NEUROPHYSIOLOGICAL MECHANISMS OF SPEECH INTELLIGIBILITY UNDER MASKING AND DISTORTION

by

Vibha Viswanathan

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Weldon School of Biomedical Engineering

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Michael G. Heinz, Co-Chair

Weldon School of Biomedical Engineering

Dr. Barbara G. Shinn-Cunningham, Co-Chair

Neuroscience Institute, Carnegie Mellon University

Dr. Edward L. Bartlett

Weldon School of Biomedical Engineering

Dr. Joshua M. Alexander

Department of Speech, Language, and Hearing Sciences

Approved by:

Dr. Andrew O. Brightman

ACKNOWLEDGMENTS

I owe thanks to several people who have contributed in various ways to this dissertation. I would like to start by thanking my advisors Michael Heinz and Barbara Shinn-Cunningham for their excellent mentorship, support, and friendship over the years that I have worked with them. I greatly appreciate them giving me complete freedom to pursue my research ideas while making themselves available whenever I needed advice. I also thank Edward Bartlett and Joshua Alexander for taking the time to be on my thesis committee and for valuable feedback on this dissertation.

I'm grateful to the Lab in Multisensory Neuroscience (BU/CMU) for introducing me to the wonderful, collegial world of Auditory Neuroscience research, and for encouraging me to apply to Ph.D. programs in this field. I thank them, as well as all members of the Auditory Neurophysiology and Modeling lab (Purdue) and the larger hearing science community at Purdue for contributing many insightful discussions as well as their friendship, support, and encouragement over my years as a Ph.D. student.

I would like to thank the National Institute of Deafness and Communication Disorders for awarding me a Ruth L. Kirschstein National Research Service Award (NRSA) Individual Predoctoral Fellowship (F31DC017381), which funded the majority of my dissertation work. Additional research funding was provided by other grants from the National Institutes of Health [R01DC009838 (to M.G.H.), 9605702, R01DC013825 (to B.G.S-C.), R01DC015988 (to B.G.S-C.), and R01DC015989 (to H.M.B.)], Action on Hearing Loss [G72 (to M.G.H.)], and Office of Naval Research [ONR N00014-20-12709 (to B.G.S.-C.)].

Finally, I am deeply grateful to my family, especially my parents and my husband Hari, whose constant support, encouragement, love, faith, and patience over the years have been invaluable for the successful completion of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Difficulty understanding speech in background noise is the most common hearing complaint. Elucidating the neurophysiological mechanisms underlying speech intelligibility in everyday environments with multiple sound sources and distortions is hence important for any technology that aims to improve real-world listening. Using a combination of behavioral, electroencephalography (EEG), and computational modeling experiments, this dissertation provides insight into how the brain analyzes such complex scenes, and what roles different acoustic cues play in facilitating this process and in conveying phonetic content. Experiment #1 showed that brain oscillations selectively track the temporal envelopes (i.e., modulations) of attended speech in a mixture of competing talkers, and that the strength and pattern of this attention effect differs between individuals. Experiment #2 showed that the fidelity of neural tracking of attended-speech envelopes is strongly shaped by the modulations in interfering sounds as well as the temporal fine structure (TFS) conveyed by the cochlea, and predicts speech intelligibility in diverse listening environments. Results from Experiments #1 and #2 support the theory that temporal coherence of sound elements across envelopes and/or TFS shapes scene analysis and speech intelligibility. Experiment #3 tested this theory further by measuring and computationally modeling consonant categorization behavior in a range of background noises and distortions. We found that a physiologically plausible model that incorporated temporal-coherence effects predicted consonant confusions better than conventional speech-intelligibility models, providing independent evidence that temporal coherence influences scene analysis. Finally, results from Experiment #3 also showed that TFS is used to extract speech content (voicing) for consonant categorization even when intact envelope cues are available. Together, the novel insights provided by our results can guide future models of speech intelligibility and scene analysis, clinical diagnostics, improved assistive listening devices, and other audio technologies.

1. INTRODUCTION

Understanding speech in environments with interfering sound sources and distortions (e.g., reverberation, wind-noise reduction in cell phones and assistive listening devices, etc.) is one of the most complex tasks our brains solve everyday (Cherry, 1953), and is an ability unparalleled by machine algorithms (Loizou, 2013). Yet the mechanisms supporting this process are still poorly understood. Elucidating the neurophysiological mechanisms underlying speech intelligibility in such everyday environments is important not just from a basic-science perspective, but also for clinical applications and audio technologies. Indeed, clinical diagnostics, individualized interventions for speech-in-noise communication problems (e.g., hearing-aid fitting versus cognitive aural training), signal-processing strategies for assistive listening devices (e.g., cochlear implants and hearing aids), speech-denoising algorithms (e.g., in cell phones), and room acoustics design are all applications that may benefit from accurate characterization and modeling of speech intelligibility mechanisms.

Various physiological processes shape the internal representation of input speech sounds along the auditory pathway from the periphery to the cortex. Extracting intelligible information from a target speech source in a sound mixture crucially depends on the robust encoding of the input acoustics by the auditory periphery, the subsequent segregation of the target from other interfering sounds, selective attention to the segregated target source while effectively ignoring distracting sound sources, and further cognitive processing such as categorical perception (B. Shinn-Cunningham, 2008). Several fundamental questions remain about the precise mechanisms underlying scene segregation and target selection in everyday listening, as well as the roles of different acoustic cues in facilitating these processes and in conveying phonetic content. This dissertation uses a combination of behavioral, electrophysiological, and computational modeling experiments to address this significant gap. Each chapter in this dissertation is an individual study, and is written in the format of a journal article with sufficient background and discussion to stand on its own. The following paragraphs of this introductory chapter highlight the contents of the chapters to follow.

The cortical mechanisms underlying speech-on-speech selective attention were investigated in the electroencephalography (EEG) study presented in Chapter 2. This study examined

how a mixture of two speech streams is represented in the brain as subjects attended to one stream or the other. In particular, because brain rhythms are intimately associated with sensory processing (Buzsáki & Draguhn, 2004), we systematically investigated how brain oscillations in each of the canonical frequency bands are related to speech fluctuations, comparing when the speech is attended versus when it is ignored. In doing so, we addressed an important gap in the field where no prior studies had reported how the full complement of canonical brain oscillations track speech sounds in a mixture of competing sources. We show that EEG oscillations in the delta, theta, and low-gamma bands selectively track attended speech, a result that mechanistically links computational models of information routing in cortical circuits (Börgers et al., 2008) with attentive listening. Moreover, we found individual differences in the overall strength of the attention effects as well as in the specific speech and EEG features (channels and frequency bands) that were most informative about attentional focus.

The study described in Chapter 3 characterized the neural encoding of target speech in diverse acoustic scenes and examined the relationship between temporal coding, scene segregation, and target-speech intelligibility. Previous psychophysical studies suggested that slow temporal fluctuations (envelopes or modulations) convey most speech content (Shannon et al., 1995; Smith et al., 2002), and that the masking of envelopes in attended speech by those in interfering sounds (i.e., modulation masking) influences speech intelligibility (Bacon & Grantham, 1989; Dubbelboer & Houtgast, 2008; Jørgensen et al., 2013; Stone & Moore, 2014). Here, we evaluated this theory neurophysiologically using EEG and simultaneous speech-intelligibility measurements. We found that the neural envelope-domain signal-to-noise ratio (SNR) in target-speech encoding, which is shaped by modulations in interfering sounds, predicts intelligibility over a variety of realistic listening conditions. This result provides neurophysiological evidence for modulation masking. However, we also found that envelope coding of target speech in the brain is influenced not only by cochlear-level envelopes, but also by faster stimulus fluctuations (temporal fine structure; TFS), which support scene segregation (Darwin, 1997; A. J. Oxenham & Simonson, 2009). These results are consistent with the theory that temporal coherence of sound elements (Elhilali et al., 2009; Singer &

Gray, 1995) across envelopes and/or TFS shapes scene analysis and attentive selection of a target sound.

To further test the temporal coherence theory, we conducted a follow-up study that used a combination of physiologically plausible computational modeling and a psychophysical experiment to measure consonant categorization across diverse noises and distortions (Chapter 4). Rather than limiting our analyses to overall speech intelligibility—as had been done in prior studies of scene analysis—we analyzed consonant confusion patterns (Miller & Nicely, 1955). Consonant confusions provide a more detailed characterization of error patterns in speech categorization compared to percent-correct scores, and can thus better constrain models of scene analysis. We tested whether modulation masking within individual frequency channels (Jørgensen et al., 2013; Relaño-Iborra et al., 2016) is sufficient to predict consonant confusions across the tested conditions, or if model predictions are improved by the addition of across-channel temporal-coherence processing (Elhilali et al., 2009) that accounts for interference from masker elements that are temporally coherent with target elements but in different frequency channels. Our results provide independent evidence for the role of temporal-coherence processing in scene analysis and speech perception. Moreover, they address the important question of whether this processing may start earlier in the auditory pathway than previously thought (Teki et al., 2013), such as the cochlear nucleus where physiological correlates of across-channel comodulation masking release (CMR) are apparent (Pressnitzer et al., 2001).

Another important problem in the study of speech perception in everyday complex environments is to understand the relative contributions of different acoustics cues in transmitting speech content. The classic view in the literature is that there is a dichotomy in auditory perception where envelopes are thought to convey most speech content and to be sufficient to understand speech in quiet (given adequate frequency resolution; Shannon et al., 1995), whereas TFS is thought to convey other sound attributes such as fundamental frequency (B. C. Moore et al., 2006). However, whether TFS can convey speech content in background noise and when redundant envelope cues are available was poorly understood. Chapter 5 describes our psychophysical study to address this important gap. In this study, we measured consonant categorization in ecologically relevant multi-talker babble for stimuli with intact and degraded

TFS. Confusion patterns in consonant categorization revealed that TFS conveys voicing information beyond what is conveyed by envelopes for intact speech in everyday listening environments with multiple competing talkers. This result suggests that in addition to influencing speech intelligibility in noise indirectly by aiding source segregation, TFS can also influence intelligibility directly by conveying phonetic content.

Chapter 6 concludes this dissertation with a summary of the main findings and how they advance the field, a brief discussion of the significance of this body of work, and some future research directions.

2. ELECTROENCEPHALOGRAPHIC SIGNATURES OF THE NEURAL REPRESENTATION OF SPEECH DURING SELECTIVE ATTENTION

Abstract[1]

The ability to selectively attend to speech in the presence of other competing talkers is critical for everyday communication; yet the neural mechanisms facilitating this process are poorly understood. Here, we use electroencephalography (EEG) to study how a mixture of two speech streams is represented in the brain as subjects attend to one stream or the other. To characterize the speech-EEG relationships and how they are modulated by attention, we estimate the statistical association between each canonical EEG frequency band (delta, theta, alpha, beta, low-gamma, and high-gamma) and the envelope of each of ten different frequency bands in the input speech. Consistent with previous literature, we find that low-frequency (delta and theta) bands show greater speech-EEG coherence when the speech stream is attended compared to when it is ignored. We also find that the envelope of the low-gamma band shows a similar attention effect, a result not previously reported with EEG. This is consistent with the prevailing theory that neural dynamics in the gamma range are important for attention-dependent routing of information in cortical circuits. In addition, we also find that the greatest attention-dependent increases in speech-EEG coherence are seen in the mid-frequency acoustic bands (0.5–3 kHz) of input speech and the temporal-parietal EEG sensors. Finally, we find individual differences in the following: (1) the specific set of speech-EEG associations that are the strongest, (2) the EEG and speech features that are the most informative about attentional focus, and (3) the overall magnitude of attentional enhancement of speech-EEG coherence.

