRTs. As shown in Table 1, the central neural-observer data for the $\rho_{ENV}$ coefficient of 1 predicted the human data most accurately in steady noise. As more weighting was applied to $\rho_{TFS}$, the central model prediction showed a substantial decrease in SRTs in steady noise (Fig. 6), resulting in greater differences between the data from the model and human subjects (Table 1). Somewhat different patterns were observed for modulated noise. When the $\rho_{ENV}$ weighting coefficient of 0.75 was used, the smallest MSE value was observed between the human and central neural-observer model data.

### Speech-masking release predicted by neural observers

Figure 7 shows speech- masking release predicted by the central model, plotted with the data observed from human subjects. To compute the central neural-observer model prediction of speech-masking release, the $\rho_{ENV}$ weighting coefficients that showed the

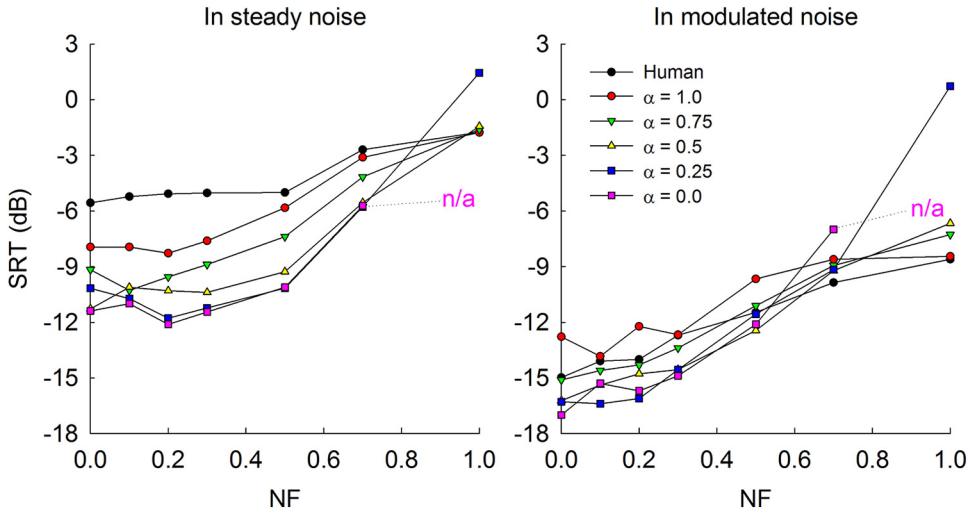


**Figure 6.** Predicted SRTs by the central neural observer in steady and modulated noise as a function of NF. Black circles represent mean SRTs averaged across six human listeners. Symbols (red circles, green reversed triangles, yellow triangles, blue squares, and purple squares) show different $\rho_{ENV}$ weighting coefficients for the neural-observer predictions. Error bars are not plotted for clarity, but SEs across six model runs were generally below 1 dB. SEs across six human subjects can be found in Figure 3*A*. Note that because the neural-observer model was unable to perform the speech-identification task for NF = 1 and the $\rho_{ENV}$ weighting of 0, the data point for that condition is shown as n/a.

**Table 1. MSE between the SRTs observed from human subjects and predicted by the central neural-observer model**

| $\rho_{ENV}$:$\rho_{TFS}$ | 1:0 | 0.75:0.25 | 0.5:0.5 | 0.25:0.75 | 0:1 |
|---|---|---|---|---|---|
| Steady noise | 4.40 | 12.05 | 19.68 | 24.41 | 31.45 |
| Modulated noise | 1.86 | 0.59 | 1.57 | 14.62 | 3.64 |

