Systems/Circuits

Speech Categorization Reveals the Role of Early-Stage Temporal-Coherence Processing in Auditory Scene Analysis

Vibha Viswanathan,[1] Barbara G. Shinn-Cunningham,[2] and Michael G. Heinz[1,3]
[1]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana 47907, [2]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, and [3]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana 47907

Temporal coherence of sound fluctuations across spectral channels is thought to aid auditory grouping and scene segregation. Although prior studies on the neural bases of temporal-coherence processing focused mostly on cortical contributions, neurophysiological evidence suggests that temporal-coherence-based scene analysis may start as early as the cochlear nucleus (i.e., the first auditory region supporting cross-channel processing over a wide frequency range). Accordingly, we hypothesized that aspects of temporal-coherence processing that could be realized in early auditory areas may shape speech understanding in noise. We then explored whether physiologically plausible computational models could account for results from a behavioral experiment that measured consonant categorization in different masking conditions. We tested whether within-channel masking of target-speech modulations predicted consonant confusions across the different conditions and whether predictions were improved by adding across-channel temporal-coherence processing mirroring the computations known to exist in the cochlear nucleus. Consonant confusions provide a rich characterization of error patterns in speech categorization, and are thus crucial for rigorously testing models of speech perception; however, to the best of our knowledge, they have not been used in prior studies of scene analysis. We find that within-channel modulation masking can reasonably account for category confusions, but that it fails when temporal fine structure cues are unavailable. However, the addition of across-channel temporal-coherence processing significantly improves confusion predictions across all tested conditions. Our results suggest that temporal-coherence processing strongly shapes speech understanding in noise and that physiological computations that exist early along the auditory pathway may contribute to this process.

Key words: **cochlear nucleus; comodulation masking release; computational modeling; consonant confusions; cross-channel processing; wideband inhibition**

Significance Statement

Temporal coherence of sound fluctuations across distinct frequency channels is thought to be important for auditory scene analysis. Prior studies on the neural bases of temporal-coherence processing focused mostly on cortical contributions, and it was unknown whether speech understanding in noise may be shaped by across-channel processing that exists in earlier auditory areas. Using physiologically plausible computational modeling to predict consonant confusions across different listening conditions, we find that across-channel temporal coherence contributes significantly to scene analysis and speech perception and that such processing may arise in the auditory pathway as early as the brainstem. By virtue of providing a richer characterization of error patterns not obtainable with just intelligibility scores, consonant confusions yield unique insight into scene analysis mechanisms.

Introduction

An accumulating body of evidence suggests that temporal-coherence processing is important for auditory scene analysis (Elhilali et al., 2009). Indeed, a rich psychophysical literature on grouping (Darwin, 1997), comodulation masking release (CMR; Schooneveldt and Moore, 1987), cross-channel modulation interference (Apoux and Bacon, 2008), and pitch-based masking release (Oxenham and Simonson, 2009) supports the theory that temporally coherent sound modulations can bind together sound elements across distinct spectral channels to form a perceptual object, which can help perceptually

Received Aug. 6, 2021; revised Oct. 18, 2021; accepted Oct. 26, 2021.
Author contributions: V.V., B.G.S.-C., and M.G.H. designed research; V.V. performed research; V.V. analyzed data; V.V., B.G.S.-C., and M.G.H. wrote the paper.
This research was supported by National Institutes of Health Grants F31DC017381 (V.V.) and R01DC009838 (M.G.H.) and Office of Naval Research Grant ONR N00014-20-12709 (B.G.S.-C.). We thank Hari Bharadwaj for access to online psychoacoustics infrastructure. We also thank Andrew Sivaprakasam, François Deloche, Hari Bharadwaj, and Ravinderjit Singh for feedback on an earlier version of this paper.
The authors declare no competing financial interests.
Correspondence should be addressed to Vibha Viswanathan at viswanav@purdue.edu.
https://doi.org/10.1523/JNEUROSCI.1610-21.2021

Table 1. Rationale for the different stimulus conditions included in this study

No.	Stimulus condition	Rationale for inclusion in study
1	SiQuiet	Used as a control condition
2	SiSSN at −8 dB SNR	Widely used in the literature; used for calibration of prediction model
3	SiB at −8 dB SNR	Simulates ecologically relevant cocktail-party listening
4	SiDCmod at −18 dB SNR	To obtain a different modulation masking profile from stationary noise (which contains relatively more high-frequency modulation energy) and babble (which contains relatively more low-frequency modulation power; Viswanathan et al., 2021a)
5	SiB at 0 dB SNR subjected to 64-channel envelope vocoding (Vocoded SiB)	Used to compare performance across models that consider TFS and those that do not (as TFS can influence scene analysis and can convey consonant voicing information in noise; Viswanathan et al., 2021a,b)

The different listening conditions were chosen to span a range of modulation masking spectral profiles and TFS information, which allows for theories of scene analysis based on within-channel modulation masking and across-channel temporal coherence to be tested in a rigorous manner. Collectively, these conditions represent a diversity of scene acoustics, including important examples in our environment and clinical applications. The SNR levels were chosen to give approximately equal overall intelligibility across SiSSN, SiB, SiDCmod, and Vocoded SiB using a behavioral pilot study with three subjects who did not participate in the online consonant identification experiment. This was done to obtain roughly equal variance in the consonant confusion estimates for these conditions, which allows us to fairly compare confusion patterns across them. Equalizing intelligibility also maximizes the statistical power for detecting differences in the pattern of confusions. The overall intelligibility in each of these conditions was ~60%, which yielded a sufficient number of confusions for analysis.

segregate different sources in an acoustic mixture. This theory may help explain how we perform speech separation in a multisource environment (Krishnan et al., 2014), as speech naturally has common temporal fluctuations across channels, particularly in the syllabic (0–5 Hz), phonemic (5–64 Hz), and pitch (64–300 Hz) ranges (Crouzet and Ainsworth, 2001; Swaminathan and Heinz, 2011).

Object binding and scene segregation are perceptually defined phenomena, whose neural correlates are yet to be definitively established. These phenomena may in general be supported by a cascade of mechanisms throughout the auditory pathway (Pressnitzer et al., 2008; Shinn-Cunningham, 2020; Mishra et al., 2021). Prior studies on the neural bases of temporal-coherence processing mostly focused on cortical contributions (Elhilali et al., 2009; Teki et al., 2013; O'Sullivan et al., 2015). However, single-unit measurements and computational modeling of across-channel CMR effects suggest that temporal-coherence-based scene analysis may start early in the auditory pathway; for instance, the cochlear nucleus has the physiological mechanisms (e.g., wideband inhibition) needed to support such analysis (Pressnitzer et al., 2001; Meddis et al., 2002). Moreover, attention, which operates on segregated auditory objects (Shinn-Cunningham, 2008), affects responses in the early auditory cortex (Hillyard et al., 1973). Given this, binding and scene segregation likely start even earlier, such as brainstem, and accumulate along the auditory pathway. However, no prior studies have directly tested the hypothesis that speech understanding in noise may be shaped by aspects of temporal-coherence processing that exist in early auditory areas.

Previous studies of temporal-coherence processing mostly used nonspeech stimuli (Elhilali et al., 2009; Teki et al., 2013; O'Sullivan et al., 2015). Moreover, a parallel literature on modeling speech-intelligibility mechanisms typically focused on overall intelligibility to test predictions of performance (Jørgensen et al., 2013; Relaño-Iborra et al., 2016). A detailed characterization of error patterns in speech categorization—crucial to rigorously examine any theory of speech perception—has not been previously used in studies of scene analysis. In contrast, confusion patterns in speech categorization, such as consonant confusion matrices (Miller and Nicely, 1955), have been widely used in the speech acoustics and cue-weighting literatures and can provide deeper insight into underlying mechanisms if used to test theories of scene analysis.

To address these gaps, we used a combination of online consonant identification experiments and computational modeling of temporal-coherence processing that is physiologically plausible in the cochlear nucleus (Pressnitzer et al., 2001), the first auditory area where cross-channel processing over a wide frequency range is supported. We asked whether the masking of target-speech envelopes by distracting masker modulations (i.e., modulation masking; Bacon and Grantham, 1989; Stone and Moore, 2014) within individual frequency channels (as implemented in current speech-intelligibility models; Jørgensen et al., 2013; Relaño-Iborra et al., 2016) is sufficient to predict consonant categorization, or if across-channel temporal-coherence processing improves predictions by accounting for interference from masker elements that are temporally coherent with target elements but in different frequency channels. Crucially, instead of just trying to predict perceptual intelligibility measurements from model outputs, we predicted consonant confusion patterns in various listening conditions. Considering the error patterns in consonant categorization provided a richer characterization of the processes engaged during speech perception compared to looking only at percent correct scores. Our combined use of consonant confusions and physiologically plausible computational modeling provides independent evidence for the role of temporal-coherence processing in scene analysis and speech perception. Moreover, it suggests that this processing may start earlier in the auditory pathway than previously thought.

Table 2. Phonetic features of the 20 English consonants used in this study

Consonant	Voicing	MOA	POA
/b/	Voiced	Stop	Bilabial
/tʃ/	Unvoiced	Affricative	Palatal
/d/	Voiced	Stop	Alveolar
/ð/	Voiced	Fricative	Dental
/f/	Unvoiced	Fricative	Labiodental
/g/	Voiced	Stop	Velar
/dʒ/	Voiced	Affricative	Palatal
/k/	Unvoiced	Stop	Velar
/l/	Voiced	Liquid	Alveolar
/m/	Voiced	Nasal	Bilabial
/n/	Voiced	Nasal	Alveolar
/p/	Unvoiced	Stop	Bilabial
/r/	Voiced	Liquid	Palatal
/s/	Unvoiced	Fricative	Alveolar
/ʃ/	Unvoiced	Fricative	Palatal
/t/	Unvoiced	Stop	Alveolar
/θ/	Unvoiced	Fricative	Dental
/v/	Voiced	Fricative	Labiodental
/z/	Voiced	Fricative	Alveolar
/ʒ/	Voiced	Fricative	Palatal

Materials and Methods

Stimulus generation. The stimuli used in the present study draw from and expand on the materials and methods previously described in Viswanathan et al. (2021b). Twenty consonants from the Speech Test Video (STeVi) corpus (Sensimetrics) were used. The consonants were

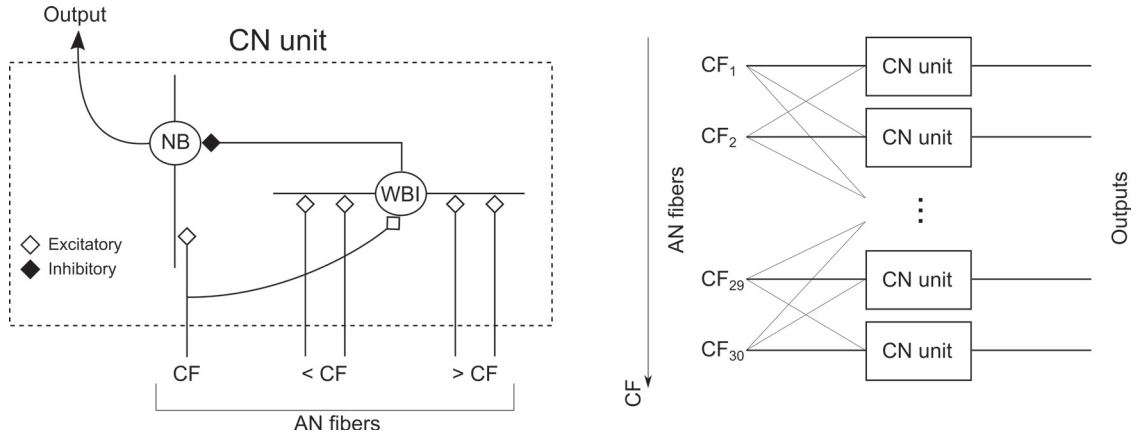

Figure 1. CMR circuit based on wideband inhibition in the cochlear nucleus. This physiologically plausible circuit was proposed by Pressnitzer et al. (2001) to model CMR effects seen in the cochlear nucleus (CN). CN units at different CFs form the building blocks of this circuit. Each CN unit consists of a narrowband cell (NB) that receives narrow on-CF excitatory input from the auditory nerve (AN) and inhibitory input from a wideband inhibitor (WBI). The WBI in turn receives excitatory inputs from AN fibers tuned to CFs spanning 2 octaves below to 1 octave above the CF of the NB that it inhibits. The time constants for the excitatory and inhibitory synapses are 5 ms and 1 ms, respectively. The WBI input to the NB is delayed with respect to the AN input by 2 ms. Note that our model simulations were rate based; that is, they used AN PSTHs rather than spikes. Thus, all outputs were half-wave rectified (i.e., firing rates were positive at every stage). All synaptic filters were initially normalized to have unit gain, then the gain of the inhibitory input was allowed to vary parametrically to implement different excitation-to-inhibition (EI) ratios between 3:1 and 1:1. The EI ratio was adjusted to obtain the best consonant confusion prediction accuracy for SiSSN (i.e., the calibration condition), and the optimal ratio for the calibration condition was found to be 1.75:1.