2.1 Introduction

Most of us take for granted our ability to understand speech amidst the cacophony we encounter every day (Cherry, 1953), an ability that is unparalleled by machine algo-

[1]↑This chapter was published following peer review in eNeuro (Viswanathan et al., 2019).

rithms (Loizou, 2013). Yet, 3–5% of children and approximately one in five adults find communicating in noisy social situations extremely challenging (Chermak & Musiek, 1997; Lin et al., 2011), including some listeners who have clinically normal or near-normal thresholds (Kumar et al., 2007). The brain mechanisms that support this auditory "selective attention" process are poorly understood. Identifying correlates of how speech is represented in the brain during selective attention would give us insight into the mechanisms of this process, and how it fails in different clinical populations. Here, we use electroencephalography (EEG) to probe how attended and ignored speech streams in a sound mixture are represented in the brain. Specifically, our goal was to characterize which acoustic features of the speech streams are related to which features of the EEG response, and how such relationships differ for attended and ignored streams.

Neurophysiological experiments using EEG and MEG (magnetoencephalography) show that brain rhythms are intimately associated with sensory processing (Buzsáki & Draguhn, 2004). Electrophysiological studies and computational models suggest that gamma rhythms (30–90 Hz) support the formation of cell assemblies (Cannon et al., 2014). Such assemblies likely mediate stimulus competition and attentional selection of task-relevant representations (Börgers et al., 2008). In contrast, delta (1–3 Hz) and theta (3–7 Hz) oscillations may reflect synchronous interactions between assemblies (White et al., 2000). Strikingly, speech also has spectro-temporal features that are quasiperiodic over similar time scales. Perceptually, the energy envelopes of different frequencies spanning the hearing range carry important information about speech content (Elliott & Theunissen, 2009; Shannon et al., 1995). Importantly, the time scales of phonemic, syllabic, and phrase/sentence level rhythmic fluctuations in speech parallel the EEG gamma, theta, and delta frequencies, respectively. This has led researchers to speculate that the canonical cortical network oscillations are involved in the processing of speech sounds (Doelling et al., 2014; Giraud & Poeppel, 2012). For speech in isolation, brain oscillations phase lock to the speech fluctuations, or more precisely, the fluctuations conveyed at the output of cochlear processing of speech sounds (Ghitza et al., 2012; Gross et al., 2013). It has been suggested that the temporal match between inherent cortical network oscillations and the natural fluctuations in communication sounds may help

the listener parse input speech (Ghitza & Greenberg, 2009; Gross et al., 2013; Luo & Poeppel, 2007).

Fundamental to our understanding of everyday communication is the question of how the neural computations generating brain oscillations relate to the perceptual processes of scene segregation and attentional selection (B. Shinn-Cunningham, 2008). EEG/MEG studies show that when a *mixture* of speech sources is presented, low-frequency cortical responses (matching canonical delta and theta bands) preferentially track the temporal envelopes of attended speech compared to simultaneously presented ignored speech (Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). Similarly, electrocorticography (ECoG) studies show that the power of brain oscillations in the high-gamma (70–150 Hz) band preferentially phase locks to attended speech more than ignored speech (Golumbic et al., 2013; Mesgarani & Chang, 2012). While non-invasive studies have focused on low-frequency portions of the EEG, invasive studies have focused on the high-frequency bands. To the best of our knowledge, no non-invasive studies to date have reported how the full complement of canonical brain oscillations track speech sounds in a mixture of competing sources, when attention is selectively directed to one source stream.

Here, we systematically study how brain oscillations in each of the canonical frequency bands are related to speech fluctuations, comparing when the speech is attended versus when it is ignored. Specifically, we analyze EEG data recorded during a realistic selective attention task, and replicate previous findings that low-frequency EEG bands (in the delta and theta range) show enhanced synchrony with a speech stream when it is attended compared to when it is ignored. In addition, we find that the envelope of the low-gamma EEG band also shows enhanced synchrony with the target speech. Finally, we observe individual differences in the strength and pattern of attention effects. We discuss the implications of our findings for basic neuroscience, and their potential for informing brain-computer interface (BCI) applications such as EEG-guided hearing aids (Fiedler et al., 2017; Fuglsang et al., 2017; J. O'Sullivan et al., 2017; Van Eyndhoven et al., 2017).

2.2 Materials and Methods

2.2.1 Participants

Data was collected from twelve human subjects (six female), aged 23–41 years, recruited from the Boston University community. All subjects had pure-tone hearing thresholds better than 20 dB hearing level (HL) in both ears at standard audiometric frequencies between 250 Hz and 8 kHz. Subjects provided informed consent in accordance with protocols established at Boston University. Of the twelve subjects who participated, data from two were excluded from analysis for reasons described below.

2.2.2 Experimental design

In each listening block, two running speech streams (narrated whole stories), one spoken by a male and the other by a female (from one of "The Moth" storytelling events, New York), were presented simultaneously to the subject. The stories were each lateralized using interaural time delays (ITDs). The root-mean-square intensities of the male and female speech streams were equalized dynamically using a sliding window of length 2 s. A total of four stories were used in the experiment. Each subject performed four blocks; at the beginning of each block, subjects were verbally instructed to attend to one of the two talkers throughout that block. Subjects were also asked to stay still with their eyes blinking naturally during the experiment; however, their eye gaze was not restricted. EEG was measured simultaneously with the behavioral task in each block. The individual stories were about 9-12 min long; thus the blocks were also 9-12 min long each.

At the end of each block, subjects were given a quiz on the attended story. If a subject answered at least 90% of the quiz questions correctly, they passed the quiz. Based on the responses to the quiz, one subject was excluded due to their inability to accurately recall details of the attended story. All of the remaining eleven subjects were able to recount details of the attended story accurately, and reported being largely unaware of the details of the other (ignored) story.

All the subjects were presented with the same set of speech stories. However, which story was attended in a given block was varied randomly across listeners, with the constraint that each listener heard every story once when it was to be ignored and once when it was to be attended. This design allowed us to directly compare attended and ignored conditions for the same acoustic input to the subject. Furthermore, the two presentations of each speech story (once when the story was to be attended, and the other when it was to be ignored) were separated by at least one block for every subject.

2.2.3 Data acquisition

A personal desktop computer controlled all aspects of the experiment, including triggering sound delivery and storing data. Special-purpose sound-control hardware (System 3 real-time signal processing system, including digital-to-analog conversion and amplification; Tucker Davis Technologies) presented audio through insert earphones (ER-1; Etymotic) coupled to foam ear tips. The earphones were custom shielded using a combination of metallic tape and metal techflex to attenuate electromagnetic artifacts. The absence of measurable electromagnetic artifact was verified by running intense click stimuli through the transducers with the transducers positioned in the same location relative to the EEG cap as actual measurements, but with foam tips left outside the ear. All audio signals were digitized at a sampling rate of 24.414 kHz. The EEG signals were recorded at a sampling rate of 2.048 kHz using a BioSemi ActiveTwo system. Recordings were done with 32 cephalic electrodes, additional electrodes on the earlobes, and a bipolar pair of electrodes adjacent to the outer left and right canthi to measure saccadic eye movements.

2.2.4 Data preprocessing

The EEG signals were re-referenced to the average of all the channels. The signal-space projection method was used to construct spatial filters to remove eye blink and saccade artifacts (Uusitalo & Ilmoniemi, 1997). The broadband EEG was then band-pass filtered between 1 Hz and 120 Hz for further analysis. For computing associations between speech and EEG, the EEG data were segmented into 5-s-long epochs. Epochs with movement

artifacts were identified as those with a peak-to-peak swing that exceeded twenty median absolute deviations compared to the median epoch. All such epochs were rejected to eliminate movement artifacts. Of the eleven subjects who successfully passed our behavioral screening, one subject was excluded because more than 20% of their EEG data was contaminated by movement artifacts. The data from the remaining ten subjects were used in all further analyses.

2.2.5 Estimating speech-EEG associations

Our goal was to understand the relationships between features of input speech and EEG responses, and how these relationships vary depending on whether speech is attended to or ignored. For the speech features, we considered envelope fluctuations in ten different frequency bands. For the EEG features, we considered different EEG bands corresponding to the canonical cortical rhythms, and different scalp locations of the 32-channel EEG recording. The rationale for the choice of these speech and EEG features, along with the procedure for extracting them are described below.

The auditory periphery can be approximated as a filter bank that decomposes speech into different frequency bands; the envelope at the output of each cochlear filter is conveyed to the brain by auditory-nerve fibers tuned to the corresponding frequency band (Khanna & Leonard, 1982; Smith et al., 2002). We used a bank of ten gammatone filters that mimic cochlear frequency selectivity (Slaney et al., 1993), with center frequencies spanning 100–8533 Hz. The filters were spaced roughly logarithmically, such that their center frequencies had best places that are spaced uniformly along the length of the cochlea according to an established place-frequency map (Greenwood, 1990). The amplitude envelope at the output of each filter, extracted using the Hilbert transform, was treated as a distinct speech feature. For the speech signals used in our experiment, the envelopes at the different filters were not strongly correlated. In analyzing the speech envelopes extracted from different bands, we found that the variance explained in the envelope of one band by any other band was about 8% or less (estimated by calculating squared coherence between speech envelopes). This suggests that

the speech envelopes in the ten different cochlear bands provide somewhat complementary speech information.

Previous EEG/MEG studies show that cortical responses to speech mixtures preferentially track the spectro-temporal features of the attended speech during selective listening (Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). Specifically, the low-frequency speech envelope elicits phase-locked EEG responses at corresponding frequencies (delta band: 1–3 Hz, and theta band: 3–7 Hz). Furthermore, ECoG studies show that the slowly varying envelopes of high-frequency neural responses (high-gamma band: $>$ 70 Hz) also track the attended speech (Golumbic et al., 2013; Mesgarani & Chang, 2012). Thus, we systematically studied the relationship between speech and the corresponding neural responses by decomposing the EEG signal from each of the 32 channels into six canonical frequency bands (delta: 1–3 Hz, theta: 3–7 Hz, alpha: 7–15 Hz, beta: 13–30 Hz, low-gamma: 30–70 Hz, and high-gamma: 70–120 Hz; Buzsáki & Draguhn, 2004). In the delta, theta, alpha, and beta bands, the filtered EEG signal was treated as a feature. On the other hand, for the higher-frequency gamma bands, we were motivated by the results from the ECoG studies to extract and use the amplitude envelopes in those bands instead (discarding phase information). For the alpha and beta bands, we considered the amplitude envelopes of those bands as additional features separately from the filtered EEG. This choice was motivated by the finding that alpha power fluctuates coherently with the attended stimulus (Wöstmann et al., 2016), and that beta-band power fluctuates in a task-specific way across many cognitive and motor tasks (Engel & Fries, 2010). To extract the envelopes of the alpha, beta, low-gamma, and high-gamma bands, we used the Hilbert transform. Overall, a total of 256 EEG features were considered: the filtered EEG in the delta, theta, alpha, and beta bands, and the envelopes of alpha, beta, low-gamma, and high-gamma bands, across the 32 EEG channels. Throughout this report, we will use the term EEG bands to denote the EEG signals or envelope signals in different frequency bands. Thus, the analyzed EEG bands consist of the delta, theta, alpha, and beta bands, and the amplitude envelopes of alpha, beta, low-gamma, and high-gamma bands.