minimum MSE values were considered, i.e., the "best-fit" to observed SRTs was used. For steady noise, the $\rho_{ENV}$ weighting coefficient of 1 was used, and for modulated noise, the $\rho_{ENV}$ weighting coefficient 0.75 was used as these weighting coefficients produced the minimum MSE values. In addition, $\rho_{ENV}$ weighting coefficients of 1 and 0.5 also showed fairly small MSE values for modulated noise, so these ratios were also considered. In Figure 7, the central neural-observer model prediction of speech-masking release followed the pattern of human data; that is, speech-masking release decreased as NF increased. This result suggests that both human subjects and the neural-observer model increased their masking release when accurate $TFS_{acoust}$ cues were present (i.e., NF = 0) compared with a listening condition where $TFS_{acoust}$ cues were absent (i.e., NF = 1). Comparing the three model predictions, greater speech-masking release was shown when a higher weighting was given to $\rho_{TFS}$. Quantitatively, MSE values for the model prediction with the $\rho_{ENV}$ weighting coefficient of 0.5 for modulated noise was 2.11, whereas MSE values for the model prediction with the $\rho_{ENV}$ weighting coefficients of 0.75 and 1 for modulated noise were 3.89 and 10.16, respectively.

### Effects of model parameters

In the current study, all model simulations were obtained with the functionality of inner and outer hair cells set to normal hearing (i.e., $C_{IHC}$ and $C_{OHC}$ = 1.0) and the AN fibers had a high spontaneous rate (50 spikes per second). Input stimuli were resampled to 100 kHz before presentation to the model and scaled to best modulation level (BML). Here, BML refers to the sound level that produces maximal neural ENV coding. BMLs are typically ~15 dB above AN fiber threshold (Joris and Yin, 1992). In this study, BMLs were computed at four different CFs for multiple input sentences to determine the overall BML, which was finally set to 35 dB SPL for all simulations in the present study. Thus, the sound level of the individual sentences used in the modeling was 35 dB SPL, which was 30 dB below the sound level used in the human speech-identification tests. However, differences in input sound level between the auditory model and human behavioral testing are consistent with many previous AN modeling studies that attempted to simulate performance of human subjects in speech-identification tests using neural responses of high spontaneous rate AN fibers (Swaminathan and Heinz, 2012; Chintanpalli and Heinz, 2013). In the AN model used in this study (Zilany and Bruce, 2006, 2007), thresholds of high spontaneous rate fibers were fitted to the lowest thresholds observed in cats (−5 dB SPL at mid-frequencies), and these fibers have rate-level functions with the typical 30–40 dB dynamic range (Miller et al., 1997). Thus, the sound levels that produced the maximum amount of neural ENV coding (best modulation levels) are quite low in the model fibers used in this study. However, given the fact that there is typically a wide range of thresholds across AN fibers, it is reasonable to expect that some fibers would exhibit essentially the same maximal neural ENV coding properties at 65 dB SPL (Joris and Yin, 1992).

To compare the effects of input sound level on the neural TFS and ENV coding, ρTFS and ρENV values were computed across four CFs between 200 and 2500 Hz. These simulations were performed using 10 different IEEE sentences presented in modulated noise at −15 dB SNR and processed with NF = 0, since the human subjects showed an average SRT of −15 dB for this listening condition. Figure 8*A* shows mean ρTFS and ρENV values averaged across four CFs as a function of NF for input stimulus levels of 35 and 65 dB SPL. As discussed above, a similar pattern of change in ρTFS and ρENV values as a function of NF was observed at both 35 and 65 dB SPL, suggesting that the absolute difference in input sound level between the model and behavioral speech-identification tests is unlikely to limit the conclusions from the current study.

In the current study, only four CFs were included between 200 and 2500 Hz for the model simulations to reduce computation time, but it may be more realistic to include higher CFs to reflect the different patterns of the AN encoding across different CFs. To evaluate the potential effect of including higher CFs, ρENV was computed for seven CFs between 200 and 8000 Hz. Note that