/b/, /tʃ/, /d/, /ð/, /f/, /g/, /dʒ/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /ʃ/, /t/, /θ/, /v/, /z/, and /ʒ/. The consonants were presented in consonant-vowel (CV) context, where the vowel was always /a/. Each consonant was spoken by two female and two male talkers (to reflect real-life talker variability). The CV utterances were embedded in the following carrier phrase: "You will mark /CV/ please" (i.e., in natural running speech).

Stimuli were created for five experimental conditions: (1) Speech in quiet (SiQuiet): Speech in quiet was used as a control condition. (2) Speech in speech-shaped stationary noise (SiSSN): Speech was added to stationary Gaussian noise at −8 dB signal-to-noise ratio (SNR). The long-term spectra of the target speech (including the carrier phrase) and that of stationary noise were adjusted to match the average (across instances) long-term spectrum of the four-talker babble. A different realization of stationary noise was used for each SiSSN stimulus. (3) Speech in babble (SiB): Speech was added to four-talker babble at −8 dB SNR. The long-term spectrum of the target speech (including the carrier phrase) was adjusted to match the average (across instances) long-term spectrum of the four-talker babble. Each SiB stimulus was created by randomly selecting a babble sample from a list of 72 different four-talker babble maskers obtained from the QuickSIN corpus (Killion et al., 2004). (4) Speech in a masker with only DC modulations (SiDCmod) (Stone et al., 2012): In line with the procedure described in Stone et al. (2012), the target speech was filtered into 28 channels between 100 and 7800 Hz, and a sinusoidal masker centered on each channel was added to the channel signal at −18 dB SNR. To minimize peripheral interactions between maskers, odd-numbered channels were presented to one ear and even to the other; this procedure effectively yields an unmodulated masker (i.e., a masker with a modulation spectrum containing only a DC component). Thus, the SiDCmod condition presented stimuli that were dichotic, unlike the other conditions, which presented diotic stimuli. The long-term spectra of the target speech (including the carrier phrase) and that of the masker were adjusted to match the average (across instances) long-term spectrum of the four-talker babble. (5) Vocoded speech in babble (Vocoded SiB): SiB at 0 dB SNR was subjected to 64-channel envelope vocoding. A randomly selected babble sample was used for each Vocoded SiB stimulus, similar to what was done for SiB. In accordance with prior work (Qin and Oxenham, 2003; Viswanathan et al., 2021b), our vocoding procedure retained the cochlear-level envelope in each of 64 contiguous frequency channels with center frequencies equally spaced on an equivalent rectangular bandwidth (ERB)-number scale (Glasberg and Moore, 1990) between 80 and 6000 Hz; however, the stimulus temporal fine structure (TFS) in each channel was replaced with a noise carrier. The envelope in each channel was extracted by half-wave rectification of the frequency content in that channel followed by low-pass filtering with a cutoff frequency of 300 Hz, or half of the channel bandwidth, whichever was lower. The envelope in each channel was then used to modulate a random Gaussian white noise carrier; the result was band-pass filtered within the channel bandwidth and scaled to match the level of the original signal. We verified that the vocoding procedure did not significantly change envelopes at the cochlear level, as described in Viswanathan et al. (2021b). Table 1 describes the rationale behind including these different stimulus conditions in our study.

The stimulus used for online volume adjustment was running speech mixed with four-talker babble. The speech and babble samples were obtained from the QuickSIN corpus (Killion et al., 2004); these were repeated over time to obtain a ~20 s total stimulus duration to give subjects sufficient time to adjust their computer volume with the instructions described in Experimental design. The root mean square value of this stimulus corresponded to 75% of the dB difference between the softest and loudest stimuli in the consonant identification experiment, which ensured that no stimulus was too loud for subjects once they had adjusted their computer volume to a comfortable level.

Participants. Full details of participant recruitment and screening are provided in Viswanathan et al. (2021b) and are only briefly reviewed here. Anonymous subjects were recruited for online data collection using Prolific.co. A three-part subject-screening protocol developed and validated by Mok et al. (2021) was used to restrict the subject pool. This protocol included a survey on age, native-speaker status, presence of persistent tinnitus, and history of hearing and neurologic diagnoses, followed by headphone/earphone checks and a speech-in-babble-based hearing screening. Subjects who passed this screening protocol were invited to participate in the consonant identification study, and when they returned, headphone/earphone checks were performed again. Only subjects who satisfied the following criteria passed the screening protocol: (1) 18–55 years old; (2) self-reported no hearing loss, neurologic disorders, or persistent tinnitus; (3) born and residing in the United States/Canada and a native speaker of North American English; (4) experienced Prolific subject; and (5) passed the headphone/earphone checks and speech-in-babble-based hearing screening (Mok et al., 2021). Subjects provided informed consent in accordance with remote testing protocols approved by the Purdue University Institutional Review Board (IRB).

Experimental design. The online consonant identification experiment was previously described in Viswanathan et al. (2021b). Subjects performed the experiment using their personal computers

and headphones/earphones. Our online infrastructure included checks to prevent the use of mobile devices. The experiment consisted of the following parts: (1) headphone/earphone checks, (2) demonstration (Demo), and (3) Test. Each of these three parts had a volume-adjustment task at the beginning. In this task, subjects were asked to make sure that they were in a quiet room and wearing wired (not wireless) headphones or earphones. They were instructed not to use desktop/laptop speakers. Headphone/earphone use was checked using the procedures in Mok et al. (2021). They were then asked to set their computer volume to 10–20% of the full volume, after which they were played a speech-in-babble stimulus and asked to adjust their volume up to a comfortable but not too loud level. Once subjects had adjusted their computer volume, they were instructed not to adjust the volume during the experiment, as that could lead to sounds being too loud or soft.

The Demo stage consisted of a short training task designed to familiarize subjects with how each consonant sounds and with the consonant identification paradigm. Subjects were instructed that in each trial they would hear a voice say, "You will mark *something* please." They were told that they would be given a set of options for *something* at the end of the trial, and that they should click on the corresponding option. After subjects had heard all consonants sequentially (i.e., the same order as the response choices) in quiet, they were tasked with identifying consonants presented in random order and spanning the same set of listening conditions as the Test stage. Subjects were instructed to ignore any background noise and only listen to the particular voice saying, "You will mark *something* please." To ensure that all subjects understood and were able to perform the task, only those subjects who scored $\geq$85% in the Demo's SiQuiet control condition were selected for the Test stage.

Subjects were given instructions in the Test stage similar to those in the Demo but told to expect trials with background noise from the beginning. The Test stage presented, in random order, the 20 consonants (with one stimulus repetition per consonant) across all four talkers and all five experimental conditions. In both Demo and Test, the masking noise, when present, started 1 s before the target speech and continued for the entire duration of the trial. This was done to cue the subjects' attention to the stimulus before the target sentence was played. In both the Demo and Test parts, subjects received feedback after every trial on whether their response was correct to promote engagement with the task. However, subjects were not told what consonant was presented to avoid overtraining to the acoustics of how each consonant sounded across the different conditions; the only exception to this rule was in the first subpart of the Demo, where subjects heard all consonants in quiet in sequential order.

Separate studies were posted on Prolific.co for the different talkers. When a subject performed a particular study, they would be presented with the speech stimuli for one specific talker consistently over all trials. Thus, each subject was not just trained with one talker but also tested with that same talker to avoid training-testing disparities. To obtain results that are generalizable, we used 50 subjects per talker (subject overlap between talkers was not controlled); with four talkers, this yielded 200 subject-talker pairs or data samples. Within each talker and condition, all subjects performed the task with the same stimuli. Moreover, all condition effect contrasts were computed on a within-subject basis and averaged across subjects.

Data preprocessing. Only samples with intelligibility scores $\geq$85% for the SiQuiet control condition in the Test stage were included in results reported here. All conditions for the remaining samples were excluded from further analyses as a data quality control measure. This yielded a final $N = 191$ samples.

Quantifying confusion matrices from perceptual measurements. The 20 English consonants used in this study were assigned the phonetic features described in Table 2. The identification data collected in the Test stage were used to construct consonant confusion matrices (pooled over samples) for the different conditions; these matrices in turn were used to construct voicing, place of articulation (POA), and manner of articulation (MOA) confusion matrices by pooling over all consonants.

Given that all psychophysical data were collected online, we performed data quality checks; the analyses performed and the results are described in detail in Mok et al. (2021) and Viswanathan et al. (2021b), and are only briefly presented here. We compared consonant confusions for SiSSN, a commonly used condition in the literature, with previous lab-based findings. Phatak and Allen (2007) found that for a given overall intelligibility, recognition scores vary across consonants. They identified three groups of consonants, C1, C2, and C3 with low, high, and intermediate recognition scores, respectively, in speech-shaped noise. The SiSSN data that we collected online closely replicated that key trend for the groups they identified. Moreover, based on a graphical analysis of confusion patterns in speech-shaped noise, Phatak and Allen (2007) identified perceptual clusters (i.e., sets where one consonant is confused most with another in the same set). In the current study too, we identified perceptual clusters for SiSSN by subjecting the consonant confusion matrix to a hierarchical clustering analysis (Ward, 1963); our results closely replicated the lab-based clustering results of Phatak and Allen (2007). As previous lab-based results were not readily available for the