Spectral coherence (also simply referred to as coherence) was chosen as the measure of statistical dependence between the speech and EEG signals. High coherence indicates a consistent phase relationship between signals (Dobie & Wilson, 1989; Hannan, 1970; Thomson,

1982). Moreover, when artifactual trials are excluded, spectral coherence is likely to be more sensitive than the phase-locking value (Lachaux et al., 1999), as coherence computation assigns greater weights to trials with larger signal amplitude (Dobie & Wilson, 1994). A multi-taper approach (with five tapers, resulting in a frequency resolution of 1.2 Hz) was used to estimate the spectral coherence between each speech and EEG feature from the 5-s-long epochs segmented from the raw EEG data (Slepian, 1978; Thomson, 1982). A total of 108 epochs were used in the computation of each coherence spectrum. The multi-taper estimate minimizes spectral leakage (i.e., reduces mixing of information between far-away frequencies) for any given spectral resolution, and is calculated from the Fourier representations of two signals $X(f)$ and $Y(f)$ as follows:

$$C_{XY}(f) = \frac{S_{XY}(f)}{\sqrt{S_{XX}(f)S_{YY}(f)}} \tag{2.1}$$

where

$$S_{XY}(f) = \frac{1}{K_{tapers}N_{epochs}} \sum_{k=1}^{K_{tapers}} \left| \sum_{n=1}^{N_{epochs}} X_{kn}(f)Y_{kn}^*(f) \right| \tag{2.2}$$

$$S_{XX}(f) = \frac{1}{K_{tapers}N_{epochs}} \sum_{k=1}^{K_{tapers}} \left| \sum_{n=1}^{N_{epochs}} X_{kn}(f)X_{kn}^*(f) \right| \tag{2.3}$$

$$S_{YY}(f) = \frac{1}{K_{tapers}N_{epochs}} \sum_{k=1}^{K_{tapers}} \left| \sum_{n=1}^{N_{epochs}} Y_{kn}(f)Y_{kn}^*(f) \right| \tag{2.4}$$

For each pair of speech and EEG features, a single measure of coherence was obtained by averaging the coherence spectrum obtained via the multi-taper estimation procedure as follows: For the regular coherence in the delta, theta, alpha, and beta bands, the coherence values were averaged over the canonical frequency ranges of the respective bands (i.e., 1–3 Hz for delta, 3–7 Hz for theta, 7–15 Hz for alpha, and 13–30 Hz for beta). For the envelope coherences of the alpha, beta, low-gamma, and high-gamma bands, the averaging was performed over envelope frequencies of 1–7 Hz (corresponding to the frequency range at which previous studies report phase locking between the speech envelope and the envelope of the neural

response in the gamma band; Gross et al., 2013). Figure 2.1 summarizes the steps used to extract speech and EEG features, and to estimate the coherence between them.

In this way, we characterized the relationships between different features of input speech (i.e., the speech envelopes in different cochlear bands) and different features of the EEG response (each of which corresponds to a specific EEG band and channel). In particular, we characterized these relationships in an attention-specific manner, i.e., both when the input speech was attended and also when it was ignored. This allowed us to examine the effects of attention on the speech-EEG relationships separately in different EEG bands, different scalp locations, and different speech bands, and also to characterize individual differences in the attentional enhancement of speech-EEG associations. Further methodological details are presented alongside each result description as needed.

2.2.6 Visualizing individual subject results as a network graph

The full set of speech-EEG relationships is a high-dimensional data set (with EEG bands, scalp channels, and speech bands constituting the different dimensions) that can be conceived of as a network. In many domains, bipartite graphs have been successfully used to represent and characterize the complex pattern of associations between two types of variables ("nodes") in a relational network (e.g., group-member relationships in a social network (T. P. Wilson, 1982), genotype-phenotype relationships in a biological network (Goh & Choi, 2012), etc.). To visualize the relationships between all pairs of speech and EEG features simultaneously in each individual subject, we constructed bipartite graphs with the ten speech features forming the nodes in one partition, and the 256 EEG features (32 scalp locations $\times$ eight EEG bands) forming the nodes in the other. An edge (i.e., connection) between a speech feature and an EEG feature in our bipartite graph construction signifies a statistical dependence between them, such as a significant coherence value. We constructed separate attended and ignored graphs for each individual subject in our study using the following procedure. First, the speech-EEG coherences for each subject were averaged across all speech stories for the attended and ignored conditions separately. Next, edges were drawn between those pairs of speech-EEG features whose coherence values met a particular threshold. The resulting graph

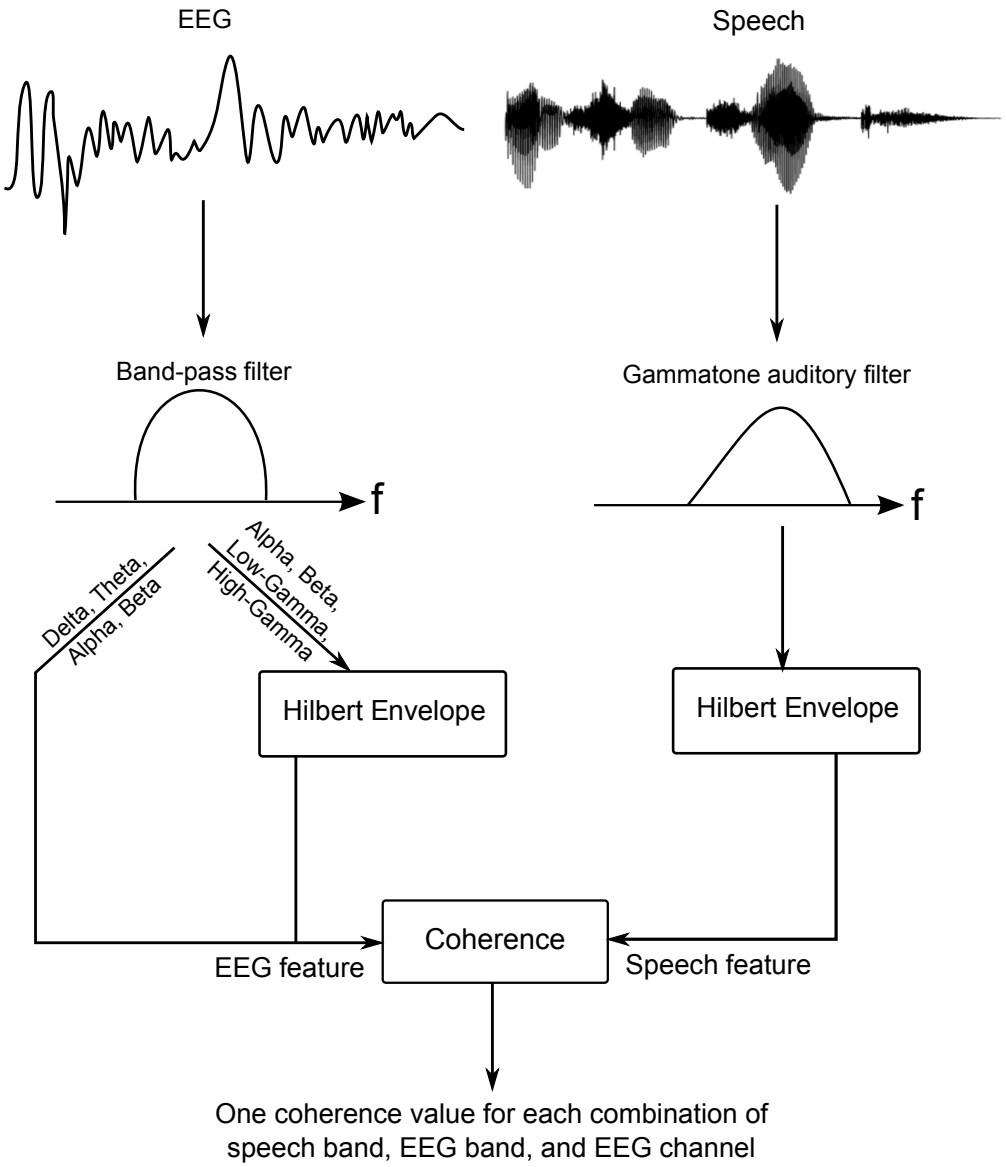

Figure 2.1. : **Illustration of the steps used to extract speech and EEG features and to estimate the association between them.** The speech signal is passed through a gammatone filter bank simulating cochlear processing, and the envelope at the output of each filter (i.e., the envelope of each speech band) is extracted as a speech feature. Similarly, different bands of the EEG and different sensor channels together form the different EEG features. For the lower-frequency bands (delta and theta), the EEG signals are used as is. For the alpha and beta bands, both the signals in those bands, and their envelopes are extracted as separate features. For the higher-frequency gamma bands, only the envelopes of the EEG signals in those bands are considered. These EEG features are then compared with the speech features using spectral coherence.

representations of speech-EEG relationships were visualized to qualitatively compare the two attention conditions and different individuals. To quantitatively compare attended and ignored graphs, we computed the average difference in the number of graph edges between the attended and ignored conditions, for different coherence thresholds. The results were compared with permutation-based null distributions to obtain p-values, as described in Section 2.2.7.

The bipartite graph formulation also has the advantage that the complex set of dependencies between speech and EEG, and how those dependencies are modulated by attention, can be summarized using rigorous metrics developed in network science. Accordingly, we take advantage of network summary measures that use the entire network structure to find those speech and EEG features that best capture attentional focus in an individual-specific manner. This is done with the view of informing attention-decoding applications as to which EEG and stimulus features may provide the best decoding performance at the individual level. For this, we first computed the differential ("attended - ignored") coherence for each speech-EEG pair for each individual subject (but averaged across speech stories). For each individual, the full set of speech and EEG features and their associated differential coherences can be represented as a weighted "differential" speech-EEG bipartite graph, with the differential coherence associated with each speech-EEG pair forming the edge weight for that pair. Note that this weighted graph representation of the differential coherences contrasts with the unweighted graph representations for the attended and ignored conditions that were described previously. For the attended and ignored graphs, we had used a coherence threshold to define an edge. On the other hand, to obtain the differential graphs, we did not use any thresholding procedure. Instead, the differential coherence values across all speech-EEG feature pairs were retained, and used to define graph edge weights. Finally, to find those speech and EEG features that are the most informative about an individual's attentional focus, we computed the eigenvector-based graph centrality measure for each speech and EEG feature in every individual's differential graph. For a discussion on the notion of network centrality, and how it may be computed in bipartite graphs to identify the most informative nodes in the network, see Faust, 1997.

2.2.7 Statistical analysis

The primary question that this study is concerned with is whether the neural representation of speech is modulated by attention. For this, the null hypothesis is that attention does not alter speech-EEG relationships. We used a non-parametric *within-subjects* randomization procedure to perform statistical inference against this null hypothesis. This procedure was applied to two separate analyses, as described below.

For the analysis performed to characterize which EEG bands show attention-dependent changes in coherence with speech (results in Figure 2.3 Panel A), the specific null is that the speech-EEG coherence in each of the EEG bands is the same on average for the attended and ignored conditions. Thus, under the null hypothesis, the attended and ignored conditions are equivalent and the labels "attended" and "ignored" can be swapped randomly to generate examples of coherence differences that would be observed under the null hypothesis. Note that our experimental design of randomly assigning which of the two stories in each block is attended provides the necessary exchangeability criterion, justifying the permutation procedure (Nichols & Holmes, 2002). That is, every permutation of the order in which the stimuli and attention conditions occurred was equally likely to occur during data acquisition. Thus, under the null hypothesis, the condition labels corresponding to the measurements can be randomly permuted. To generate a single realization from the null distribution, a random sign was assigned to the coherence difference between the attended and ignored conditions for each subject and speech story, then the results were averaged across subjects and stories. This procedure was repeated with 500,000 distinct randomizations to generate the full null distribution for the average coherence difference. A separate null distribution was generated for each of the eight EEG bands using band-specific data. For each band, the corresponding null distribution was used to assign a p-value to the observed average coherence difference obtained with the correct labels. Finally, to correct for multiple comparisons across the eight EEG bands, the conservative Bonferroni procedure was used. In addition to being used to obtain p-values, the null distributions were also used to express each individual's coherence-difference values as a z-score, which provided an easy-to-interpret quantification of effect sizes. We used a similar permutation procedure to generate noise floors for computing

the z-scores shown in Figure 2.3 Panels B and C, and in the differential scalp map of Figure 2.4. A separate noise floor was generated for each speech band in Figure 2.3 Panel B, for each pixel (corresponding to a distinct speech band and EEG band) in Figure 2.3 Panel C, and for each electrode in Figure 2.4.

For the analysis on the number of edges in the graph representation of speech-EEG coherence (Figure 2.7), a similar permutation procedure was used. Here, the specific null hypothesis is that the graph has the same number of edges in the attended and ignored conditions on average. Thus, for each subject, a random sign was assigned to the difference in the number of edges between the attended and ignored conditions, then the result was averaged over subjects. This randomization procedure was repeated 500,000 times to generate the full null distribution. A separate null distribution was generated for each of the coherence thresholds shown in Figure 2.7. The observed average differences in the number of edges between the correctly labeled attended and ignored conditions were then compared to the corresponding null distributions to assign p-values.