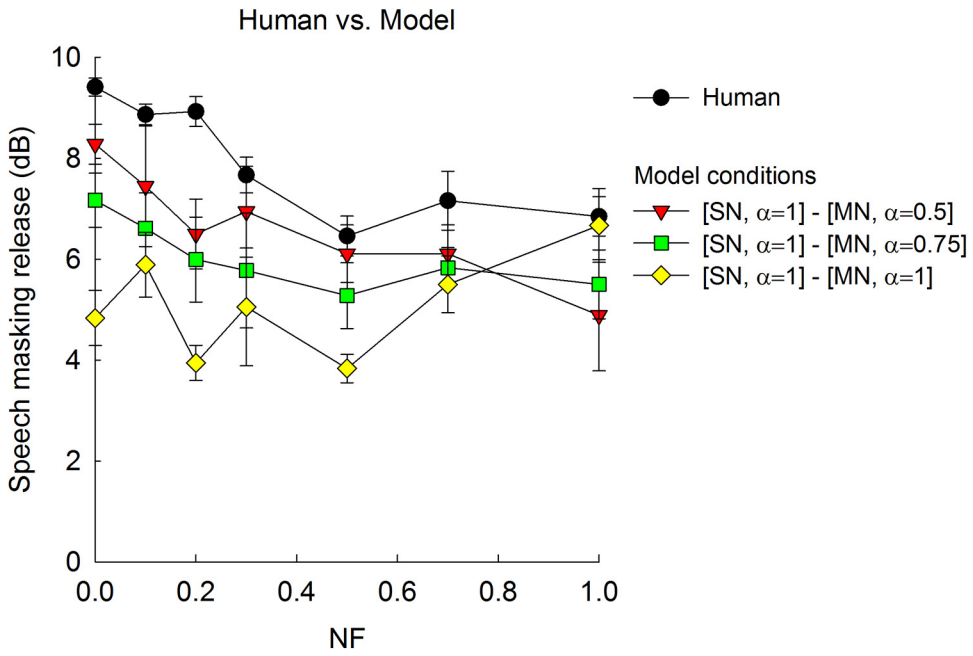


**Figure 7.** Predicted speech-masking release by the central neural-observer model. Black circles represent mean speech-masking release found for six human listeners. Red triangles represent speech-masking release predicted using the $\rho_{ENV}$ weighting of 1 for steady noise (SN) and 0.5 for modulated noise (MN). Green squares represent speech-masking release predicted using the $\rho_{ENV}$ weightings of 1 for SN, and 0.75 for MN. Yellow rhombuses represent speech-masking release predicted using the $\rho_{ENV}$ weighting of 1 for SN as well as for MN. Error bars represent 1 SE across six subjects (for human data) and across six repetitions (for the model data).