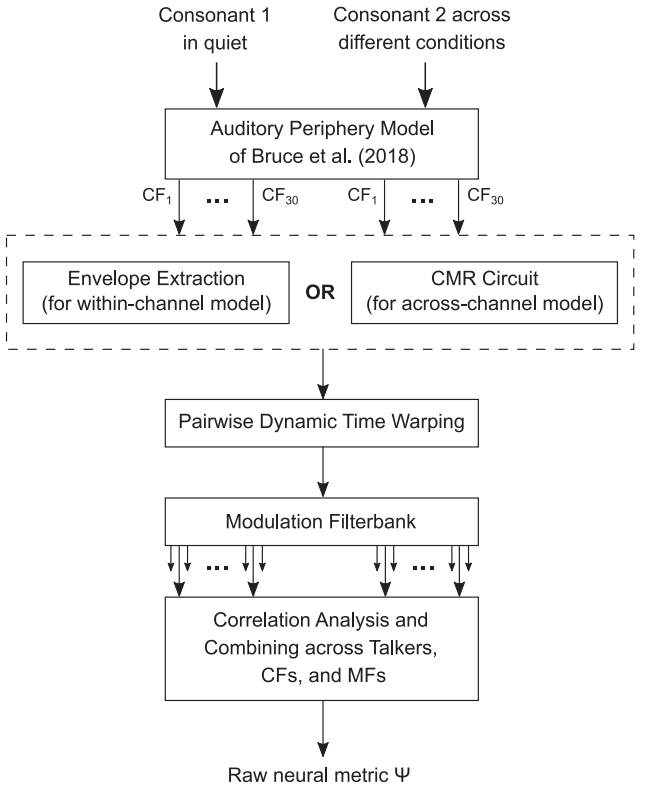

Figure 2. Schematic of the within- and across-channel scene analysis models. The speech stimuli were input into the Bruce et al. (2018) model, which simulated a normal auditory periphery with 30 cochlear filters having CFs equally spaced on an ERB-number scale (Glasberg and Moore, 1990) between 125 Hz and 8 kHz. PSTHs from the periphery model were processed to retain only the time segments when the target consonants were presented. For the within-channel model, these results were filtered within a 1 ERB bandwidth (Glasberg and Moore, 1990) to extract band-specific envelopes; however, for the across-channel model, the results were instead input into the CMR circuit model (Fig. 1). Pairwise dynamic time warping was performed to align the outputs from the previous step across time for each pair of consonants. A modulation filterbank (Ewert and Dau, 2000; Jørgensen et al., 2013) was then used to decompose the results at each CF into different MF bands. This filterbank consists of a low-pass filter with a 1 Hz cutoff in parallel with eight bandpass filters with octave spacing, a quality factor of 1, and center frequencies between 2 and 256 Hz. For each condition, talker, CF, MF, and consonant, Pearson correlation coefficients were computed between the filterbank output for that consonant in that particular condition and the output for each of all 20 consonants in quiet. Each of the individual correlations was squared to obtain the variance explained; the results were averaged across talkers, CFs, and MFs to obtain a raw neural metric ψ for each experimental condition. A separate ψ value was obtained for each condition, and every pair of consonant presented and option for consonant reported. The ψ values were normalized such that their sum across all options for consonants reported for a particular consonant presented was equal to one, which yielded a condition-specific neural consonant confusion matrix.

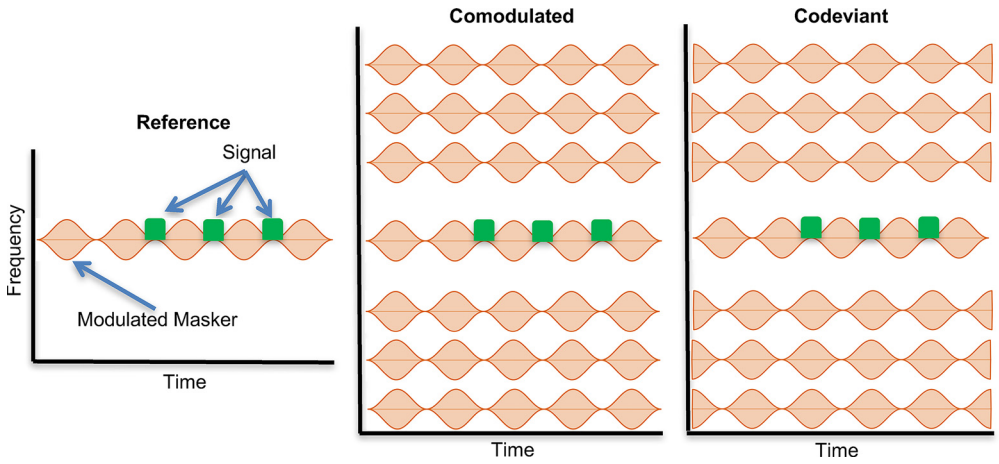

Figure 3. Stimuli used to validate the CMR circuit model. The stimuli used were from Pressnitzer et al. (2001), and consisted of a target signal in a 10 Hz 100% SAM tonal complex masker. The masker differed depending on the experimental condition. In the Reference condition, the masker was a 1.1-kHz-carrier SAM tone (referred to as the OFC). In the Comodulated and Codeviant conditions, six flanking components were presented in addition to the OFC. The flanking components were SAM tones at the same level as the OFC. The flanking components were separated from the OFC by −800, −600, −400, 400, 600, and 800 Hz, respectively. The modulation of each flanking component was in phase with the OFC modulation in the Comodulated condition, but 180° out of phase with the OFC modulation in the Codeviant condition. The target signal was a 50-ms-long 1.1 kHz tone pip that was presented in the dips of the OFC modulation during the last 0.3 s of the stimulus period (i.e., in the last 3 dips) at different values of SCR (defined as the signal maximum amplitude over the amplitude of the OFC before modulation).

remaining masking conditions in our study, we instead examined whether subjects randomly chose a different consonant from what was presented when they made an error, or if there was more structure in the data. The percent errors in our data fell outside the distributions expected from random confusions, suggesting that the error patterns have a nonrandom structure. Together, these results support the validity of our online-collected data.

We wished to test whether there are any significant differences in consonant confusion patterns across the different masking conditions, namely, SiSSN, SiB, SiDCmod, and Vocoded SiB. If so, these differences could then be predicted by computational modeling to test our hypothesis about the role of temporal-coherence-based across-channel masking of target speech by noise fluctuations. As the SiQuiet condition was intended to primarily be used as a control condition to ensure data quality (see Data preprocessing), SiQuiet data were not subjected to this analysis. To test whether confusion patterns differed across the masking conditions, we first normalized the overall intelligibility for these conditions to 60% by scaling the consonant confusion matrices such that the sum of the diagonal entries was the desired intelligibility (note that overall intelligibility was not normalized for the main modeling analyses of this study). By matching intelligibility in this manner, differences in confusion matrices across conditions could be attributed to changes in consonant categorization and category errors rather than differences in overall error counts because of one condition being inherently easier at a particular SNR. Because overall intelligibility was similar across the masking conditions to start with (see Fig. 5), small condition differences in intelligibility could be normalized without loss of statistical power. Confusion-matrix differences between the intelligibility-matched conditions were then compared with appropriate null distributions of zero differences (see Statistical analysis) to extract statistically significant differences (see Fig. 6).

Auditory periphery modeling. The auditory-nerve model of Bruce et al. (2018) was used to simulate processing by the auditory periphery. The parameters of this model were set as follows. Thirty cochlear filters with characteristic frequencies (CFs) equally spaced on an ERB-number scale (Glasberg and Moore, 1990) between 125 and 8000 Hz were used. Normal function was chosen for the outer and inner hair cells. The species was chosen to be human with the Shera et al. (2002) cochlear tuning at low sound levels; however, with suppression, the Glasberg and Moore (1990) tuning is effectively obtained for our broadband, moderate-level stimuli (Heinz et al., 2002; Oxenham and Shera, 2003). The noise type parameter for the inner-hair-cell synapse model was set to fixed fractional Gaussian noise to yield a constant spontaneous auditory-nerve firing rate. To avoid single-fiber saturation effects, the spontaneous rate of the auditory-nerve fiber was set to 10, corresponding to that of a medium-spontaneous-rate fiber. An approximate implementation of the power-law adaptation dynamics in the synapse was used. The absolute and relative refractory periods were set to 0.6 ms.

The periphery model was simulated with the same speech stimuli used in our psychophysical experiment (i.e., CV utterances that spanned 20 consonants, four talkers, and five conditions, and were embedded in a carrier phrase) as input. The level for the target speech was set to 60 dB SPL across all stimuli, as this produced sufficient (i.e., firing rate greater than spontaneous rate) model auditory-nerve responses for consonants in quiet and also did not saturate the response to the loudest stimulus. The periphery model was provided with just one audio channel input for all conditions except SiDCmod, as that was the only condition that was dichotic rather than diotic. Instead, for SiDCmod, the model was separately simulated for each of the two audio channels. Two hundred stimulus repetitions were used to derive peristimulus time histograms (PSTHs) from model auditory-nerve outputs. The model was simulated for the full duration of each stimulus, as opposed to just the time period when the target consonant was presented. A PSTH bin width of 1 ms (i.e., a sampling rate of 1 kHz) was used. This was done to capture fine-structure phase locking up to and including the typical frequency range of human pitch for voiced sounds. In the case of the SiDCmod condition, a separate PSTH was computed for each of the two dichotic audio channels.

Although the full speech stimuli (including the carrier phrase and CV utterances) were used as inputs to the periphery model, the responses to the target consonants were segmented out from the model PSTHs before being input into the scene analysis models. This segmentation had to be performed manually because the duration of the carrier phrase varied across consonants and talkers, and the start and end times corresponding to any given target consonant were unknown a priori. The time segment corresponding to when the target consonant was presented was calculated for each speech-in-quiet stimulus by visualizing speech spectrograms computed by gammatone filtering (Patterson et al., 1987) followed by Hilbert-envelope extraction (Hilbert, 1906). One hundred twenty-eight gammatone filters were used for this purpose, with center frequencies between 100 and 8000 Hz and equally spaced on an ERB-number scale (Glasberg and Moore, 1990). A fixed duration of 104.2 ms was used for each consonant segment. Segmentation accuracy was verified by listening to the segmented consonant utterances. The time segments thus derived were used to extract model auditory-nerve responses to the different target consonants across the different conditions and talkers. These responses were then used as inputs to the scene analysis models described below.