The noise floor parameters used for computing the z-scores shown in the attended and ignored scalp maps of Figure 2.4 were theoretically derived. This was done by using the mean and variance expressions for multi-taper coherence estimates provided in Bokil et al., 2007, and adjusting the variance parameter to account for pooling across EEG frequencies and speech bands.

2.2.8 Software accessibility

Stimulus presentation was controlled using custom MATLAB (The MathWorks, Inc., Natick, MA) routines. EEG data preprocessing was performed using the open-source software tools MNE-Python (Gramfort et al., 2013; Gramfort et al., 2014) and SNAPsoftware (Bharadwaj, 2018). All further analyses were performed using custom software in Python (Python Software Foundation, www.python.org). Network visualizations were created using the SAND package (Kolaczyk & Csárdi, 2014) in R (R Core Team, www.R-project.org). Copies of all custom code can be obtained from the authors.

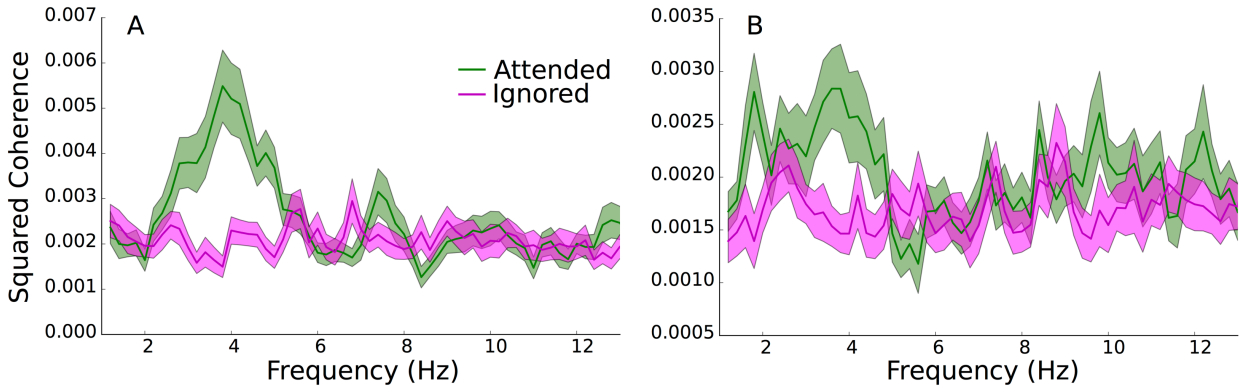

Figure 2.2. : **Illustration of the effect of attention on the average speech-EEG magnitude squared coherence spectra, for (A) the envelope of the 1014 Hz speech band, and the low-frequency portions (overlapping with the delta and theta bands) of EEG channel C3, and for (B) the envelope of the 3733 Hz speech band, and the envelope of the low-gamma band of EEG channel CP1.** Note that the y-axis ranges differ between panels A and B. The shaded regions indicate values within one standard error around the mean. The delta- and theta-band EEG responses (Panel A), and the low-gamma-band EEG envelope fluctuations (Panel B) selectively track features of the attended speech over the ignored speech.

2.3 Results

Figure 2.2 shows magnitude squared coherence spectra (averaged over subjects and speech stories) for two example speech-EEG pairings: the envelope of the 1014 Hz speech band and the low-frequency EEG in sensor C3 (Panel A), and the envelope of the 3733 Hz speech band and the envelope of the low-gamma EEG band in sensor CP1 (Panel B). The coherence in the attended condition is greater than that in the ignored condition in the 2–5 Hz frequency range (overlapping with the delta and theta bands) for the example in Panel A. The slow envelopes of the low-gamma band also preferentially track attended speech in the 2–5 Hz frequency range (Panel B).

As described in Section 2.2.5, the coherence spectrum for each pair of speech-EEG features was averaged across frequencies to obtain a single coherence value for that feature pair; this was done separately for the attended and ignored conditions. One key question we wished to answer was which EEG bands showed the greatest attention effects. To address this question, we averaged the differential coherences ("attended - ignored") for each EEG band across all speech bands and across the 32 EEG channels. The results obtained from this

analysis are shown in Figure 2.3 Panel A. For each EEG band, we statistically tested whether the coherence increase in the attended condition was significant using the permutation procedure described previously. To correct for multiple comparisons across the eight EEG bands that were considered, we used a Bonferroni correction with a familywise error rate of 0.05. Thus, for each of the eight tests, only p-values less than 0.05/8 were considered to be statistically significant. Based on the statistical tests, we find that both delta and theta bands of the EEG show greater coherence with a speech stream when that stream is attended compared to when it is ignored (i.e., a positive "attended - ignored" difference). This replicates previously reported results (Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). Aside from the attention-dependent increase in low-frequency coherence, we also observe that the envelope of the low-gamma band shows greater coherence to speech in the attended condition. The preferential synchrony of gamma-band envelopes with attended speech has previously been reported only in invasive recordings (Golumbic et al., 2013; Mesgarani & Chang, 2012). For speech in isolation, some non-invasive studies have found gamma-band envelopes to be synchronous with input speech (Gross et al., 2013); however, to the best of our knowledge an attention-dependent increase of this coherence has previously not been reported with non-invasive recordings.

In addition to identifying the EEG bands that showed the greatest attention effects, we were also interested in characterizing which speech bands contribute most to attention-dependent increases in coherence. To address this question, we averaged the differential coherences for each speech band across the 32 scalp locations and across all EEG bands. This yielded a profile of attention-dependent increases in coherence across the ten different speech bands. The results are shown in Figure 2.3 Panel B. The strongest attention effects appear to occur in the 0.5–3 kHz range, which contains spectro-temporal speech features (formants and formant transitions) that convey many vowel and certain consonant cues (Gold & Morgan, 2002), and is also the range thought to be the most important for speech intelligibility (Kryter, 1962).

To examine whether the attention effects for different speech bands varied with the EEG bands that they were paired with, we visualized the differential coherence for the full matrix of speech bands versus EEG bands, averaged across EEG channels. The results are shown in

Figure 2.3 Panel C. While the 0.5–3 kHz speech frequency range shows hot spots in the delta, theta, and low-gamma EEG bands, the lower-frequency speech bands (e.g., 200 Hz) show a hot spot only in the theta range corresponding to the syllabic rate. This could be because the pitch conveyed by the resolved harmonics of the syllabic voicing may be an important cue based on which attention is directed.

We also wished to find the EEG scalp locations that show the greatest coherence and attention effects. To address this question, we averaged the coherence values across the ten speech bands, and the delta, theta, and low-gamma EEG bands (i.e., the bands showing significant attention effects in Figure 2.3 Panel A). The results are plotted as a topographic map of coherence values (i.e., one value for each of the 32 scalp locations) for the attended, ignored, and differential conditions, respectively, in Figure 2.4. The spatial profiles are hard to distinguish between the attended and ignored maps; however, note that the coherences are larger in the attended condition than the ignored, on an absolute scale. The differential map quantifies these differences across the scalp. Temporal-parietal regions appear to show the largest coherence differences between the attended and ignored conditions; however, this pattern is not symmetric between the hemispheres. This result is consistent with previous studies that found that areas such as the superior temporal gyrus and the inferior parietal lobule contribute to attention effects (Golumbic et al., 2013). In addition to plotting scalp maps averaged across EEG bands, we also looked at band-specific scalp maps for the differential condition. However, the spatial patterns in those maps were not easily interpretable, and are hence not shown here. Because we only used 32 channels, a detailed exploration of which brain sources contribute to the observed differential coherences cannot be done with our data. This should be a focus of future studies.

The results shown so far were mainly concerned with attention-dependent coherences averaged across different sets of speech and EEG features (i.e., across speech bands, and/or EEG bands, and/or scalp locations). In addition to this, we also constructed speech-EEG bipartite graphs for each individual to examine the full set of coherence values corresponding to all pairs of speech-EEG features simultaneously. Figure 2.5 shows attended and ignored graphs (averaged over speech stories) for all individual subjects in our study. In this figure, each square denotes a speech feature, and each circle denotes an EEG feature. An edge

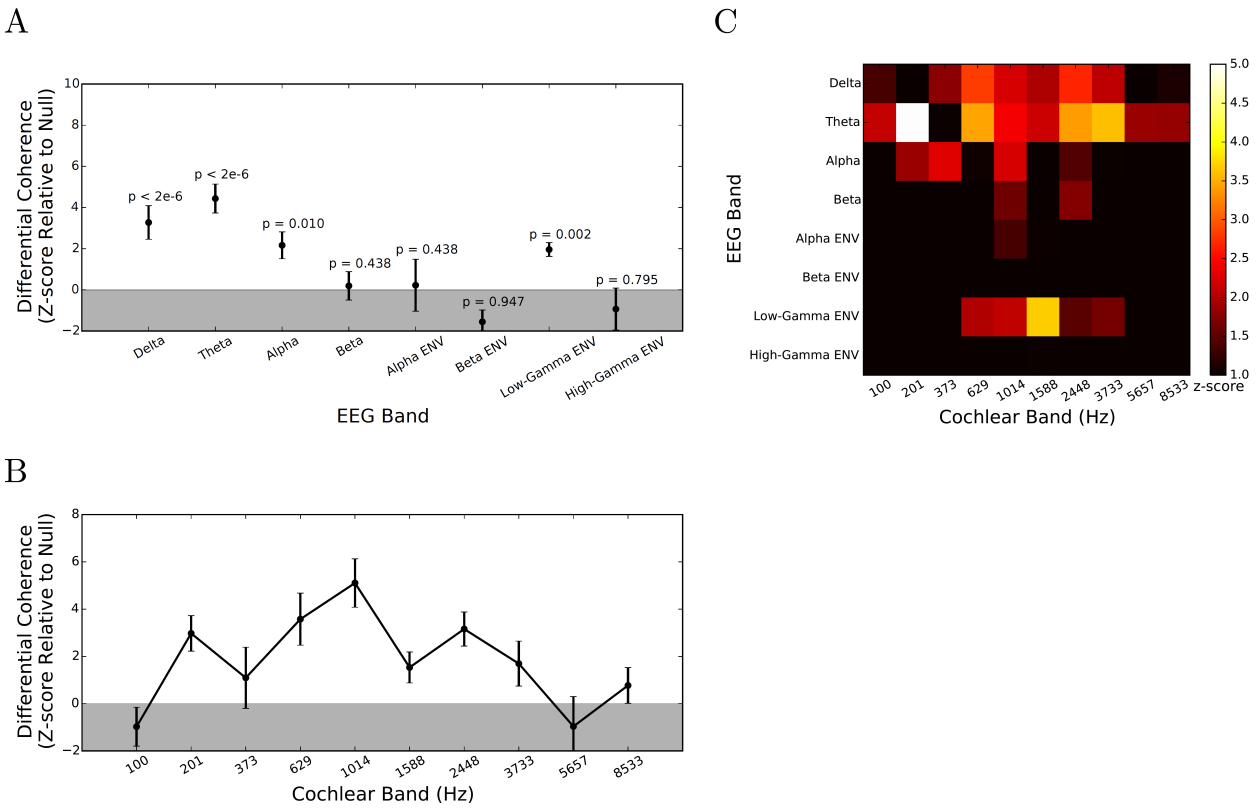

Figure 2.3. : **Differential effects of attention on speech-EEG coherences in different EEG bands (Panel A), different speech bands (Panel B), and the full matrix of EEG bands versus speech bands (Panel C). Panel A** shows the differential ("attended - ignored") coherence averaged across speech bands and EEG channels (shown as a z-score) for each of the EEG bands. Uncorrected p-values obtained from the permutation test are displayed for the different EEG bands. When a Bonferroni-corrected p-value threshold of $0.05/8 = 0.006$ is applied to each band, we find that the delta and theta bands show significantly higher coherence with speech when it is attended compared to when it is ignored. In addition, we also find that the envelope of the low-gamma band shows greater coherence with attended versus ignored speech. **Panel B** shows the differential coherence averaged across all EEG bands and EEG channels (shown as a z-score) for each input speech band. The strongest attention effects appear to occur in the 0.5–3 kHz range, which contains spectro-temporal speech features (formants and formant transitions) that convey many vowel and certain consonant cues, and is also the range thought to be the most important for speech intelligibility. In **Panel C**, the differential coherence averaged across EEG channels is shown as a z-score for each EEG band and speech band for completeness. While the 0.5–3 kHz speech frequency range shows hot spots in the delta, theta, and low-gamma EEG bands, the lower-frequency speech bands (e.g., 200 Hz) show a hot spot only in the theta range corresponding to the syllabic rate. This could be because the pitch conveyed by the resolved harmonics of the syllabic voicing may be an important cue based on which attention is directed. In all three panels, z-scores shown are averaged across speech stories and individual subjects, with error bars representing the standard error.