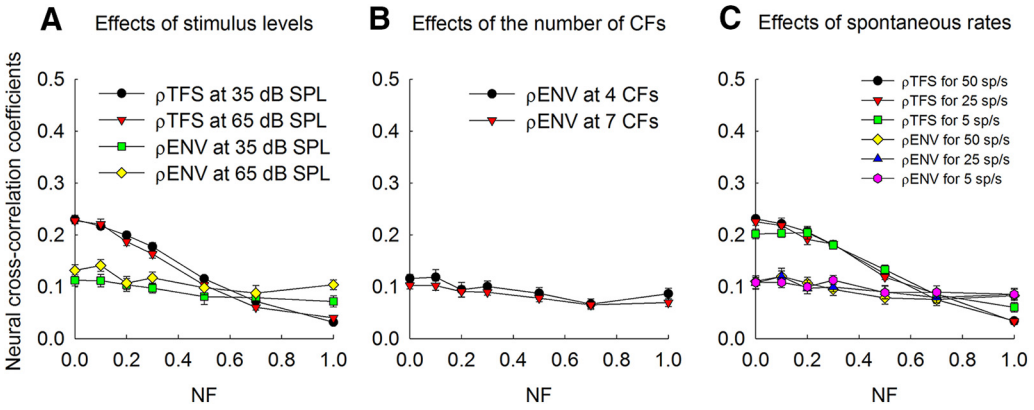


**Figure 8.** Effects of input sound level, the number of CFs, and the spontaneous rate of the AN fibers on the neural TFS and ENV are evaluated. For these simulations, 10 different IEEE sentences were used. IEEE sentences were presented in modulated noise at −15 dB SNR. $\rho$TFS and $\rho$ENV values were computed between the phase vocoded (with NF = 0) signals of the mixture (i.e., target IEEE sentence in the presence of modulated noise) and the corresponding target IEEE sentence (processed with NF = 0). ***A***, Mean $\rho$TFS and $\rho$ENV values averaged across four CFs (200–2500 Hz) for input stimulus of 35 and 65 dB SPL. The AN fibers were set to have a spontaneous rate of 50 spikes per second for these simulations. ***B***, Mean $\rho$ENV values averaged across four CFs (200–2500 Hz) and seven CFs (200–8000 Hz). For these simulations, stimuli were presented at 35 dB SPL and the model was run with the spontaneous rate of 50 spikes per second. ***C***, Mean $\rho$TFS and $\rho$ENV values averaged across four CFs (200–2500 Hz) for three different spontaneous rates (5, 25, and 50 spikes per second). For these simulations, stimuli were presented at 35 dB SPL. Error bars represent 1 SE across 10 IEEE sentences.

neural TFS processing is substantially decreased above 2500 Hz (Johnson, 1980), thus only $\rho$ENV values were compared for four CFs (200–2500 Hz) and seven CFs (200–8000 Hz). Figure 8*B* shows that $\rho$ENV values averaged across four CFs or seven CFs were very similar, suggesting that including more CFs would not likely change the conclusions of the present study.

In the current study, all model simulations were obtained with high spontaneous rate (50 spikes per second) fibers, because the AN model has not been validated for low spontaneous rate fibers. In Figure 8*C*, mean $\rho$TFS and $\rho$ENV values averaged across four CFs for three different spontaneous rates (5, 25, and 50 spikes per second) were compared as a function of NF. For these three different spontaneous rates, similar patterns of $\rho$TFS and $\rho$ENV values were observed.

## Discussion

### Neural ENV and TFS coding for speech identification

In this study, a long-standing question of the neural mechanisms of speech identification in noise was explored. Our approach was unique in two aspects. From the acoustic perspective, previous vocoder studies had generally assessed the contribution of $TFS_{neural}$ and $ENV_{neural}$ cues to speech intelligibility by removing either $TFS_{acoust}$ or $ENV_{acoust}$ within each frequency band. However, this all-or-none approach may preclude a systematic assessment of the contribution of $TFS_{neural}$ or $ENV_{neural}$ on speech intelligibility. We used a phase vocoder that allows for a more gradual degradation in $TFS_{acoust}$ within each frequency band. From the physiological perspective, our approach provides a link

between the physical sound and its representation in the AN, and a link between the peripheral sensory information and central processing for the purpose of speech perception. There are several ways to quantify $TFS_{neural}$ and $ENV_{neural}$ information in the AN. The overall strength of $TFS_{neural}$ and $ENV_{neural}$ coding in individual fibers can be quantified with synchrony-based indices (Young and Sachs, 1979; Johnson, 1980; Joris and Yin, 1992), whereas correlational or coincidence-based approaches can be used to evaluate across-fiber temporal coding (Shamma, 1985; Deng and Geisler, 1987; Carney et al., 2002). However, neither of these approaches allows for the direct assessment of the effects of degradations of $TFS_{acoust}$ or $ENV_{acoust}$ cues, because they quantify the strength of the entire temporal responses. For example, if listeners are presented with the speech signals containing $TFS_{acoust}$ only (Lorenzi et al., 2006), the entire $ENV_{neural}$ response is a combination of the independent $ENV_{neural}$ from the noise carrier and the speech-related $ENV_{neural}$ responses recovered from the $TFS_{acoust}$. Cross-correlation-based metrics allows for the direct comparison of the temporal responses to original and vocoded speech, and thus provides a direct assessment of the effects of vocoding on the fidelity of speech-related cues.