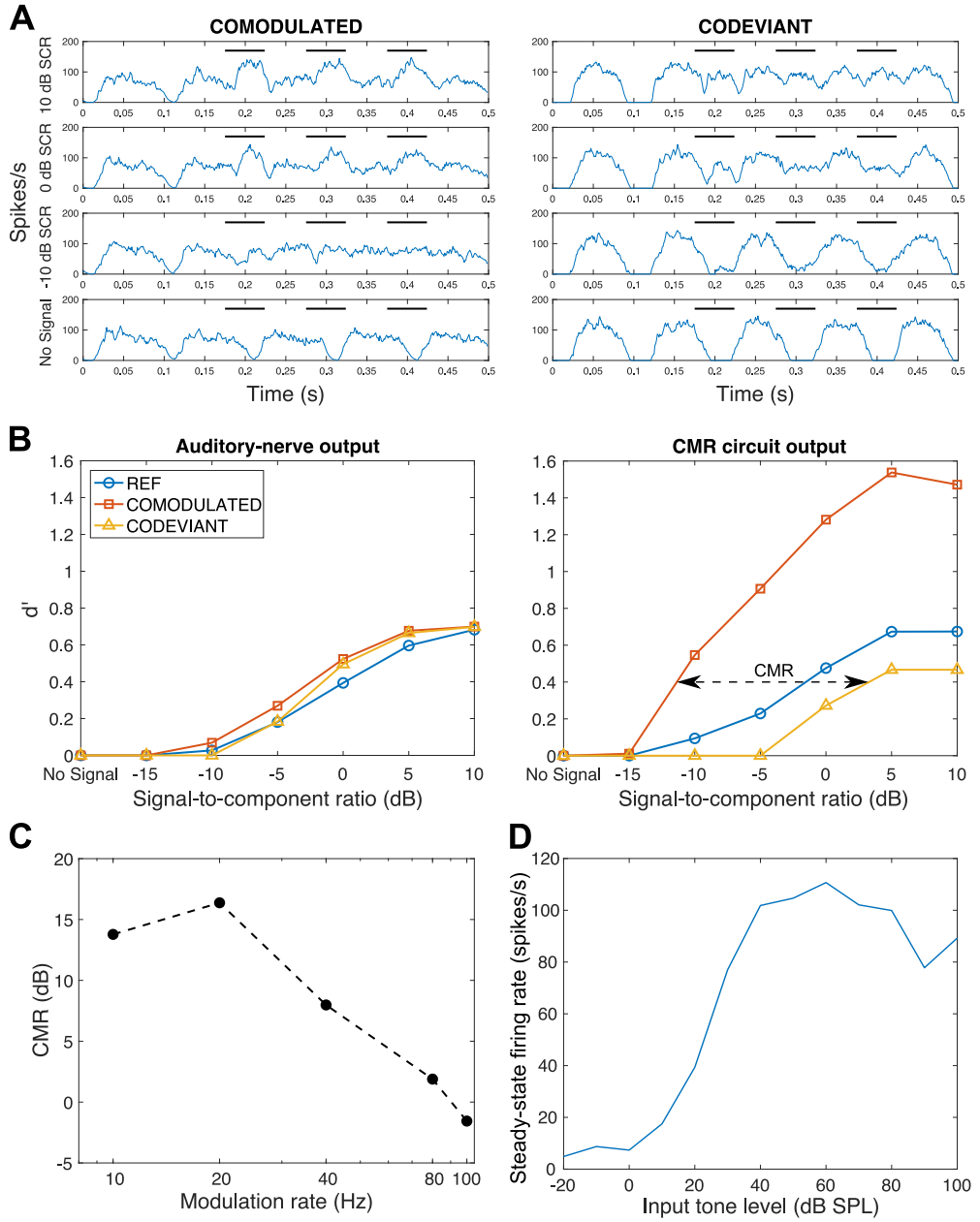

Figure 4. CMR circuit model validation. ***A***, PSTH outputs from the CMR circuit model at 1.1 kHz CF for the stimuli in Figure 3. Results are shown separately for the Comodulated and Codeviant conditions and at different SCRs. The black horizontal bars indicate the time points corresponding to when the target signal was presented. ***B***, Summary of the results from ***A*** showing the neurometric sensitivity, d', as a function of SCR for the auditory-nerve and CMR circuit model outputs (both at 1.1 kHz CF). The CMR circuit model shows a clear separation between the Comodulated and Codeviant conditions, that is, a CMR effect. This is not seen at the level of the auditory nerve. ***C***, The variation in the CMR obtained from the circuit model as a function of modulation rate. ***D***, The pure-tone rate-level function (i.e., mean steady-state firing rate versus input tone level) for the CMR circuit model.

Scene analysis modeling to predict consonant confusions. To study the contribution of across-channel temporal-coherence processing to consonant categorization, we constructed two different scene analysis models. The first is a within-channel modulation-masking-based scene analysis model inspired by Relaño-Iborra et al. (2016), and the second is a simple across-channel temporal coherence model mirroring the physiological computations that are known to exist in the cochlear nucleus (Pressnitzer et al., 2001).

In the within-channel modulation-masking-based model, the auditory-nerve PSTHs (i.e., the outputs from the periphery model; see Auditory periphery modeling) corresponding to the different consonants, conditions, and talkers were filtered within a 1 ERB bandwidth (Glasberg and Moore, 1990) to extract band-specific envelopes. Note that the envelopes extracted from auditory-nerve outputs may contain some TFS converted to envelopes via inner-hair-cell rectification (assuming envelope and TFS are defined at the output of the cochlea), but that is the processing that is naturally performed by the auditory system as well. Pairwise dynamic time warping (Rabiner, 1993) was performed to align the results for each pair of consonants across time. Dynamic time warping can help compensate for variations in speaking rate across consonants. A modulation filterbank (Ewert and Dau, 2000; Jørgensen et al., 2013) was then used to decompose the results at each CF into different modulation frequency (MF) bands. This filterbank consists of a low-pass filter with a cutoff frequency of 1 Hz in parallel with eight bandpass filters with octave spacing, a quality factor of 1, and center frequencies ranging from 2 to 256 Hz. For each condition, talker, CF, MF, and consonant, Pearson correlation coefficients were computed between the filterbank output for that consonant in that particular condition and the output for each of all 20 consonants in quiet. Each of the individual correlations was squared to obtain the variance explained; the

Table 3. Pearson correlation coefficients between within-channel model predictions and perceptual measurements

Condition	Diagonal entries		Off-diagonal entries		All entries	
	Correlation	*p*-value	Correlation	*p*-value	Correlation	*p*-value
SiSSN	83%	0.0002***	67%	10^{-8}***	87%	10^{-21}***
SiB	72%	0.0026**	64%	10^{-7}***	87%	10^{-21}***
SiDCmod	66%	0.0072**	64%	10^{-7}***	83%	10^{-17}***
Vocoded SiB	4%	0.4445	40%	0.0019**	75%	10^{-13}***

Results are listed separately for the diagonal entries of the confusion matrix (i.e., proportion correct for the different consonant phonetic categories), off-diagonal entries (i.e., true confusions), and across all entries. Note that *p*-value ranges are mapped to symbols as follows: *** indicates $0 \leq p < 0.001$, ** indicates $0.001 \leq p < 0.01$, and * indicates $0.01 \leq p < 0.05$.

Table 4. Pearson correlation coefficients between across-channel model predictions and perceptual measurements

Condition	Diagonal entries		Off-diagonal entries		All entries	
	Correlation	*p*-value	Correlation	*p*-value	Correlation	*p*-value
SiSSN	89%	10^{-5}***	81%	10^{-13}***	92%	10^{-27}***
SiB	85%	0.0001***	73%	10^{-10}***	90%	10^{-24}***
SiDCmod	88%	10^{-5}***	72%	10^{-9}***	86%	10^{-20}***
Vocoded SiB	63%	0.0103*	70%	10^{-9}***	86%	10^{-20}***

Results are listed separately for the diagonal entries of the confusion matrix (i.e., proportion correct for the different consonant phonetic categories), off-diagonal entries (i.e., true confusions), and across all entries. Note that *p*-value ranges are mapped to symbols as follows: *** indicates $0 \leq p < 0.001$, ** indicates $0.001 \leq p < 0.01$, and * indicates $0.01 \leq p < 0.05$.

results were averaged across talkers, CFs, and MFs to obtain a raw neural metric ψ for each experimental condition. A separate ψ value was obtained for each condition, and every pair of consonant presented and option for consonant reported. For the dichotic SiDCmod condition, the variance explained was separately computed for the left and right ears at each CF, then the maximum across the two ears (i.e., the better-ear contribution) was used for that CF (Zurek, 1993). Finally, for each condition, the ψ values were normalized such that their sum across all options for consonants reported for a particular consonant presented was equal to one; this procedure yielded a condition-specific neural consonant confusion matrix.

We wanted to test whether across-channel temporal-coherence processing of input fluctuations could better predict consonant categorization than a purely within-channel modulation masking model. To simulate across-channel temporal-coherence processing, we modeled a physiologically plausible wideband-inhibition-based temporal-coherence processing circuit proposed by Pressnitzer et al. (2001) to account for physiological correlates of CMR in the cochlear nucleus. A schematic of this circuit is provided in Figure 1. Note that the circuit model parameter corresponding to the excitation-to-inhibition ratio cannot be readily compared to its physiological correlate because the model is rate based and lacks important membrane conductance properties that spiking models can be endowed with. The overall across-channel scene analysis model is similar to the within-channel model, except that the envelope extraction stage of the within-channel model is replaced with the CMR circuit model in the across-channel model. Thus, the across-channel model can account for both within-channel modulation masking effects as well as across-channel temporal-coherence processing. Figure 2 shows schematics of both the within- and across-channel models.

To verify that the CMR circuit model (Fig. 1) produced physiological correlates of CMR similar to those reported by Pressnitzer et al. (2001), we used the same complex stimuli that they used (Fig. 3). The stimuli consisted of a target signal in a 100% sinusoidally amplitude-modulated (SAM) tonal complex masker. There were three experimental conditions: Reference, Comodulated, and Codeviant. In the Reference condition, the masker had just one component, a SAM tone with a carrier frequency of 1.1 kHz (to allow comparison to data from Pressnitzer et al., 2001); this masking component is also referred to as the on-frequency component (OFC). The Comodulated and Codeviant conditions presented the OFC along with six flanking components that were SAM tones at the same level as the OFC. The carrier frequency separation between the different flanking components and the OFC were −800, −600, −400, 400, 600, and 800 Hz, respectively. The flanking components were modulated in phase with the OFC in the Comodulated condition, and 180° out of phase with the OFC in the Codeviant condition. A 10 Hz modulation rate was used for all SAM tones. The target signal consisted of a 50-ms-long (i.e., half of the modulation time period) tone pip at 1.1 kHz that was presented in the dips of the OFC modulation during the last 0.3 s of the stimulus period (i.e., in the last three dips) at different values of signal-to-component ratio (SCR; defined as the signal maximum amplitude over the amplitude of the OFC before modulation). These stimuli were presented to the periphery model, and the corresponding model outputs were passed into the CMR circuit model.

The rate-level function at the output of the CMR circuit model (Fig. 4*D*) closely matches physiological data for chopper units in the ventral cochlear nucleus (Winter and Palmer, 1990) and was used to set the masker level for the CMR stimuli. The firing-rate threshold was 0 dB SPL for pure-tone inputs at CF; thus, a fixed level of 40 dB SPL (i.e., 40 dB SL) was used for the OFC. The PSTH outputs from the CMR circuit model (at 1.1 kHz CF) are shown in Figure 4*A*. The time-averaged statistics of the firing rate during the last 0.3 s of the stimulus period and in the absence of the target signal were used as the null distribution against which the neurometric sensitivity, d', was calculated; a separate null distribution was derived for each condition. The average firing rate during the target signal periods was compared with the corresponding null distribution to estimate a separate d' for each SCR and condition (Fig. 4*B*). The d' of 0.4 was used to calculate SCR thresholds and the corresponding CMR (threshold difference between the Codeviant and Comodulated conditions). Note that the absolute d' values cannot be interpreted in a conventional manner given that the choice of window used to estimate the null-distribution parameters introduces an arbitrary scaling; thus, our choice of the d' criterion to calculate CMR was instead based on avoiding floor and ceiling effects. Results indicate that the CMR circuit model shows a CMR effect consistent with actual cochlear nucleus data in that signal detectability is best in the Comodulated condition, followed by the Reference and Codeviant conditions (compare Figs. 4*A* and *B* with Figs. 2 and 6*A* from Pressnitzer et al., 2001). The size of the predicted CMR effect is also consistent with perceptual measurements (Mok et al., 2021). As expected, no CMR effect is seen at the level of the auditory nerve. Thus, the CMR circuit model accounts for the improved signal representation in the Comodulated condition where the masker is more easily segregable from the target signal, an advantage that derives from the fact that the different masking components are temporally coherent with one another. In addition, it also accounts for the greater cross-channel interference in the Codeviant condition, where the flanking components are temporally coherent with the target signal that is presented in the dips of the OFC. Finally, when the modulation rate of the input SAM tones was varied, CMR effects were still seen (Fig. 4*C*) and followed the same low-pass trend as human perceptual data (Carlyon et al., 1989).

Each scene analysis model was separately calibrated by fitting a logistic/sigmoid function mapping the neural consonant confusion matrix entries from that model for the SiSSN condition to corresponding perceptual measurements. The mapping derived from this calibration was used to predict perceptual consonant confusion matrices from the corresponding neural confusion matrices for unseen conditions. Voicing, POA, and MOA confusion matrices were then derived by pooling over all consonants. Finally, the Pearson correlation coefficient was used to compare model predictions to perceptual measurements across the voicing, POA, and MOA categories. The prediction accuracy for the different models is reported in Results.

Statistical analysis. Permutation testing (Nichols and Holmes, 2002) with multiple-comparisons correction at 5% false discovery rate (FDR; Benjamini and Hochberg, 1995) was used to extract significant differences in the SiSSN, SiB, SiDCmod, and Vocoded SiB consonant confusion matrices quantified earlier (see Quantifying confusion matrices from perceptual measurements). The null distributions for permutation testing were obtained using a nonparametric shuffling procedure, which

Table 5. Improvement in prediction accuracy offered by the across-channel model compared to the within-channel model

Condition	Diagonal entries			Off-diagonal entries		
	Improvement	Uncorrected *p*-value	Significant under 5% FDR threshold?	Improvement	Uncorrected *p*-value	Significant under 5% FDR threshold?
SiB	12%	0.0225	Yes	8%	0.0406	Yes
SiDCmod	22%	$<10^{-5}$	Yes	8%	0.1006	No
Vocoded SiB	59%	$<10^{-5}$	Yes	30%	$<10^{-5}$	Yes

The across-channel model showed improved correlations between model predictions and perceptual measurements for all the unseen conditions, with the largest improvement apparent for Vocoded SiB.