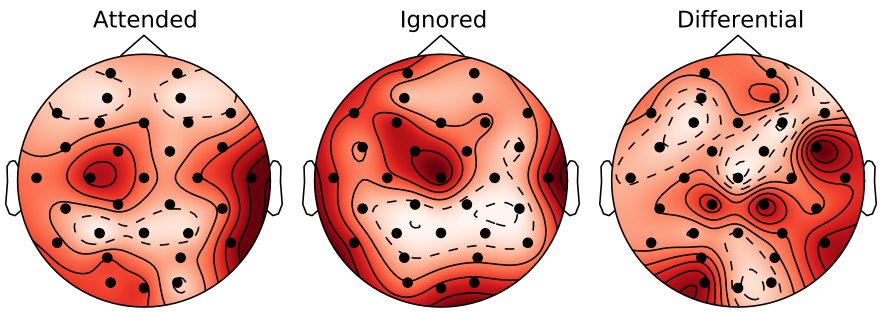

Figure 2.4. : **Scalp maps showing the average coherence (shown as a z-score) at each of the different EEG electrodes in the attended, ignored, and differential conditions.** To obtain the scalp maps, the speech-EEG coherence values were averaged across the delta, theta, and low-gamma EEG bands (i.e., the bands showing significant attention effects in Figure 2.3 Panel A), and all speech bands, and expressed as a z-score. The intensity shown at each electrode is the mean of the z-score across speech stories and individual subjects. Note that the scalp maps are scaled to their respective minimum and maximum z-score values, so as to best show the spatial patterns. The spatial profiles are hard to distinguish between the attended and ignored maps; however, note that the coherences are larger in the attended condition than the ignored, on an absolute scale. The differential map shown in the right column quantifies these differences across the scalp. Temporal-parietal regions appear to show the largest coherence differences between the attended and ignored conditions; however, this pattern is not symmetric between the hemispheres.

is shown connecting a pair of speech-EEG features if the coherence between them meets a certain threshold. Here, a coherence threshold of 3 standard deviations from the average coherence (pooled across attended and ignored conditions) is arbitrarily chosen, and only edges whose coherence meets that threshold are shown. One pattern that is immediately apparent from Figure 2.5 is that there are many more edges in the attended condition than in the ignored condition for eight of the ten subjects in this study. This suggests that a larger number of speech-EEG feature pairs become coherent when the speech is attended. Also apparent from Figure 2.5 is the fact that the graph structure is variable across subjects. This means that the particular speech-EEG feature pairs that show the greatest coherence values are not the same across subjects. As described in Section 2.2.6, we used the eigenvector centrality measure for bipartite graphs to find those EEG and speech features that are the most informative about an individual's attentional focus. We find that the most central features differ between individuals, as shown in Figure 2.5. This suggests that for applications such as BCIs that aim to decode attention from EEG, individual-specific customization of features might be necessary to obtain optimal decoding performance.

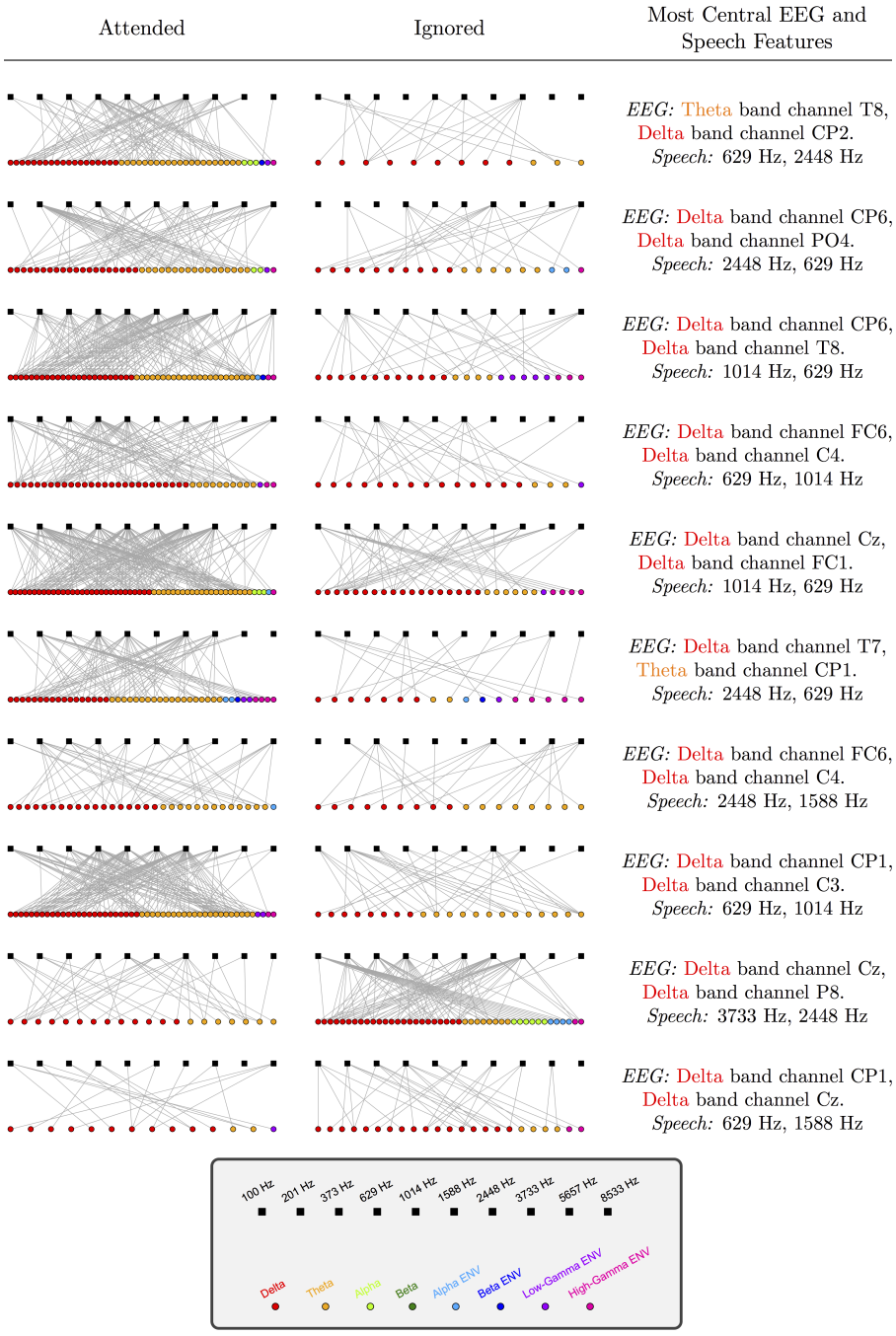

Figure 2.5. : **Graph representation of speech-EEG coherence in the attended and ignored conditions for all individual subjects.** Rows represent different individuals. Squares denote speech features (i.e., the envelopes from the ten speech bands; shown in the order of increasing center frequency). Each circle denotes an EEG feature (i.e., a particular EEG band from a particular scalp location). An edge between a speech and EEG feature indicates that the coherence between them meets a threshold of 3 standard deviations from the mean. Only EEG features with one or more edges that survive the thresholding procedure are shown. Attended graphs exhibit greater number of edges compared to ignored graphs for all but two subjects (see bottom two rows). Additionally, the graph structure is variable across subjects. The top two EEG and speech features that are most informative (as obtained using eigenvector centrality) about an individual's attentional focus also vary across subjects (rightmost column).

Figure 2.6 shows individual differences in the overall magnitude of attentional enhancement of speech-EEG coherences, separately for the delta, theta, and low-gamma EEG bands (i.e., the bands showing significant attention effects in Figure 2.3 Panel A). Here, each individual's "attentional boost" was computed as their percentage change in squared coherence going from the ignored condition to the attended, averaged across the 32 EEG channels, all speech bands, and the different speech stories. This attentional boost metric represents the percentage change in the proportion of EEG signal energy that is correlated with a speech signal, when the speech is attended to versus ignored. The distribution of the attentional boost across individuals is skewed above zero in all three EEG bands, consistent with positive attentional boost in the neural coding of target speech. Furthermore, there is considerable variation across subjects almost uniformly over the range of boosts. Finally, where a particular individual falls relative to the overall distribution is somewhat consistent across the three EEG bands (the rank correlation between the attentional boosts in the delta and theta bands is 0.78, and between the boosts in the delta and low-gamma bands is 0.38).

Although Figure 2.5 is visualized for a particular coherence threshold, the observation that there are many more edges in the attended condition than in the ignored condition did not depend strongly on the choice of threshold. To illustrate this, we quantified the percentage of edges (i.e., coherences that meet a given threshold) for the attended and ignored conditions, for three different threshold values. The results are shown in Figure 2.7. For all three thresholds shown, the number of edges in the attended condition is significantly greater than the number of edges in the ignored condition, which confirms the generality of this result. The p-values for this statistical comparison were obtained using a permutation test as described in Section 2.2.7. While Figure 2.3 showed that specific speech-EEG associations are strengthened by attention, the present result suggests that a greater number of distinct speech-EEG associations are induced by attention.

2.4 Discussion

We systematically studied the attention-dependent relationships between input speech envelopes in different frequency bands and the neural response in different EEG channels

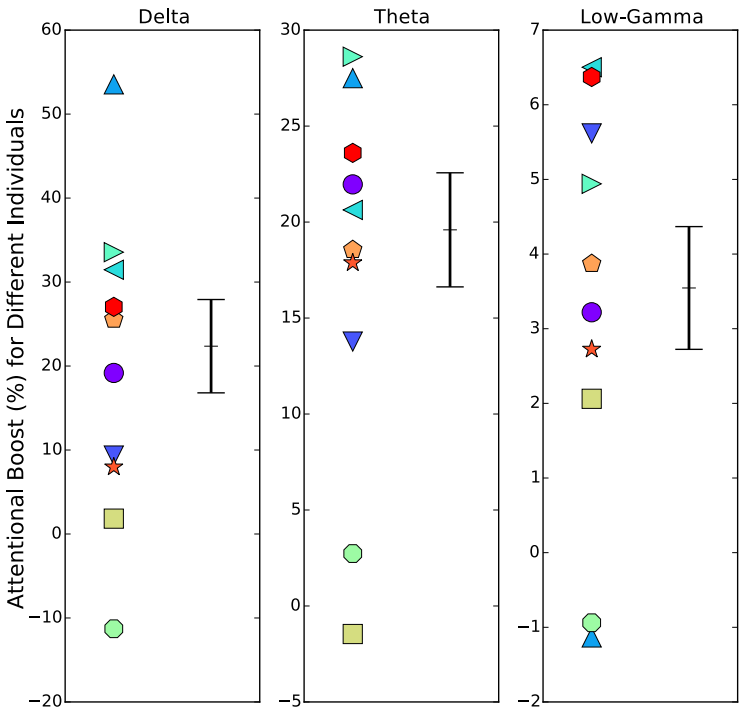

Figure 2.6. : **Individual differences in the overall magnitude of attentional enhancement of speech-EEG coherences in different EEG bands.** Each individual's "attentional boost" in coherence is shown (with an individual-specific marker symbol and color) for the delta, theta, and low-gamma EEG bands (i.e., the bands showing significant attention effects in Figure 2.3 Panel A). The mean and standard error across individuals are also indicated in black. Note that the y-axis ranges differ between the three panels of the figure. The attentional boost was computed as the percentage change in squared coherence going from the ignored condition to the attended, averaged across EEG channels, speech bands, and the different speech stories. The distribution of the attentional boost across individuals is skewed above zero in all three EEG bands, consistent with positive attentional boost in the neural coding of target speech. Furthermore, there is considerable variation across subjects almost uniformly over the range of boosts.