Using the SCC analyses with the simulated AN responses, we showed a systematic degradation of the $TFS_{neural}$ and $ENV_{neural}$ information as a function of NF and SNR. This result is consistent with recent modeling work suggesting that jittering phase cues of the acoustic stimuli degrade the encoding of both $TFS_{neural}$ and $ENV_{neural}$ information (Heinz and Swaminathan, 2009; Shamma and Lorenzi, 2013). This dynamic interaction between the acoustic phase and the $TFS_{neural}$ and $ENV_{neural}$ information was used by the central neural-observer model. For example, at NF = 1, the original phase information was completely destroyed by the vocoder, and hence, $\rho_{TFS}$ was near 0 (Figs. 4, 5). Thus, this condition may force listeners to use $ENV_{neural}$ information only. This is consistent with the fact that the central neural-observer model was unable to predict sentence identification with the $\rho_{ENV}$ weighting coefficient of 0 (Fig. 6, n/a) at NF = 1.

The central neural-observer model predicts lower SRTs when $\rho_{TFS}$ was weighted more, demonstrating that the model prediction of speech-identification performance in noise clearly benefits from $TFS_{neural}$ cues. The extent to which $TFS_{neural}$ cues contribute to speech identification could differ for different types of maskers. For steady noise, the best fit between the SRTs observed from humans and predicted from the model was found with the $\rho_{ENV}$ weighting coefficient of 1, while for modulated noise, the $\rho_{ENV}$ weighting coefficient of 0.75 showed the best fit (Fig. 6). Furthermore, speech-masking release observed from humans was best explained when the $\rho_{ENV}$ weighting coefficient of 0.5 was used for the prediction of SRTs in modulated noise. Together, the sole use of either $ENV_{neural}$ or $TFS_{neural}$ cues does not account for human performance in all masking conditions. It may be possible that $ENV_{neural}$ and $TFS_{neural}$ cues are used optimally by the central auditory system, depending on the acoustic environment where a listener is situated.

### Limitations of the current modeling approach

The model data demonstrate that it is possible to account for the effect of phase and masker type on speech identification by listeners using relatively simple decision rules that involve the use of $ENV_{neural}$ and $TFS_{neural}$ cues. Although the central model generally outperformed humans, particularly for steady noise, the fact that the model prediction replicated the pattern of human data (effects of NF and masker type) is encouraging because these models could be used in future studies investigating the effects of hearing loss on using such $ENV_{neural}/TFS_{neural}$ cues or assessing the possible benefits of novel signal processing for hearing prostheses. However, it is important to acknowledge that the central model used in the present study was built on several assumptions. For example, only within-channel temporal cues were considered in the model. Also, humans were tested in open set, meaning that they do not have any exemplar sentence options available for them, but the central model was provided exemplar sentences. One can implement a closed-set speech-identification test both for humans and model simulations to minimize any procedural bias. Human listeners are able to use the top-down process for speech identification in degraded listening conditions (Warren, 1970), but the neural-observer model does not take into account such top-down processing.

The AN model used in the present study is stochastic in nature. Thus, this model takes into account the effect of internal noise at the level of the AN (Javel and Viemeister, 2000), but the model does not take into account internal noise that is generated more centrally (Vogels et al., 1989). Psychoacoustic studies on spectral and temporal auditory perception showed that human subjects (e.g., normal-hearing vs hearing-impaired subjects, or young vs older subjects) may differ strongly in terms of the central "processing efficiency," which is their ability to use optimally the information encoded peripherally (Hall and Grose, 1994; Vinay and Moore, 2007). The model framework presented herein can readily be expanded to include the effects of reduced processing efficiency, which could potentially exist in listeners with various forms of hearing loss. For example, Lopez-Poveda and Barrios (2013) predicted the effects of the deafferentation process on temporal coding of AN fibers and how it could subsequently affect speech identification for hearing-impaired listeners. The work of Lopez-Poveda and Barrios (2013) and the present modeling study suggest that the reduction of speech-masking release that is typically associated with sensorineural hearing loss (Lorenzi et al., 2006) might result (at least in part) from the deafferentation process and the subsequent degradation of TFS coding. Altogether, the current approach could help better link the sensory information processing (i.e., from the acoustic to the peripheral neural representation) with perceptual outcomes (i.e., following processing of peripheral information into central neural information).