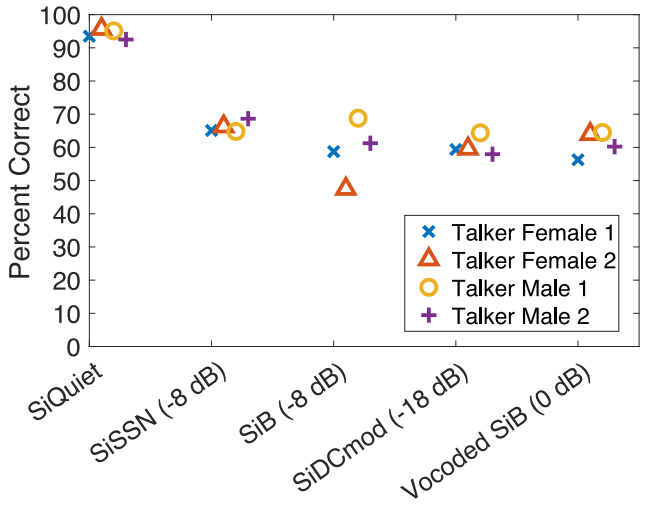

Figure 5. Overall intelligibility measured in the online consonant identification study for different conditions and talkers. Approximately equal overall intelligibility was achieved for SiSSN, SiDCmod, SiB, and Vocoded SiB ($N = 191$).

ensured that the data used in the computation of the null distributions had the same statistical properties as the measured confusion data. A separate null distribution was generated for each consonant. Each realization from each null distribution was obtained by following the same computations used to obtain the actual differences in the confusion matrices across conditions but with random shuffling of condition labels corresponding to the measurements. This procedure was independently repeated with 10,000 distinct randomizations for each null distribution.

The *p*-values for the Pearson correlation coefficients between model predictions and perceptual measurements (Tables 3, 4) were derived using Fisher's approximation (Fisher, 1921).

To test whether the improvements in prediction accuracy (i.e., the correlation between model predictions and perceptual measurements) offered by the across-channel model compared to the within-channel model are statistically significant, a permutation procedure was used once again. Under the null hypothesis that the within- and across-channel models are equivalent in their predictive power, the individual entries of the confusion matrices predicted by the two models can be swapped without effect on the results. Thus, to generate each realization of the null distribution of the correlation improvement, a randomly chosen half of the confusion matrix entries were swapped; this permutation procedure was independently repeated 100,000 times. A separate null distribution was generated in this manner for each condition. The actual improvements in correlation were compared with the corresponding null distributions to estimate (uncorrected) *p*-values. To adjust for multiple testing, an FDR procedure (Benjamini and Hochberg, 1995) was used. Table 5 indicates whether each test met criteria for statistical significance under an FDR threshold of 5%.

Code Accessibility. Subjects were directed from Prolific.co to the SNAPlabonline psychoacoustics infrastructure (Bharadwaj, 2021; Mok et al., 2021) to perform the study. Offline data analyses were performed using custom software in Python (https://www.python.org) and MATLAB (MathWorks). The code for our computational models is publicly available on GitHub at https://github.com/vibhaviswana/modeling-consonant-confusions.

Results

Our aim was to test the hypothesis that speech understanding in noise is shaped by aspects of temporal-coherence processing that exist in early auditory areas. For this, we used a combination of online consonant identification experiments and computational modeling. In particular, we compared consonant confusion predictions from a model of across-channel temporal-coherence processing that is physiologically plausible in the cochlear nucleus with predictions from a purely within-channel model inspired by current speech-intelligibility models (see Scene analysis modeling to predict consonant confusions).

Figure 5 shows speech intelligibility measurements from the online consonant identification study. Approximately equal overall intelligibility was achieved for SiSSN, SiDCmod, SiB, and Vocoded SiB because of our careful choice of SNRs for these conditions based on piloting (Table 1). This was done to obtain roughly equal variance in the consonant confusion estimates for these conditions, which allows us to fairly compare confusion patterns across them. Equalizing intelligibility also maximizes the statistical power for detecting differences in the pattern of confusions. ~60% overall intelligibility was obtained in each condition, which yielded a sufficient number of confusions for analysis.

The identification data collected in the online experiment were used to construct a consonant confusion matrix for each condition, then statistically significant differences in these matrices across conditions were extracted (see Quantifying confusion matrices from perceptual measurements and Statistical analysis). Results (Fig. 6) show significant differences in the confusion patterns across (1) conditions with different masker modulation statistics, and (2) stimuli with intact versus degraded TFS information. Computational modeling was then used to predict these differences across conditions to test our hypothesis about the role of temporal-coherence processing in scene analysis.

We constructed the following different models of scene analysis: (1) a within-channel model, which simulates masking of target-speech envelopes by distracting masker modulations within individual frequency channels, and (2) an across-channel model, which simulates across-channel temporal-coherence processing to account for interference from masker elements that are temporally coherent with target elements but in different frequency channels (see Scene analysis modeling to predict consonant confusions). We derived a separate neural confusion matrix for each model and listening condition. Then, each scene analysis model was separately calibrated by fitting a nonlinear mapping relating the neural consonant confusion matrix entries derived from that model for the SiSSN condition to corresponding perceptual measurements. Once fit, this mapping was used to quantitatively predict perceptual consonant confusions for novel conditions not used in calibration. Figure 7 shows results from the calibration step. In this figure, the different entries of the measured

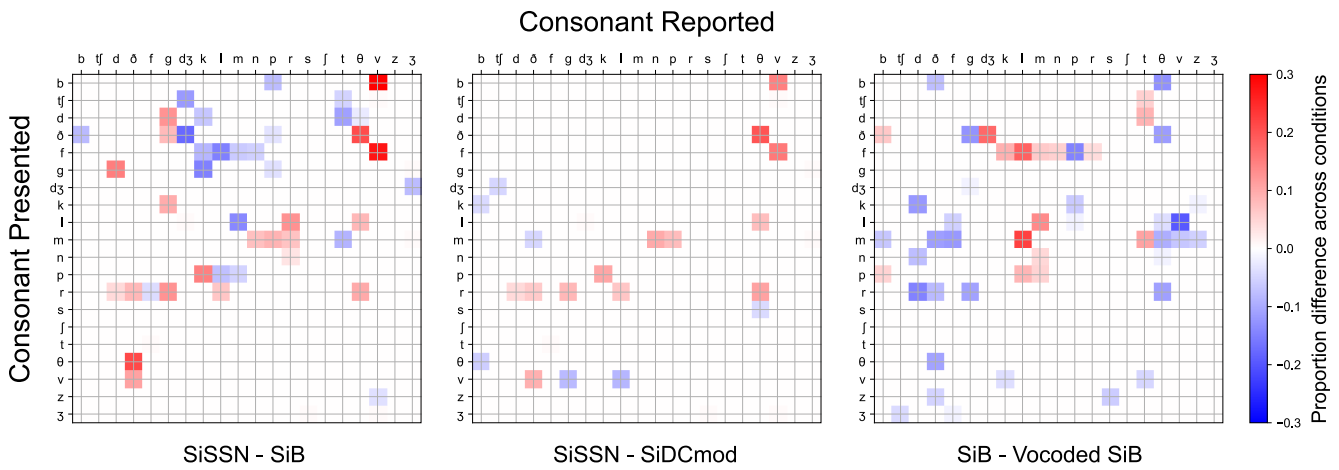

Figure 6. Measured consonant confusion-matrix differences across conditions (pooled over samples; $N = 191$). The first two plots represent differences across maskers with different modulation spectra, whereas the third plot shows the difference across stimuli with intact versus degraded TFS information. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR). As the modulation statistics of the masker or the TFS content were varied, statistically significant differences emerged in the confusion patterns across conditions. Overall intelligibility was normalized to 60% for this analysis (see Quantifying confusion matrices from perceptual measurements) so that differences in confusion matrices across conditions could be attributed to changes in consonant categorization and category errors rather than differences in overall error counts because of one condition being inherently easier at a particular SNR.

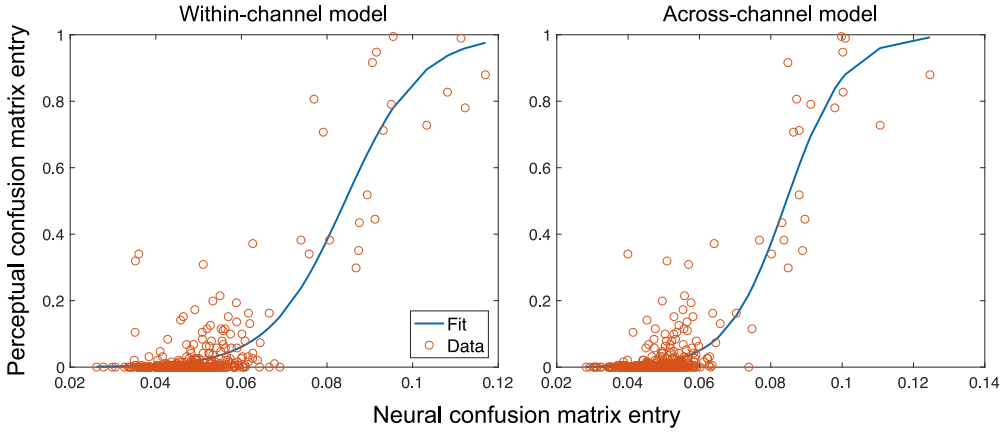

Figure 7. Calibration result for the within- and across-channel models of scene analysis. The different entries of the measured perceptual confusion matrix for the SiSSN condition are plotted (open circles) against the corresponding entries of the neural confusion matrix for SiSSN derived from the within-channel model (left) and across-channel model (right). The nonlinear relationship between these neural and perceptual data was fit using a sigmoid/logistic function (thick curve) separately for each model.

perceptual confusion matrix for SiSSN are plotted against the corresponding entries of the neural confusion matrix from each model for SiSSN. From this figure, it can be seen that the data show floor and ceiling effects, that is, as the neural metric increases (or decreases), the perceptual metric concomitantly increases (or decreases) but only up to a point, after which it saturates. This phenomenon is common to psychometric measurements. We fit this nonlinear relationship between the neural and perceptual data for SiSSN using a sigmoid/logistic function (Fig. 7; commonly used in the literature to obtain psychometric curves) separately for each model.