and frequency bands. Importantly, we investigated selective attention effects in all canonical (Buzsáki & Draguhn, 2004) EEG frequency bands simultaneously. In doing so, we found that low-frequency delta- and theta-band EEG showed the strongest attention effects (i.e., the greatest speech-EEG coherence increases for the attended condition compared to the ignored). This result is consistent with the preferential phase locking to attended rather than ignored speech in the delta and theta bands reported in previous EEG/MEG studies (Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). Using stationary masking noise, Ding and Simon, 2013 found that the delta band was the most robust in carrying target information at poorer SNRs (-3 dB and lower), whereas both delta and theta bands were equally robust in

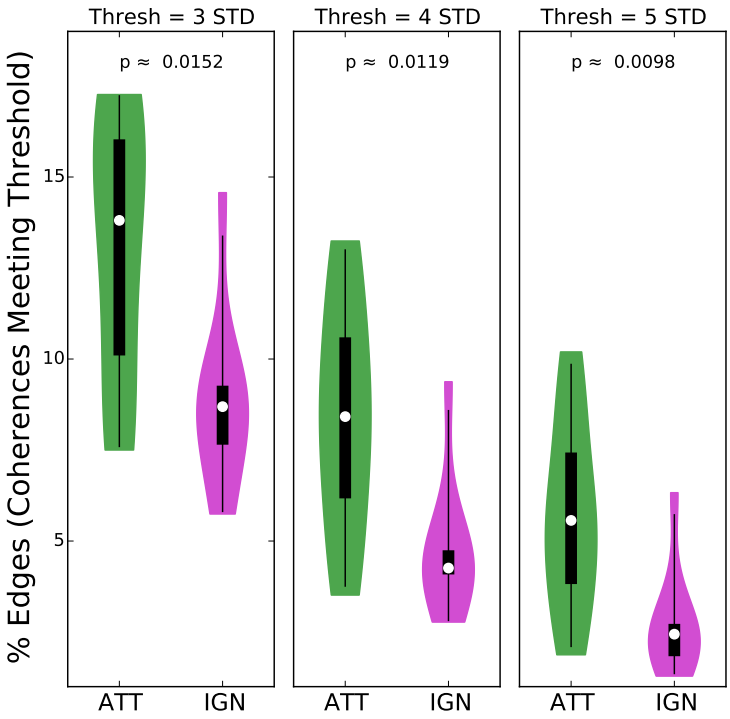

Figure 2.7. : **Percentage of edges (i.e., coherences meeting threshold) in attended (ATT) and ignored (IGN) speech-EEG bipartite graphs, at different coherence thresholds**. The across-subject distribution of the percentage of graph edges is shown as a violin plot, separately for the attended and ignored conditions, and for three different coherence thresholds. In addition, the median (white dot), 50% confidence limits (thick black box), and 95% confidence limits (black whiskers) of each distribution are shown. Across all three threshold values, the number of edges is significantly larger for the attended condition (based on a permutation test; p-values are shown). While Figure 2.3 showed that specific speech-EEG associations are strengthened by attention, the present result suggests that a greater number of distinct speech-EEG associations are induced by attention.

conveying target information at higher SNRs. These findings are consistent with our present results from using a speech masker at 0 dB SNR. One possible factor contributing to the strong delta- and theta-band attention effects is that the power in the acoustic envelope of natural speech is maximal below 8 Hz (corresponding to the prosodic and syllabic rates; Ding et al., 2017). Moreover, in the presence of background noise, the SNR in the envelope domain at the auditory-nerve level is strongest for slow modulation frequencies (Rallapalli & Heinz, 2016). Thus, the strength of the delta- and theta-band effects may be a reflection of the neural computations that take advantage of the high power and SNR in speech at slow envelope frequencies. Yet another possible factor could be that attention mechanisms might be geared towards boosting the representation of those temporal modulations that are the

most important for speech intelligibility; previous studies suggest that modulations below 8 Hz are perhaps the most important (Drullman et al., 1994; Elliott & Theunissen, 2009).

A novel finding of the present study is that the power fluctuations (i.e., envelope) of the low-gamma band of the EEG show significantly higher coherence with the attended speech stream versus the ignored. In contrast to cortical theta-band activity, activity in the gamma band has relatively small amplitude (Pritchard, 1992). This may explain why previous EEG studies have not reported attention effects in the gamma band. Despite the relatively low amplitude and the conservative statistical thresholding that we adopted (i.e., using Bonferroni corrections across EEG bands), we found the low-gamma envelope to fluctuate coherently with the attended speech. This finding supports the view that gamma activity plays an important role in the underlying physiological computations that support selective listening (Ribary, 2005; Tallon-Baudry & Bertrand, 1999; X.-J. Wang, 2010), and demonstrates that non-invasive EEG can be used to measure these effects.

While gamma-band responses have been investigated using EEG/MEG when processing speech streams in isolation, i.e., without competition (Gross et al., 2013), prior non-invasive studies of selective attention focused on the low-frequency portions of the brain signal, which overlap with traditional evoked responses (Ding & Simon, 2012; Luo & Poeppel, 2007; J. A. O'Sullivan et al., 2014). Gamma-band power has previously been shown to fluctuate coherently with the envelope of an attended speech stream in selective attention tasks, but only from invasive (ECoG) recordings (Golumbic et al., 2013; Mesgarani & Chang, 2012). The current results replicate this finding using EEG. However, one discrepancy in the gamma-band findings between the ECoG studies and the present EEG-based study is that the ECoG studies found the high-gamma, rather than the low-gamma band to be important, while we observed no significant effects at high gamma. This may be explained by the fact that ECoG measurements are more spatially specific, reflecting local neural activity rather than the broadly distributed activity measured using EEG. For instance, the observed correlation of high gamma in the spatially summed EEG signal with attended speech could be weak even if high-gamma activity within different brain areas are each significantly correlated with the speech, but at different phases. In general, the SNR of high-gamma signals measured from ECoG is likely greater than from EEG. The fact that we observed no significant attention-

dependent change in the coherences between the high-gamma envelopes and speech signal envelopes is thus most likely due to limitations of scalp recordings.

One other study that examined the effect of attention on gamma-band EEG responses suggested that the attentional enhancement of gamma rhythms was specific to multisensory stimuli (audiovisual), and was not seen for stimuli presented solely to the auditory system (Senkowski et al., 2005); however, this study used simple tonal stimuli. Computational models (Börgers et al., 2008), in vitro studies (Llinás et al., 2002), in vivo electrophysiology (Fries et al., 2001), and modern studies using optogenetics (Cardin et al., 2009) show that gamma-band synchrony over a network of neurons can mediate sensory binding of different components that make up a perceptual object (Tallon-Baudry & Bertrand, 1999), which facilitates attentional selection and routing. Because the behavioral task in the current study involves both segregation (the grouping of input speech features into two separate coherent perceptual streams), and selection (the preferential, detailed processing of one of the two streams), the observed gamma-band effects could be related to either or both of those processes. Further studies are needed to understand the precise mechanisms involved in the generation of gamma-band activity, and how it shapes the network computations associated with segregation and selection (B. Shinn-Cunningham, 2008).

Despite the relatively high amplitude of the signals in the alpha and beta bands (e.g., compared to the gamma band), these mid-frequency bands did not show any attention effects. This is in spite of the fact that both the phase and envelope fluctuations of these bands were considered. At first glance, this result appears to be at odds with the findings of Obseler and colleagues (Obleser & Weisz, 2011; Wöstmann et al., 2016). However, the synchronous alpha variations in those studies were not of the overall alpha power, but rather the lateralization (i.e., left-right hemispherical asymmetry) of the alpha. Moreover, in Wöstmann et al., 2016, both the attended and ignored sound streams had the same temporal structure. This is in contrast to the present study, where the natural differences in the temporal envelope structure of distinct speech streams forms the basis of the analysis. Here, we did not examine any hemifield or hemisphere-specific aspects of attention on the EEG response. Instead, the goal was to examine the overall band-specific effects of attention on EEG responses. Analyses that focus on hemispheric lateralization of rhythms during

spatial selective attention may indeed reveal alpha-band effects. Further, even for speech presented in isolation, cortical processing of linguistic sounds exhibits hemispheric asymmetry with a preferential left lateralization (Morillon et al., 2010). Future work should be undertaken to investigate hemifield-specific effects of attention on EEG, and how these effects interact with asymmetric aspects of cortical processing such as the left-lateralization of phonetic and linguistic processing.

On examining the scalp topography of the speech-EEG coherence, we found that the largest differences in coherence between the attended and ignored conditions occur in temporal-parietal channels, rather than EEG channels that are sensitive to early auditory responses. For example, the N100 EEG response, which is thought to originate from the primary auditory cortex, projects to Cz and Fz channels on the scalp. These channels show a weaker attention effect than the temporal-parietal channels, suggesting that early sensory responses are less modulated by attention than are later processing regions. This is consistent with the observation that attention effects can be localized to later "components" (200–220 ms) of the EEG response by methods such as spread-spectrum analysis, which allow for the temporal signature of the attention effect to be extracted (Power et al., 2012). These results suggest that higher-order processing areas selectively process attended speech.

In the present study, we also find individual differences in the overall magnitude of attentional enhancement of speech-EEG coherences, even though all individuals scored more than 90% in the quiz. This finding is consistent with results from Choi et al., 2014, which used a selective attention task with complex-tone stimuli to show that there are large individual differences in the neural attentional boost, even when performance is at ceiling for all individuals. This study further found that as the behavioral demands became more adverse, the neural attentional boost from the easier condition was predictive of behavioral performance in the harder condition. Taken together with our results, this suggests that EEG measurements from an easier speech-based selective attention task may be used to quantify the top-down attentional contribution to individual differences in speech intelligibility in adverse listening conditions.

Finally, we visualized the coherences across all pairs of speech-EEG features as a bipartite graph, separately for each individual and for each attention condition. We found individ-

ual differences in the structures of attended and ignored graphs (i.e., which speech-EEG relationships were the strongest varied across individuals), and also in the set of EEG and speech features that are most informative about attentional focus in the entire network structure. Such an individual-specific set of just the most informative features can be used for individualized attention-decoding applications that require a compact feature set, such as attention-guided hearing aids (Fiedler et al., 2017; Fuglsang et al., 2017; J. O'Sullivan et al., 2017; Van Eyndhoven et al., 2017) and other BCIs. These features are likely to be more optimal for attention decoding than what may be extracted from more conventional analyses; however, the utility of this approach should be directly tested in future studies. One explanation for the individual differences reported here could be anatomical variations across people, which could lead to EEG measurements being differently sensitive across people to different sources. Another possibility is that every individual's listening strategy might be different. For example, while some individuals may give more weight to spatial cues to perform the task, others may rely more on voice-based cues such as speaker pitch. Finally, there could also be individual differences in the efficacy of attentional modulation of different brain sources (Choi et al., 2014). To elucidate the precise reasons for the individual differences, future studies might consider using high-density recordings and source localization techniques.

2.5 Acknowledgments

This work was supported by National Institutes of Health Grants R01DC013825 (to B.G.S.-C.) and F31DC017381 (to V.V.).