In the latest versions of the AN model used in this paper, simulations of the strength of phase locking (Zilany et al., 2009) and the discharge rates at saturation for higher CFs (Zilany et al., 2014) were improved. SRTs were predicted using four CFs in the current study, but it might be more realistic if more CFs were included in the advanced AN model simulations. However, even with the most recent model with more CFs, properties of $\rho_{ENV}$ and $\rho_{TFS}$ remained quite similar as a function of NF, masker type, and SNR (data not shown). Therefore, it is unlikely that using the most recent AN model with more CFs would change the conclusions from the present study.

### Implications for hearing devices and audio coding

One of the critical barriers to improving the performance of hearing prostheses is the large variability in patient outcomes, which makes it challenging or even impossible to predict hearing-aid or cochlear-implant outcomes. Patient outcome variability occurs because speech perception involves dynamic interactions between the acoustic signals (provided by hearing aids) or electric signals (provided by cochlear implants) and different biological conditions in the ears that received hearing prostheses. Hearing device signal processing has to be customized to an individual

patient for the best outcomes; however, individual variability in outcomes poses a challenge. The approach presented in this paper suggests an objective and innovative way for such an optimization process. For example, the peripheral model used here was set up with the functionality of inner and outer hair cells set to normal hearing (i.e., $C_{IHC}$ and $C_{OHC} = 1.0$). These model parameters could be modified to simulate a specific hearing-loss configuration of an individual patient with hearing loss. With simulated neural responses from the customized AN model along with the prediction of speech identification by the central model, extensive evaluation of signal processing would be readily possible.

This paper also suggests that understanding the encoding and decoding of neural responses for acoustic speech signals may be crucial for the development of advanced automatic speech-recognition algorithms. Today, automatic speech recognition is widely used in desktop or tablet computers, mobile phones, home appliances, and cars. However, performance by automatic speech-recognition systems is far worse than humans, particularly in background noise (Benzeghiba et al., 2007). The performance of automatic speech-recognition systems could improve if a computational AN model were implemented as front-end processing. Such efforts have already been undertaken (Jürgens et al., 2013) and show a promising path for audio coding.

## References


Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C, Rose R, Tyagi V, Wellekens C (2007) Automatic speech recognition and speech variability: a review. Speech Commun 49:763–786. CrossRef

Bruce IC, Sachs MB, Young ED (2003) An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. J Acoust Soc Am 113:369–388. CrossRef Medline

Carney LH (1993) A model for the responses of low-frequency auditory-nerve fibers in cat. J Acoust Soc Am 93:401–417. CrossRef Medline

Carney LH, Heinz MG, Evilsizer ME, Gilkey RH, Colburn HS (2002) Auditory phase opponency: a temporal model for masked detection at low frequencies. Acta Acust Acust 88:334–346.

Chintanpalli A, Heinz MG (2013) The use of confusion patterns to evaluate the neural basis for concurrent vowel identification. J Acoust Soc Am 134:2988–3000. CrossRef Medline

Deng L, Geisler CD (1987) Responses of auditory-nerve fibers to nasal consonant-vowel syllables. J Acoust Soc Am 82:1977–1988. CrossRef Medline

Ghitza O (2001) On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J Acoust Soc Am 110:1628–1640. CrossRef Medline

Glasberg BR, Moore BC (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138. CrossRef Medline

Gnansia D, Péan V, Meyer B, Lorenzi C (2009) Effects of spectral smearing and temporal fine structure degradation on speech masking release. J Acoust Soc Am 125:4023–4033. CrossRef Medline

Hall JW 3rd, Grose JH (1994) Development of temporal resolution in children as measured by the temporal modulation transfer function. J Acoust Soc Am 96:150–154. CrossRef Medline

Heinz MG, Swaminathan J (2009) Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. J Assoc Res Otolaryngol 10:407–423. CrossRef Medline

Heinz MG, Colburn HS, Carney LH (2001) Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. Neural Comput 13:2273–2316. CrossRef Medline

Hilbert D (1912) Grundzüge einer Allgemeinen Theorie der Linearen Integralgleichungen. Leipzig: B.G. Teubner.