The model-specific mapping derived in the calibration step was used to predict perceptual consonant confusion matrices for each of the scene analysis models from the neural confusion matrices for unseen conditions (not used in calibration). Then, voicing, POA, and MOA confusion matrices were derived by pooling over all consonants (see Figs. 9, 10, 11). Finally, model predictions were compared with perceptual measurements for the different confusion matrix entries across the voicing, POA, and MOA categories. The results are shown in Figure 8 for SiB, SiDCmod, and Vocoded SiB. Visual comparison of the plotted data against the line of equality in Figure 8 suggests that there is a prediction bias for the SiDCmod and Vocoded SiB conditions for both the within- and across-channel models. This bias likely arises from our choice of calibration function (i.e., the sigmoid function) and the fact that calibration parameters were fitted to SiSSN, which may be suboptimal for the other conditions. Nonetheless, it can be seen that the cluster of points is less dispersed for the across-channel model compared to the within-channel model, indicating greater predictive accuracy for the across-channel model. The SiQuiet condition is not visualized, as there were ceiling effects in the intelligibility measurements (i.e., the diagonal entries of the confusion matrix were dominant) and very few confusions (i.e., off-diagonal entries were rare), which made it infeasible to meaningfully evaluate the quality of predictions for this condition (as there was no variance across either the on- or off-diagonal entries). But overall, across all entries for SiQuiet, both models predicted diagonal entries

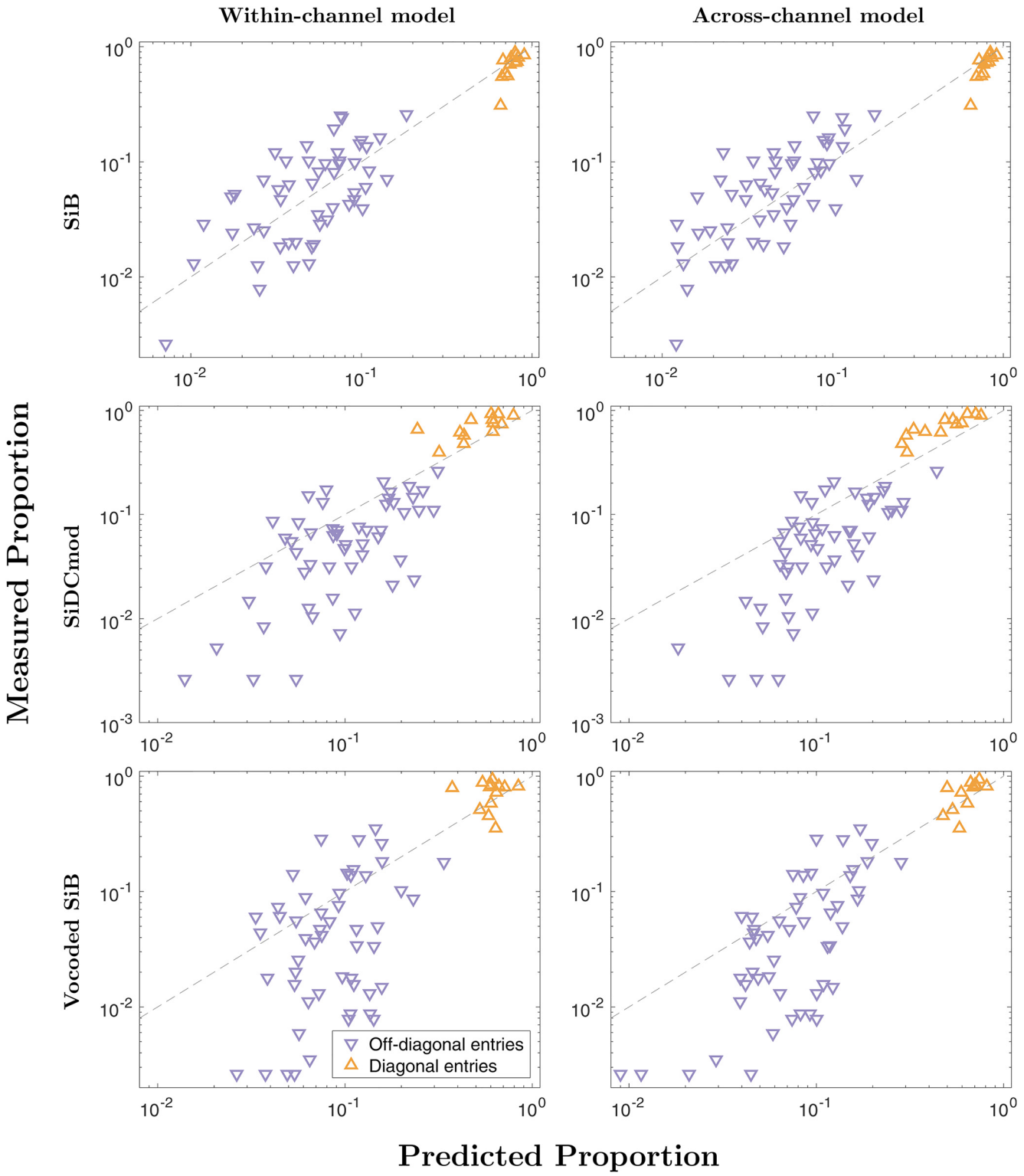

Figure 8. Within- and across-channel model predictions versus measured confusion matrix entries for the unseen conditions. Diagonal entries correspond to proportion correct scores for the different consonant phonetic categories (voicing, POA, and MOA), and off-diagonal entries correspond to true confusions. The line of equality is shown as a dashed gray line. Pearson correlation coefficients between model predictions and perceptual measurements are quantified in Tables 3 and 4.

close to one and off-diagonal entries close to zero, in line with perceptual measurements.

Pearson correlation coefficients were computed between the model predictions and perceptual measurements for the unseen conditions (Fig. 8) as well as for SiSSN (i.e., the calibration condition); the results are given in Tables 3 and 4 for the within- and across-channel models, respectively. Because the range of confusion matrix entries spanned three orders of magnitude, all comparisons were performed with log-transformed values. The correlations were statistically significant across all nonvocoded conditions for the within-channel model and across all conditions for the across-channel model (see Statistical analysis). The

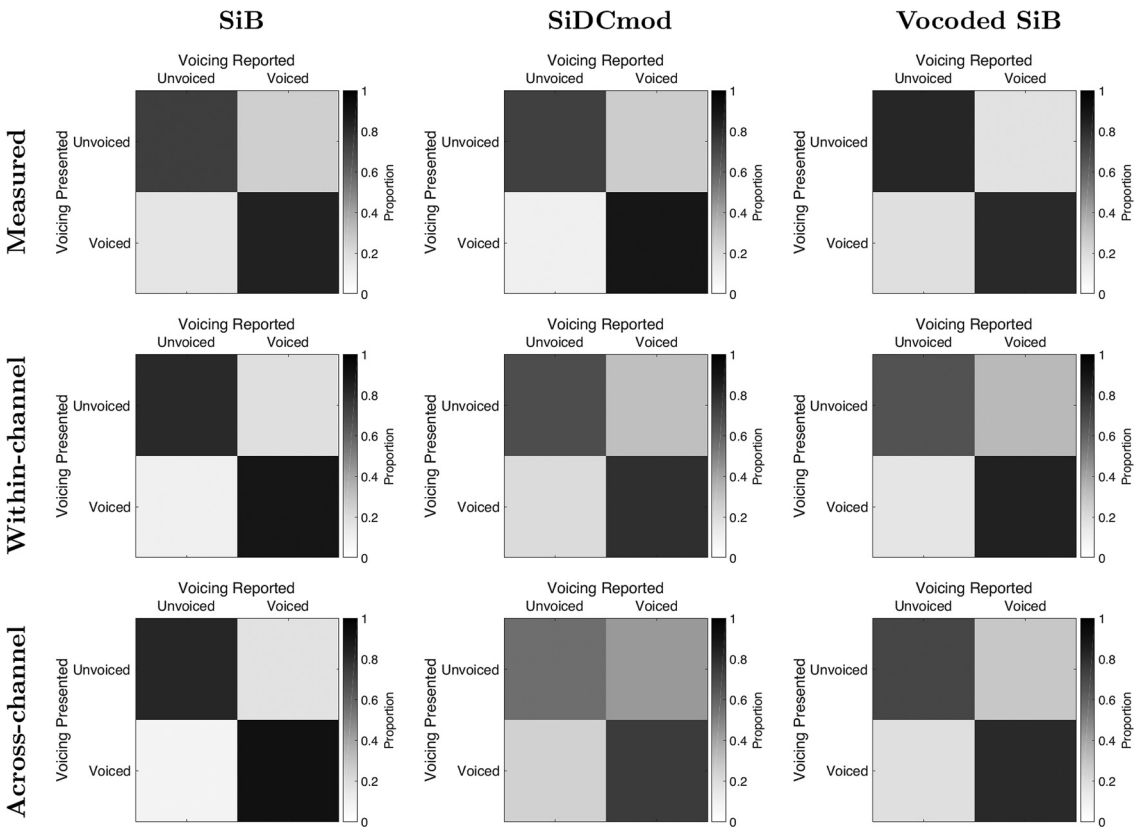

Figure 9. Full set of measured (top row) and model-predicted (middle, bottom rows) voicing confusion matrices.

strong correlation of the within-channel model predictions with perceptual data in the nonvocoded conditions (where TFS cues are preserved) provides independent evidence that speech understanding is strongly influenced by modulation masking when TFS cues are available (Viswanathan et al., 2021a); moreover, this result also suggests that modulations are used differently by the brain in the absence of natural TFS.

The across-channel model produced stronger correlation values compared to the within-channel model for all conditions, and the improvements were statistically significant across all conditions even after correcting for multiple comparisons (Table 5; see Statistical analysis). Thus, a simple physiologically plausible model of across-channel cochlear nucleus processing that shows CMR (Fig. 4) also yields category confusion predictions that match behavioral data and more specifically improves predictions compared to a within-channel model. Note that our within-channel model assumes perfect segregability of target-masker components that are separated in CF and MF (in line with current speech-intelligibility models; Jørgensen et al., 2013; Relaño-Iborra et al., 2016), and only models within-channel modulation masking. Specifically, within a particular channel (i.e., CF) and MF, masker modulations that are not in phase with the target are the only components that mask the target. However, our across-channel model simulates both within-channel modulation masking and cross-channel temporal-coherence-based interference. Specifically, masker components that are in a different channel from the target but that are temporally coherent with the target can interfere with target coding and perception. We implemented this interference via the CMR circuit model (Fig. 1), where temporally coherent pieces of the target and masker, even across distinct cochlear channels, coherently drive the wideband inhibitor, thereby enhancing outputs of the narrowband cell (which is inhibited by the wideband inhibitor) that are incoherent with the masker. Thus, our finding that model predictions are improved when cross-channel processing is added is consistent with the theory that across-channel temporal coherence shapes scene analysis (Elhilali et al., 2009). Moreover, this result also suggests that physiological computations that exist as early as the cochlear nucleus can contribute significantly to temporal-coherence-based scene analysis. Note that improvements to confusion predictions are apparent with the across-channel model for the same range of model parameters for which the CMR effect is also apparent.

Another key result from Table 5 is that the condition that showed the greatest improvement in confusion matrix predictions between the within- and across-channel models is Vocoded SiB. The masker in Vocoded SiB produces both within-channel modulation masking and cross-channel interference (as described above). These masking and interference effects are partially mitigated in SiB (and other nonvocoded conditions) compared to Vocoded SiB because the brain can use the pitch cue supplied by natural TFS to better separate the target and masker (Darwin, 1997; Oxenham and Simonson, 2009). The across-channel model is a better fit to perceptual data for all conditions, which suggests that cross-channel interference affects perceptual data. Thus, the improvement offered by this model is likely most apparent for Vocoded SiB because cross-channel interference effects contribute most to perception in this condition.