3. MODULATION MASKING AND FINE STRUCTURE SHAPE NEURAL ENVELOPE CODING TO PREDICT SPEECH INTELLIGIBILITY ACROSS DIVERSE LISTENING CONDITIONS

Abstract[1]

A fundamental question in the neuroscience of everyday communication is how scene acoustics shape the neural processing of attended speech sounds and in turn impact speech intelligibility. While it is well known that the temporal envelopes in target speech are important for intelligibility, how the neural encoding of target-speech envelopes is influenced by background sounds or other acoustic features of the scene is unknown. Here, we combine human electroencephalography with simultaneous intelligibility measurements to address this key gap. We find that the neural envelope-domain signal-to-noise ratio in target-speech encoding, which is shaped by masker modulations, predicts intelligibility over a range of strategically chosen realistic listening conditions unseen by the predictive model. This provides neurophysiological evidence for modulation masking. Moreover, using high-resolution vocoding to carefully control peripheral envelopes, we show that target-envelope coding fidelity in the brain depends not only on envelopes conveyed by the cochlea, but also on the temporal fine structure (TFS), which supports scene segregation. Our results are consistent with the notion that temporal coherence of sound elements across envelopes and/or TFS influences scene analysis and attentive selection of a target sound. Our findings also inform speech-intelligibility models and technologies attempting to improve real-world speech communication.

3.1 Introduction

A fundamental question in sensory neuroscience is how our brains parse complex scenes to organize the barrage of sensory information into perceptually coherent objects and streams. Low-level regularities in stimulus features, such as proximity and continuity of boundaries/tex-

[1]↑This chapter was published in bioRxiv (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021).

tures in vision (Gray, 1999), or rhythmicity, pitch, and harmonicity in audition (Darwin, 1997), can promote perceptual binding and scene segregation. Speech perception in complex environments is a prominent example where such feature-based scene analysis is critical for everyday communication (Cherry, 1953). Yet, the neurophysiological mechanisms supporting this process are poorly understood. Elucidating the mechanisms underlying speech intelligibility is important for both clinical applications and audio technologies, such as optimizations for cochlear-implant and hearing-aid signal processing, clinical diagnostics and individualized interventions for speech-in-noise communication problems, and speech denoising algorithms (e.g., in cell phones).

Any acoustic signal can be decomposed into a slowly varying temporal modulation, or envelope, and a rapidly varying temporal fine structure (TFS) (Hilbert, 1906). In the auditory system, the cochlea decomposes broadband inputs into a tonotopic representation, where each channel encodes the signal content in a relatively narrow band of frequencies around a different center frequency. The envelope and TFS information in each channel are then encoded through the activity of neurons in the ascending auditory pathway (Johnson, 1980; P. X. Joris & Yin, 1992). Psychophysical studies suggest that envelopes convey important information about speech content (Elliott & Theunissen, 2009; Shannon et al., 1995; Smith et al., 2002), whereas TFS is important for our perception of attributes such as pitch and location (Smith et al., 2002).

The temporal coherence theory of scene analysis (Elhilali et al., 2009; Gray, 1999) suggests that neural assemblies that fire coherently (driven by envelopes, or TFS, or both) support perceptual grouping of sound elements across distinct frequency channels, which can aid source segregation (Schooneveldt & Moore, 1987). This may also explain how masker elements that are temporally coherent with target speech, but are in a different channel from the target can perceptually interfere (Apoux & Bacon, 2008). Accordingly, the temporal coherence theory makes important predictions about how the envelopes and TFS of sources in a scene affect scene analysis, and thus how they should influence the neural representation and intelligibility of target speech. However, these predictions have not been evaluated in neurophysiological experiments for realistic listening conditions that capture the complexity of everyday "cocktail-party" environments.

A parallel psychoacoustic literature suggests that modulation masking (i.e., the internal representation of temporal modulations in the target relative to those from the background, which contains inherent distracting fluctuations) may be a key contributor to speech understanding in noise (Bacon & Grantham, 1989; Stone & Moore, 2014). Accordingly, while classic speech-intelligibility models emphasized audibility in different frequency bands (ANSI, 1969, 1997), current models that emphasize envelope coding (Steeneken & Houtgast, 1980) and modulation masking (Dubbelboer & Houtgast, 2008; Relaño-Iborra et al., 2016) have been successful in predicting performance in many realistic conditions. However, the core notion that modulation masking is important has not been validated neurophysiologically. With the exception of current speech-intelligibility models that restrict modulation masking effects to within a carrier frequency channel (Jørgensen et al., 2013; Relaño-Iborra et al., 2016), the literature on modulation masking largely does not distinguish between cross-channel interference and within-channel masking. In this sense, the theory of modulation masking is consistent with the temporal coherence theory. However, modulation masking does not consider the role of TFS, despite the consistent finding that cues conveyed by TFS (e.g., pitch) (Smith et al., 2002) critically support object formation, perceptual scene segregation, and selective attention (Darwin, 1997; B. Shinn-Cunningham, 2008). Indeed, temporal coherence across low-frequency TFS and high-frequency pitch envelopes may significantly improve speech intelligibility in noise, compared to having either cue alone (A. J. Oxenham & Simonson, 2009). While some psychophysical studies have explored the relative roles of envelope and fine-structure cues for speech intelligibility in noise (Lorenzi et al., 2006; Qin & Oxenham, 2003; Swaminathan & Heinz, 2012), few neurophysiological studies have investigated how these cues work together during selective listening.

In the present study, we bridge these gaps by measuring electroencephalography (EEG) simultaneously with intelligibility for target speech over a range of strategically chosen realistic listening conditions. The EEG measured is the response evoked by ongoing stimulus fluctuations when attending to the target speech. We hypothesized that the neural tracking of target modulations, as quantified from EEG, will depend strongly on the modulation content of the masker, in line with the temporal coherence theory and the notion of modulation masking. Furthermore, we hypothesized that the availability (or lack thereof) of TFS will

also impact this neural target-envelope coding, in line with the role of TFS in providing cues to facilitate scene analysis and attention. Finally, we hypothesized that the net neural target-envelope coding shaped by these factors (i.e., the neural signal-to-noise ratio (SNR) in the envelope domain) will predict (in a quantitative, statistical sense) speech intelligibility in conditions unseen by the predictive model. Our neurophysiological results provide evidence for all of the above hypotheses. The present study thus goes beyond comparing individual outcomes to neural measures in a particular condition (e.g., Bharadwaj et al., 2015; Ding & Simon, 2013), to elucidate what aspects of the scene acoustics and neural processing predict intelligibility across diverse real-world conditions.

3.2 Materials and Methods

3.2.1 Stimulus generation

700 Harvard/IEEE sentences (Rothauser, 1969) spoken in a female voice and recorded as part of the PN/NC corpus (McCloy et al., 2013) were chosen for the study. The Harvard/IEEE lists have relatively low semantic context compared to other commonly used speech material (Boothroyd & Nittrouer, 1988; Grant & Seitz, 2000; Rabinowitz et al., 1992). Stimuli were created for eight different experimental conditions as described below:

1-3. **Speech in Speech-shaped Stationary Noise (SiSSN):** Speech was added to spectrally matched stationary Gaussian noise, i.e., speech-shaped stationary noise, at SNRs of -2 dB, -5 dB, and -8 dB. The long-term spectra of the target speech sentences and that of stationary noise were adjusted to match the average (across instances) long-term spectrum of four-talker babble. A different realization of stationary noise was used for each SiSSN stimulus.

4-5. **Speech in Babble (SiB):** Speech was added to spectrally matched four-talker babble at SNRs of 4 dB and -2 dB. The long-term spectra of the target speech sentences were adjusted to match the average (across instances) long-term spectrum of four-talker babble. In creating each SiB stimulus, a babble sample was randomly selected from

a list comprising 72 different four-talker babble maskers obtained from the QuickSIN corpus (Killion et al., 2004).

6. **Speech in Babble with Reverberation (SiB Reverb):** SiB at 6 dB SNR was subjected to reverberation simulating the St. Albans Cathedral in England (by convolution with a binaural impulse response; see http://www.openairlib.net). The reverberation time (T60) was 2.4 s.

7. **Vocoded Speech in Babble (SiB Vocoded):** SiB at 4 dB SNR was subjected to 64-channel envelope vocoding, which left the peripheral envelopes and place coding intact, while replacing the TFS with a noise carrier in accordance with the procedure described in Qin and Oxenham, 2003. The 64 frequency channels were contiguous with their center frequencies equally spaced on an ERB-number scale (Glasberg & Moore, 1990) between 80 Hz and 6000 Hz. To verify that the vocoding procedure did not significantly change envelopes at the cochlear level, we extracted the envelopes at the output of 128 filters (using a similar procedure as in the actual vocoding process) both before and after vocoding for 50 different SiB stimuli. Note that the use of 128 filters allowed us to compare envelopes at both on-band filters (i.e., filters whose center frequencies matched with the sub-bands of the vocoder), and off-band filters (i.e., filters whose center frequencies were halfway between adjacent vocoder sub-bands on the ERB-number scale). The average correlation coefficient between envelopes before and after vocoding (across the 50 SiB stimuli and the 128 cochlear filters, and after adjusting for any vocoder group delays) is about 0.9. This suggests that our vocoding procedure leaves the cochlear-level envelopes largely intact. Indeed, as illustrated in Figure 3.1, our 64-channel vocoding procedure better preserves the within-band envelopes than the lower resolution procedures of Ding et al., 2014. Although Ding et al., 2014 suggested that TFS matters for neural envelope tracking, their methods using 4- or 8-channel vocoding do not preserve peripheral envelopes within individual cochlear bands. Consequently, a purely envelope-based explanation of their findings cannot be ruled out.

8. **Speech in Babble with ITFS (SiB ITFS):** SiB at -6 dB SNR was subjected to 64-channel ideal time-frequency segregation (ITFS), a non-linear denoising procedure that forms the basis of many machine-learning denoising strategies (D. Wang & Chen, 2018). This was performed over a frequency range of 80–8000 Hz, mirroring the procedure in Brungart et al., 2006. A local SNR criterion of 0 dB was used in the ITFS procedure.

Prior to the full study, a behavioral pilot study (with five subjects who did not participate in the actual EEG experiment) was used to determine the SNRs for the different experimental conditions. The SNRs for the three SiSSN conditions were chosen to yield intelligibility values of roughly 25%, 50% and 75%, respectively, to span the full range of intelligibility. The SNRs for the other conditions were chosen such that the intelligibility scores were between (but did not include) 0% and 100%, and were different across the different conditions.