Hopkins K, Moore BC (2009) The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. J Acoust Soc Am 125:442–446. CrossRef Medline

Javel E, Viemeister NF (2000) Stochastic properties of cat auditory nerve responses to electric and acoustic stimuli and application to intensity discrimination. J Acoust Soc Am 107:908–921. CrossRef Medline

Johnson DH (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. J Acoust Soc Am 68:1115–1122. CrossRef Medline

Joris PX (2003) Interaural time sensitivity dominated by cochlea-induced envelope patterns. J Neurosci 23:6345–6350. Medline

Joris PX, Yin TC (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. J Acoust Soc Am 91:215–232. CrossRef Medline

Jürgens T, Brand T, Clark NR, Meddis R, Brown GJ (2013) The robustness of speech representations obtained from simulated auditory nerve fibers under different noise conditions. J Acoust Soc Am 134:EL282–EL288. CrossRef Medline

Lopez-Poveda EA, Barrios P (2013) Perception of stochastically undersampled sound waveforms: a model of auditory deafferentation. Front Neurosci 7:124. CrossRef Medline

Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BC (2006) Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc Natl Acad Sci U S A 103:18866–18869. CrossRef Medline

Louage DH, van der Heijden M, Joris PX (2004) Temporal properties of responses to broadband noise in the auditory nerve. J Neurophysiol 91:2051–2065. CrossRef Medline

Miller RL, Schilling JR, Franck KR, Young ED (1997) Effects of acoustic trauma on the representation of the vowel /ɛ/ in cat auditory nerve fibers. J Acoust Soc Am 101:3602–3616. CrossRef Medline

Plack CJ, Oxenham AJ (2005) The psychophysics of pitch. In: Pitch: neural coding and perception. (Plack CJ, Oxenham AJ, Fay RR, Popper AN, eds). pp 7–55. New York: Springer.

Rothauser EH, Chapman WD, Guttman N, Nordby KS, Silbiger HR, Urbanek GE, Weinstock M (1969) I.E.E.E. recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust 17:227–246.

Shamma SA (1985) Speech processing in the auditory system. I: the representation of speech sounds in the responses of the auditory nerve. J Acoust Soc Am 78:1612–1621. CrossRef Medline

Shamma S, Lorenzi C (2013) On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. J Acoust Soc Am 133:2818–2833. CrossRef Medline

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304. CrossRef Medline

Swaminathan J, Heinz MG (2012) Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. J Neurosci 32:1747–1756. CrossRef Medline

Tan Q, Carney LH (2003) A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide. J Acoust Soc Am 114:2007–2020. CrossRef Medline

Vinay S, Moore BC (2007) Ten(HL)-test results and psychophysical tuning curves for subjects with auditory neuropathy. Int J Audiol 46:39–46. CrossRef Medline

Vogels R, Spileers W, Orban GA (1989) The response variability of striate cortical neurons in the behaving monkey. Exp Brain Res 77:432–436. CrossRef Medline

Warren RM (1970) Perceptual restoration of missing speech sounds. Science 167:392–393. CrossRef Medline

Young ED, Sachs MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J Acoust Soc Am 66:1381–1403. CrossRef Medline

Zhang X, Heinz MG, Bruce IC, Carney LH (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. J Acoust Soc Am 109:648–670. CrossRef Medline

Zilany MS, Bruce IC (2006) Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J Acoust Soc Am 120:1446–1466. CrossRef Medline

Zilany MS, Bruce IC (2007) Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: model predictions of responses in cats. J Acoust Soc Am 122:402–417. CrossRef Medline

Zilany MS, Bruce IC, Nelson PC, Carney LH (2009) A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. J Acoust Soc Am 126:2390–2412. CrossRef Medline

Zilany MS, Bruce IC, Carney LH (2014) Updated parameters and expanded simulation options for a model of the auditory periphery. J Acoust Soc Am 135:283–286. CrossRef Medline