Note that while the main difference between the two scene analysis models tested in the current study is the exclusion/inclusion of cross-channel processing, another difference is that the within-channel model discards TFS, whereas the across-channel model uses the full simulated auditory-nerve output to drive the CMR circuit model. This raises the possibility that part of the

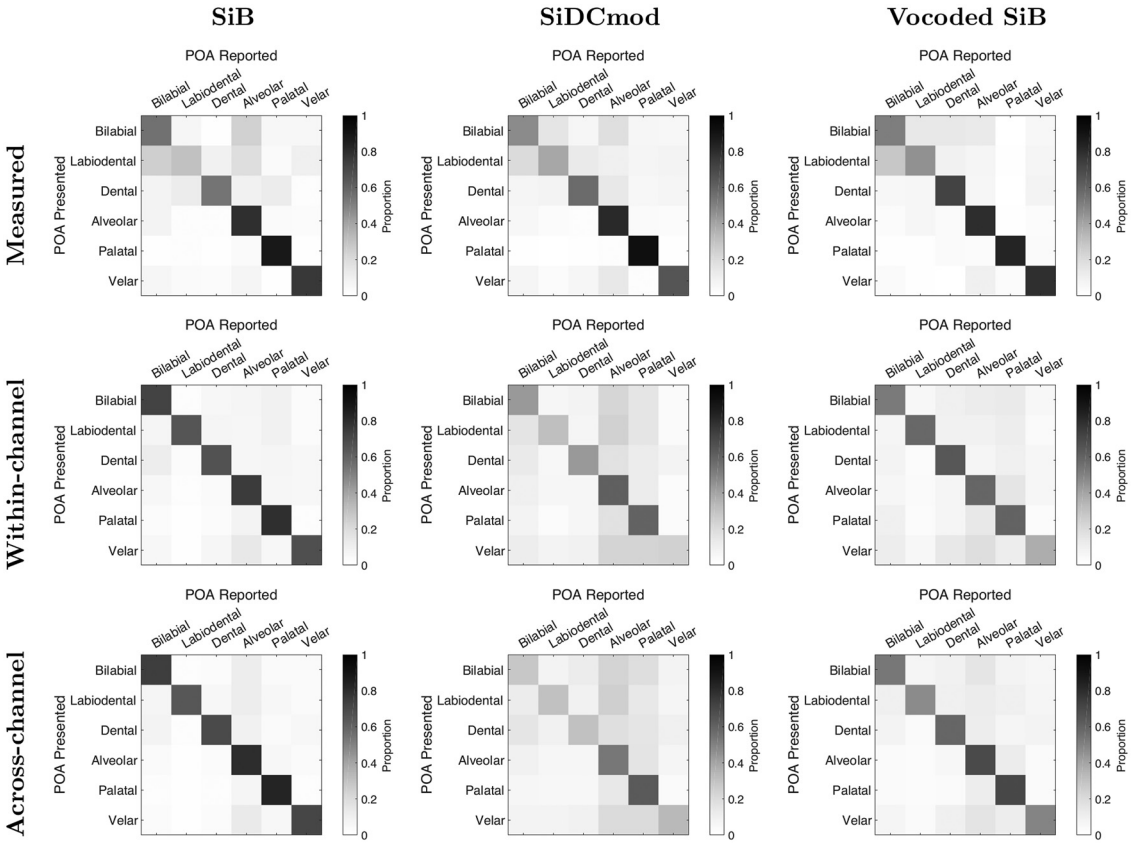

Figure 10. Full set of measured (top row) and model-predicted (middle, bottom rows) POA confusion matrices.

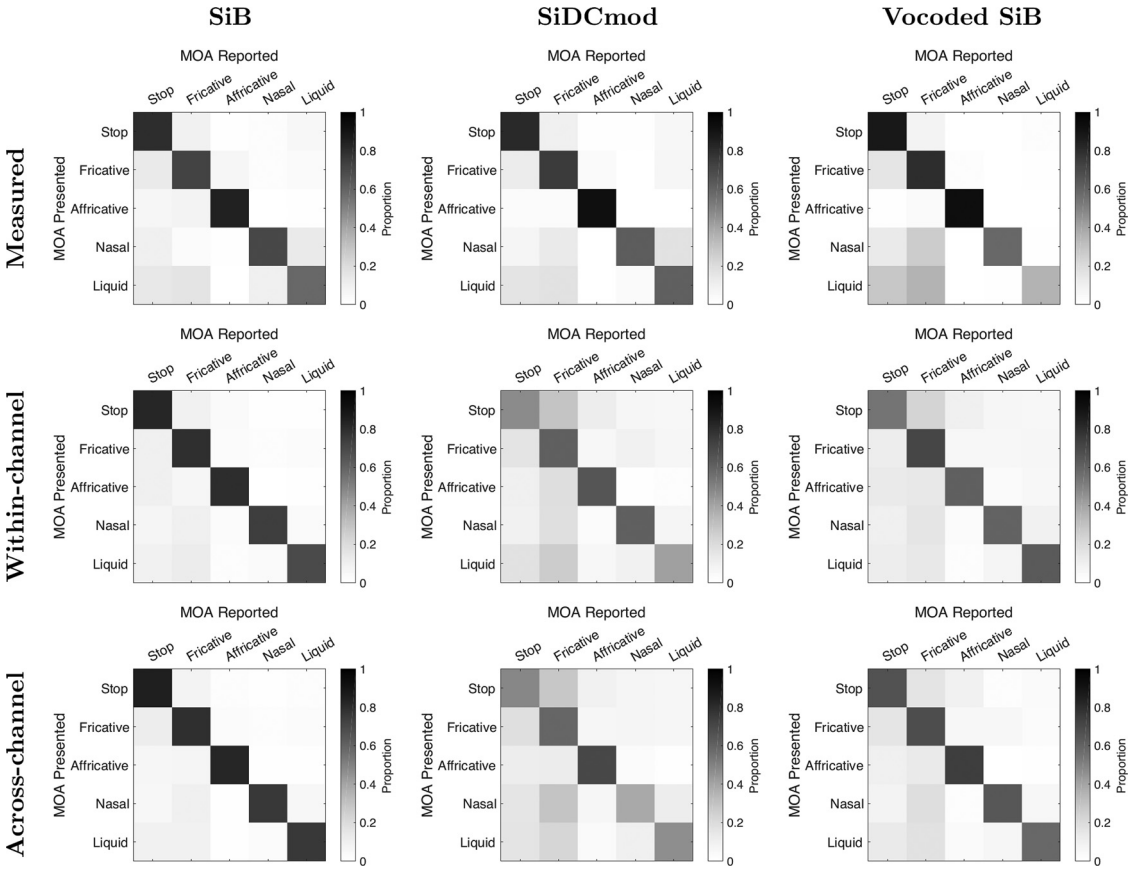

Figure 11. Full set of measured (top row) and model-predicted (middle, bottom rows) MOA confusion matrices.

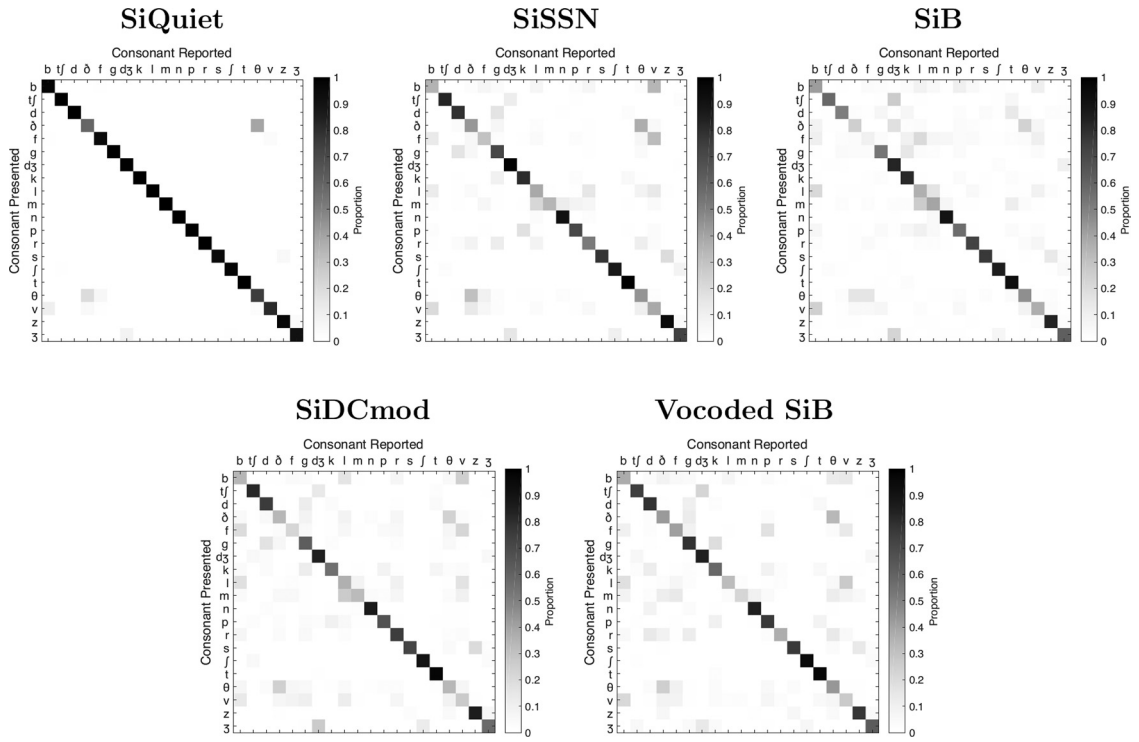

Figure 12. Raw consonant confusion matrix measurements for all conditions (pooled over samples). Overall intelligibility was ~90% for the SiQuiet condition and ~60% for the SiSSN, SiB, SiDCmod, and Vocoded SiB conditions (Fig. 5).

improvement offered by the across-channel model could come simply from the inclusion of TFS information within each channel independently. To investigate whether the poorer performance of the within-channel model was partly because of discarding TFS, we reran the within-channel model by retaining the full auditory-nerve output (data not shown). We found that the predictions from the modified within-channel model were not significantly better than those of the original within-channel model. This confirms that the improvement in predictions given by the across-channel model comes largely from across-channel CMR effects, suggesting that categorical perception is sensitive to the temporal coherence across channels. Moreover, these CMR effects were restricted to low rates (<80 Hz or so; Fig. 4*C*), consistent with perceptual data (Carlyon et al., 1989). This suggests that the cross-channel processing did not benefit much from the TFS information included in driving the CMR circuit model.

For completeness, the full set of model-predicted and measured perceptual confusion matrices are shown for the voicing, POA, and MOA categories (Figs. 9, 10, 11); results are shown only for the SiB, SiDCmod, and Vocoded SiB conditions (i.e., the conditions unseen by the calibration step and having a sufficient number of confusions for prediction). In addition, the raw consonant confusion matrix measurements for all conditions are shown in Figure 12.

Discussion

To probe the contribution of temporal-coherence processing to speech understanding in noise, the present study used a behavioral experiment to measure consonant identification in different masking conditions in conjunction with physiologically plausible computational modeling. To the best of our knowledge, this is the first study to use confusion patterns in speech categorization to test theories of auditory scene analysis. The use of confusion data provides independent constraints on our understanding of scene analysis mechanisms beyond what overall intelligibility can provide. This is because percent correct data only convey binary information about whether target coding was intact, whereas consonant categorization and confusion data provide richer information about what sound elements received perceptual weighting.

We constructed computational models simulating (1) purely within-channel modulation masking (in line with current speech-intelligibility models; Relaño-Iborra et al., 2016), and (2) a combination of within-channel modulation masking and across-channel temporal-coherence processing mirroring physiological computations that are known to exist in the cochlear nucleus (Pressnitzer et al., 2001). Our across-channel temporal-coherence circuit produced a CMR effect (Fig. 4) that is consistent with actual cochlear nucleus data (Pressnitzer et al., 2001) and perceptual measurements (Mok et al., 2021). Moreover, consonant confusion pattern predictions were significantly improved for all tested conditions with the addition of this cross-channel processing (Table 5), which suggests that temporal-coherence processing strongly shapes speech categorization when listening in noise. This result is consistent with the theory that comodulated features of a sound source are perceptually grouped together and that masker elements that are temporally coherent with target speech but in a different channel from the target perceptually interfere (Schooneveldt and Moore, 1987; Darwin, 1997; Apoux and Bacon, 2008). The only case where the within- and across-channel models were statistically equivalent was in predicting the off-diagonal entries (i.e., true confusions) for the SiDCmod condition; this may be because this condition has little coherent cross-channel interference from the masker since the masker is unmodulated (Stone et al., 2012).

An important difference between the cross- and within-channel masking simulated in our models is that while the cross-channel interference was produced by masker fluctuations that

were temporally coherent with the target, the within-channel masking was produced by masker components that were matched in both CF and MF with target components. While current speech-intelligibility models simulate the latter type of masking (Jørgensen et al., 2013; Relaño-Iborra et al., 2016), they do not account for cross-channel temporal-coherence-based masking as we have done here. This may explain why these models fail in certain conditions, including for vocoded stimuli (Steinmetzger et al., 2019). Indeed, even in the present study, although our within-channel modulation masking model reasonably accounted for category confusions, it failed when TFS cues were unavailable (Table 3). One explanation for this is that because pitch-based masking release is poorer in the vocoded condition due to degraded TFS information (Oxenham and Simonson, 2009), the effects of cross-channel interference are more salient. This may also be the reason why the Vocoded SiB condition showed the greatest improvement in confusion pattern predictions after adding cross-channel processing (Table 5), which models these interference effects.

Although the lateral inhibition network used in Elhilali et al. (2003) bears some similarities to the across-channel CMR circuit model used in the current study, the CMR circuit model was explicitly based on physiological computations present in the cochlear nucleus and their CMR properties. Thus, another implication of the results of the present study is that physiological computations that exist as early as the cochlear nucleus can contribute significantly to temporal-coherence-based scene analysis. Such effects likely accumulate as we ascend along the auditory pathway (Elhilali et al., 2009; Teki et al., 2013; O'Sullivan et al., 2015). Indeed, scene analysis may in general be supported by a cascade of mechanisms throughout the auditory pathway, including both early-stage processing as well as cortical mechanisms. For example, there are circuits in the brainstem, midbrain, and cortex that exhibit sensitivity and selectivity to different spectrotemporal regularities in the input (Nelken et al., 1999; Pressnitzer et al., 2008; Kondo and Kashino, 2009; Diepenbrock et al., 2017; Mishra et al., 2021); these properties may in turn support auditory-object formation and scene segregation. Moreover, top-down cognitive processes such as attention can also contribute to scene analysis, especially when sound elements are otherwise perceptually similar (Shinn-Cunningham, 2020). Thus, future studies should explore the contributions of scene analysis mechanisms at different levels of the hierarchy of auditory processing.

The CMR circuit model used in the current study does not perform pitch-range temporal-coherence processing, and no CMR effect was seen at high modulation rates (Fig. 4C), consistent with perceptual data in the literature (Carlyon et al., 1989). Despite this, our across-channel model significantly improved predictions of category confusions compared to the within-channel model, which suggests that temporal-coherence processing at lower modulation rates is perceptually important. A future research direction is to extend the modeling framework proposed here to study the contributions of scene analysis mechanisms beyond the specific aspects of temporal-coherence processing studied here. One such extension could be to account for pitch-based source segregation (Bregman, 1990), perhaps by modeling a combined temporal-place code for pitch processing (Shamma and Klein, 2000; Oxenham et al., 2004; Oxenham and Simonson, 2009).

One limitation of the periphery model that we used (Bruce et al., 2018) is that it was developed to match nerve responses to simple stimuli. However, this family of periphery models has been successfully used to account for complex phenomena such as synchrony capture (Delgutte and Kiang, 1984), formant coding in the midbrain (Carney et al., 2015), and qualitative aspects of evoked potentials such as auditory brainstem responses and frequency-following responses (Shinn-Cunningham et al., 2013). Although a debate exists regarding the spatiotemporal properties of different periphery models in cochlear responses (Verhulst et al., 2015; Vecchi et al., 2021), those differences are subtle compared to the slower CMR effects that are important for the present study. A more general limitation of the models used in this study is that they are simple and do not incorporate many aspects of speech perception (e.g., context effects; Dubno and Levitt, 1981) because the goal here is to test specific theories of scene analysis. Nevertheless, the contrast between the models would be unaffected by these higher-order effects.

References

Apoux F, Bacon SP (2008) Selectivity of modulation interference for consonant identification in normal-hearing listeners. J Acoust Soc Am 123:1665–1672.

Bacon SP, Grantham DW (1989) Modulation masking: effects of modulation frequency, depth, and phase. J Acoust Soc Am 85:2575–2580.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Series B Stat Methodol 57:289–300.

Bharadwaj H (2021) Haribharadwaj/SNAPlabonline: SNAPlabonline, a Django-based web application for conducting psychoacoustics on the web from the Systems Neuroscience of Auditory Perception Lab. Zenodo. doi:10.5281/zenodo.4743851.

Bregman A (1990) Auditory scene analysis: the perceptual organization of sound. Cambridge, MA: MIT Press.

Bruce IC, Erfani Y, Zilany MS (2018) A phenomenological model of the synapse between the inner hair cell and auditory nerve: implications of limited neurotransmitter release sites. Hear Res 360:40–54.

Carlyon RP, Buus S, Florentine M (1989) Comodulation masking release for three types of modulator as a function of modulation rate. Hear Res 42:37–45.

Carney LH, Li T, McDonough JM (2015) Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations. Eneuro 2:ENEURO.0004-15.2015.

Crouzet O, Ainsworth WA (2001) On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation. Paper presented at the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis. Aalborg, Denmark, September.

Darwin CJ (1997) Auditory grouping. Trends Cogn Sci 1:327–333.

Delgutte B, Kiang NY (1984) Speech coding in the auditory nerve: i. vowel-like sounds. J Acoust Soc Am 75:866–878.

Diepenbrock J-P, Jeschke M, Ohl FW, Verhey JL (2017) Comodulation masking release in the inferior colliculus by combined signal enhancement and masker reduction. J Neurophysiol 117:853–867.

Dubno JR, Levitt H (1981) Predicting consonant confusions from acoustic analysis. J Acoust Soc Am 69:249–261.

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. Speech Commun 41:331–348.

Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron 61:317–329.

Ewert SD, Dau T (2000) Characterizing frequency selectivity for envelope fluctuations. J Acoust Soc Am 108:1181–1196.

Fisher RA (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1:1–32.

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

Heinz MG, Colburn HS, Carney LH (2002) Quantifying the implications of nonlinear cochlear tuning for auditory-filter estimates. J Acoust Soc Am 111:996–1011.

Hilbert D (1906) Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen. Vierte Mitteilung. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse 1906:157–228.

Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical signs of selective attention in the human brain. Science 182:177–180.

Jørgensen S, Ewert SD, Dau T (2013) A multi-resolution envelope-power based model for speech intelligibility. J Acoust Soc Am 134:436–446.

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S (2004) Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. J Acoust Soc Am 116:2395–2405.

Kondo HM, Kashino M (2009) Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. J Neurosci 29:12695–12701.

Krishnan L, Elhilali M, Shamma S (2014) Segregating complex sound sources through temporal coherence. PLoS Comput Biol 10:e1003985.

Meddis R, Delahaye R, O'Mard L, Sumner C, Fantini DA, Winter I, Pressnitzer D (2002) A model of signal processing in the cochlear nucleus: comodulation masking release. Acta Acust united Ac 88:387–398.

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. J Acoust Soc Am 27:338–352.

Mishra AP, Peng F, Li K, Harper N, Schnupp JW (2021) Sensitivity of the neural responses to statistical features of sound textures in the inferior colliculus. Hear Res 412:108357.

Mok BA, Viswanathan V, Borjigin A, Singh R, Kafi HI, Bharadwaj HM (2021) Web-based psychoacoustics: hearing screening, infrastructure, and validation. bioRxiv. 443520. doi:10.1101/2021.05.10.443520.

Nelken I, Rotman Y, Yosef OB (1999) Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397:154–157.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

O'Sullivan JA, Shamma SA, Lalor EC (2015) Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. J Neurosci 35:7256–7263.

Oxenham AJ, Shera CA (2003) Estimates of human cochlear tuning at low levels using forward and simultaneous masking. J Assoc Res Otolaryngol 4:541–554.

Oxenham AJ, Simonson AM (2009) Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference. J Acoust Soc Am 125:457–468.

Oxenham AJ, Bernstein JG, Penagos H (2004) Correct tonotopic representation is necessary for complex pitch perception. Proc Natl Acad Sci U S A 101:1421–1425.

Patterson RD, Nimmo-Smith I, Holdsworth J, Rice P (1987) An efficient auditory filterbank based on the gammatone function. In a meeting of the IOC Speech Group on Auditory Modelling at RSRE, Vol 2, No. 7.

Phatak SA, Allen JB (2007) Consonant and vowel confusions in speech-weighted noise. J Acoust Soc Am 121:2312–2326.

Pressnitzer D, Meddis R, Delahaye R, Winter IM (2001) Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus. J Neurosci 21:6377–6386.

Pressnitzer D, Sayles M, Micheyl C, Winter I (2008) Perceptual organization of sound begins in the auditory periphery. Curr Biol 18:1124–1128.

Qin M, Oxenham A (2003) Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J Acoust Soc Am 114:446–454.

Rabiner L (1993) Fundamentals of speech recognition. Englewood Cliffs, NJ: Prentice Hall.

Relaño-Iborra H, May T, Zaar J, Scheidiger C, Dau T (2016) Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain. J Acoust Soc Am 140:2670–2679.

Schooneveldt GP, Moore BC (1987) Comodulation masking release (CMR): Effects of signal frequency, flanking-band frequency, masker bandwidth, flanking-band level, and monotic versus dichotic presentation of the flanking band. J Acoust Soc Am 82:1944–1956.

Shamma S, Klein D (2000) The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. J Acoust Soc Am 107:2631–2644.

Shera CA, Guinan JJ, Oxenham AJ (2002) Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proc Natl Acad Sci U S A 99:3318–3323.

Shinn-Cunningham B (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182–186.

Shinn-Cunningham BG (2020) Brain mechanisms of auditory scene analysis. In: The cognitive neurosciences, pp 159–166. Cambridge: MIT Press.

Shinn-Cunningham B, Ruggles DR, Bharadwaj H (2013) How early aging and environment interact in everyday listening: from brainstem to behavior through modeling. In: Basic aspects of hearing, pp 501–510. New York, NY: Springer.

Steinmetzger K, Zaar J, Relaño-Iborra H, Rosen S, Dau T (2019) Predicting the effects of periodicity on the intelligibility of masked speech: an evaluation of different modelling approaches and their limitations. J Acoust Soc Am 146:2562–2576.

Stone MA, Moore BC (2014) On the near non-existence of "pure" energetic masking release for speech. J Acoust Soc Am 135:1967–1977.

Stone MA, Füllgrabe C, Moore BC (2012) Notionally steady background noise acts primarily as a modulation masker of speech. J Acoust Soc Am 132:317–326.

Swaminathan J, Heinz MG (2011) Predicted effects of sensorineural hearing loss on across-fiber envelope coding in the auditory nerve. J Acoust Soc Am 129:4001–4013.

Teki S, Chait M, Kumar S, Shamma S, Griffiths TD (2013) Segregation of complex acoustic scenes based on temporal coherence. Elife 2:e00699.

Vecchi AO, Varnet L, Carney LH, Dau T, Bruce IC, Verhulst S, Majdak P (2021) A comparative study of eight human auditory models of monaural processing. arXiv:2107.01753.

Verhulst S, Bharadwaj HM, Mehraei G, Shera CA, Shinn-Cunningham BG (2015) Functional modeling of the human auditory brainstem response to broadband stimulation. J Acoust Soc Am 138:1637–1659.

Viswanathan V, Bharadwaj HM, Shinn-Cunningham BG, Heinz MG (2021a) Modulation masking and fine structure shape neural envelope coding to predict speech intelligibility across diverse listening conditions. J Acoust Soc Am 150:2230–2244.

Viswanathan V, Shinn-Cunningham BG, Heinz MG (2021b) Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble. J Acoust Soc Am 150:2664–2676.

Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244.

Winter IM, Palmer AR (1990) Responses of single units in the anteroventral cochlear nucleus of the guinea pig. Hear Res 44:161–178.

Zurek PM (1993) Binaural advantages and directional effects in speech intelligibility. In: Acoustical factors affecting hearing aid performance (Studebaker GA, Hochberg I, eds), pp 255–275. Boston: Allyn & Bacon.