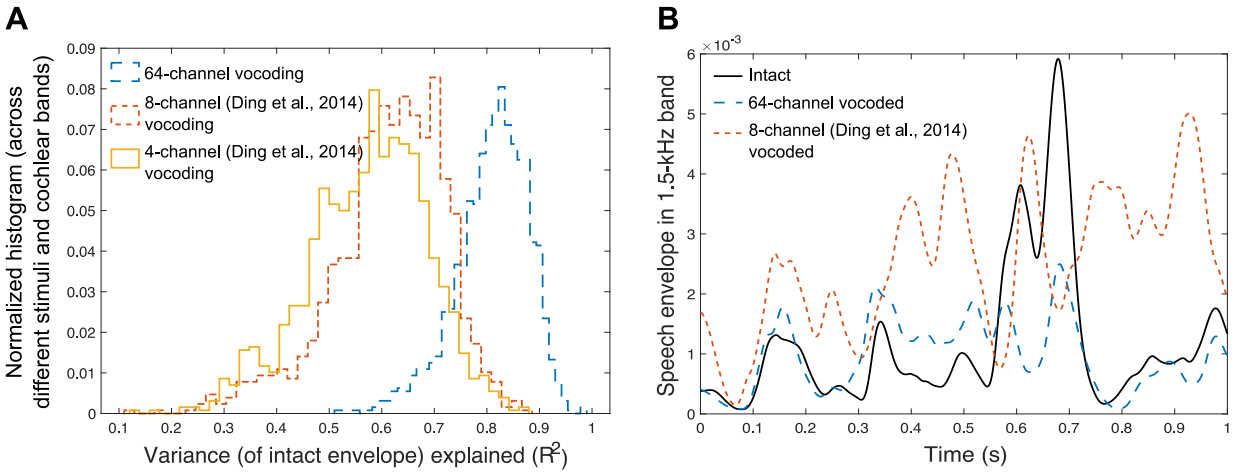

Figure 3.1. : **Illustration of the effect of 64-channel vocoding versus the lower resolution procedures of Ding et al., 2014 on envelopes within individual cochlear bands. Panel A** shows a histogram of the group-delay-adjusted squared normalized-correlation (i.e., variance explained) between the envelope in intact speech in babble (SiB) and 64-channel vocoded SiB, which is used in the present study (i), and the 8-channel (ii) and 4-channel (iii) vocoding of Ding et al., 2014 vocoded SiB. The histograms are across different speech sentences and 128 different cochlear bands equally spaced on an ERB-number scale (Glasberg & Moore, 1990) from 80-6000 Hz. The 64-channel vocoding clearly better preserves the within-band envelopes than either the 8- or 4-channel procedures of Ding et al., 2014 in that the 64-channel procedure captures an additional variance of more than 20%. This disruption of within-band envelopes using their technique was observed despite replicating their result of 0.99 correlation for the band-summed envelope (i.e., the basis for their conclusion that their vocoding preserved speech envelopes). **Panel B** shows an example envelope derived from SiB for the 1.5-kHz speech band for intact SiB, our 64-channel vocoding, and the better-resolution, 8-channel vocoding from Ding et al., 2014, to visualize how our procedure yields band-specific envelopes that more closely match those of intact SiB.

Table 3.1 lists the different stimulus conditions along with the rationale for including them in our study.

Table 3.1. : **Rationale for the different stimulus conditions included in this study.** Collectively, the different listening conditions represent a diversity of scene acoustics, including important examples in our environment and clinical applications. Moreover, they span maskers with different modulation statistics (Jørgensen et al., 2013; S. Rosen et al., 2013) and stimuli with intact and degraded TFS, which allowed us to rigorously test our hypotheses. Note that the SNR levels were chosen to span the full range of intelligibility without floor or ceiling effects.

No.	Stimulus condition	Rationale for inclusion in study
1-3	Speech in Speech-shaped Stationary Noise (SiSSN) at SNRs of -2 dB, -5 dB, and -8 dB	Widely used in the literature; used for calibration of prediction model
4-5	Speech in Babble (SiB) at SNRs of 4 dB and -2 dB	Simulates ecologically relevant cocktail-party listening; has different masker modulation statistics from SiSSN
6	SiB at 6 dB SNR subjected to reverberation (T60 = 2.4 s)	Reverberation is ubiquitous in everyday listening environments (e.g., rooms and stairwells); linearly distorts temporal information
7	SiB at 4 dB SNR subjected to 64-channel envelope vocoding	Used to investigate the role of TFS in target-speech coding and intelligibility
8	SiB at -6 dB SNR subjected to 64-channel ideal time-frequency segregation (ITFS)	ITFS is a precursor to deep-learning-based denoising algorithms that are increasingly used in many audio processing applications, including hearing aids (D. Wang & Chen, 2018); nonlinear distortion

3.2.2 Participants

Data were collected from twelve human subjects (four male), aged 19–31 years, recruited from the Purdue University community. All subjects were native speakers of American English, had pure-tone hearing thresholds better than 20 dB hearing level in both ears at standard audiometric frequencies between 250 Hz and 8 kHz, and reported no neurological disorders. All subjects also had distortion-product and click-evoked otoacoustic emissions (DPOAE and CEOAE) within the normal range of published values for individuals with normal hearing (Gorga et al., 1993), as well as normal tympanograms. Subjects provided informed consent in accordance with protocols established at Purdue University. Data were collected from each subject over the course of one or two visits (with a total visit time of ∼5 hours).

3.2.3 Experimental design

Each subject performed seven blocks of speech intelligibility testing, with 100 trials per block, and with a distinct target sentence in each trial. Subjects had a 5–10 min break between successive blocks. Different but overlapping subsets of experimental conditions were randomly assigned to each subject, such that at least 700 trials for each experimental condition were collected across the subject cohort. This design avoided confounding individual-subject effects with experimental-condition effects. The different experimental conditions were intermingled within each block.

Subjects were instructed that in each trial they would be listening for a woman's voice speaking a sentence, and that at the end of the trial, they would have to verbally repeat the sentence back to the experimenter sitting beside them in a sound-treated booth. They were told that it would be the same woman's voice every time, but that the type and level of background noise/distortion would change from trial to trial. They were also instructed that in each trial, the noise would start first and the target woman's voice $\sim$1 s later. They were encouraged to guess as many words as they could if they heard a sentence only partially.

Stimuli were presented to subjects diotically in all conditions except the reverberation condition, in which stimuli were generated with ear-specific impulse responses as described previously. 32-channel EEG was measured while subjects performed the behavioral task. The target speech sentences were presented at a sound level of 72 dB SPL, while the level of the background was varied according to the stimulus SNR.

At the beginning of each trial, subjects were presented with a visual cue that read "stay still and listen now" in red font. The audio stimulus started playing 1 s after the visual cue was presented. In every stimulus presentation, the background noise started first and continued for the entire duration of the trial, while the target speech started 1.25 s after the background started. This was done to help cue the subjects' attention to the stimulus before the target sentence was played. The target was at least 2.5 s long. After the target sentence ended, the background noise continued for a short amount of time that varied randomly from trial to trial. This was done to reduce EEG contamination from movement artifacts and motor-planning signals. 200 ms after the noise ended, subjects were presented with a

different visual cue that read “repeat now” in green font, cueing them to report the target sentence to the experimenter. Intelligibility was scored on five pre-determined keywords (which excluded articles and prepositions) for each sentence. For each experimental condition, an overall intelligibility score was obtained by averaging the percentage of key words correct (for a sentence) over all sentences used in that condition and across subjects.

Subjects performed a short training demo task before the actual EEG experiment. The demo spanned the same set of listening conditions and used the same woman’s voice as the actual experiment, but contained a different set of Harvard/IEEE target sentences, not used in the main experiment. All twelve subjects scored more than 70% on the easiest condition and got at least some words correct ($> 0\%$) on the hardest condition. All were able to stay still during the presentation of the sentences and respond on cue. This ensured that in the actual experiment, intelligibility scores showed minimal ceiling or floor effects and that movement artifacts were minimal, providing clean EEG recordings.

3.2.4 Hardware

A personal desktop computer controlled all aspects of the experiment, including triggering sound delivery and storing data. Special-purpose sound-control hardware (System 3 real-time signal processing system, including digital-to-analog conversion and amplification; Tucker Davis Technologies, Alachua, FL) presented audio through insert earphones (ER-2; Etymotic, Elk Grove Village, IL) coupled to foam ear tips. The earphones were custom shielded by wrapping the transducers in layers of magnetic shielding tape made from an amorphous cobalt alloy (MCF5; YSHIELD GmbH & Co., Ruhstorf, Germany), and then placing them in 3-mm-thick aluminum enclosures to attenuate electromagnetic interference. The signal cables driving the transducers were shielded with braided metallic techflex. All shielding layers were grounded to the chassis of the D/A converter. The absence of measurable electromagnetic artifact was verified by running intense click stimuli through the transducers with the transducers positioned in the same location relative to the EEG cap as actual measurements, but with foam tips left outside the ear. All audio signals were digitized at a sampling rate of 48.828 kHz. The EEG signals were recorded at a sampling rate of 4.096 kHz

using a BioSemi (Amsterdam, Netherlands) ActiveTwo system. Recordings were done with 32 cephalic electrodes, and two additional earlobe electrodes.

3.2.5 Data preprocessing

Of the twelve subjects who participated, one subject could not complete the task because they were sleepy, and another subject was unable to return for their second visit to complete the study. Data from these two subjects were excluded from the study. The EEG signals of the remaining 10 subjects were re-referenced to the average of the two earlobe reference electrodes. The Signal Space Projection method was used to construct spatial filters to remove eye blink and saccade artifacts (Uusitalo & Ilmoniemi, 1997). The broadband EEG was then band-pass filtered between 1 Hz and 400 Hz for further analysis. Data from three completely new subjects (who were not among the twelve that participated in the main experiment) showed that responses from the auditory cortex and brainstem are strongest in EEG channel FCz (see Fig. 3.2B); thus, we used FCz to derive all results presented in this report.

3.2.6 Quantifying EEG-based target-envelope encoding fidelity

We sought to quantify the fidelity (i.e., SNR) of neural envelope encoding of target speech relative to masker fluctuations for each of the eight experimental conditions. The EEG measured in response to our speech-in-noise stimuli reflects not only the neural responses to the target speech and masking noise, but also unrelated brain activity and other EEG measurement noise. To quantify target-envelope coding, we computed the extent to which the EEG response is phase locked to the target-speech envelope using the phase-locking value measure (PLV; Lachaux et al., 1999). We chose this metric because the PLV is monotonically related to the SNR (approximately linearly in the SNR range of ± 6 dB) in the EEG measurements (Bharadwaj & Shinn-Cunningham, 2014), and consequently also to the neural envelope-domain SNR of the target relative to the masker (as sources of noise other than the masker do not vary between conditions). A high PLV between the target-speech envelope and EEG response indicates a consistent phase relationship between those signals,

and a low PLV implies little to no relationship between the two signals. Thus, if the EEG response mostly coded target fluctuations (e.g., in a condition with low background noise levels), then the PLV between the EEG signal and the target envelope would be strong. On the other hand, if the EEG response coded mostly masker fluctuations rather than target fluctuations, the PLV would be small. Thus, the PLV captures the envelope-domain SNR with which target envelopes are internally represented relative to modulations in masking sounds and random noise. Note that envelope coding has often been quantified in the literature using a stimulus reconstruction approach, which estimates a linear filter that approximately reconstructs the input speech envelope from EEG responses (e.g., Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). Following reconstruction, the proportion of the actual stimulus envelope that is linearly related to the reconstructed envelope is computed as a metric of envelope coding. One disadvantage with this approach is that the first filter estimation step is ill conditioned and necessitates the use of arbitrary regularization techniques (Wong et al., 2018). Our phase-locking measure bypasses this filter estimation step and instead directly captures the proportion of the EEG power that is linearly related to the input speech envelope.

To derive speech envelopes for use in the PLV computation, we used a bank of 10 gammatone filters that mimic cochlear frequency selectivity (Glasberg & Moore, 1990), with center frequencies spanning 100–8500 Hz. The filters were spaced roughly logarithmically, such that their center frequencies had best places that are spaced uniformly along the length of the cochlea according to an established place-frequency map (Greenwood, 1990). Each of the 700 speech sentences used in our study were processed through this filterbank. The envelope of the output of each filter was extracted using the Hilbert transform; the results were summed across all cochlear bands to obtain one overall temporal envelope for each target speech sentence. Note that the single overall envelope obtained by summing across 10 bands is adequate to characterize envelope coding with EEG since EEG does not offer tonotopically resolved information and our focus was not on tonotopic weightings. This is in contrast to the high-resolution procedure crucial for generating vocoded stimuli, as the envelopes conveyed by the periphery are expected to influence the neural processing of target speech. To extract the EEG response to the speech-in-noise stimulus in each trial, a 2.5-s-long

epoch that corresponds to the time window during the trial when the target speech was presented was extracted from the overall EEG response in that trial. Epochs corresponding to a particular experimental condition were then pooled over all subjects who performed the condition, to yield a total of 700 epochs per condition. All EEG epochs for a particular condition were paired with the envelopes of the corresponding target speech sentences, and used to calculate the condition-specific PLV measure. PLV was computed in two different ways using custom code adapted from the MNE-Python toolbox (Gramfort et al., 2014), as described below.

The "long-term" PLV spectrum was estimated for each condition using a multi-taper approach (Zhu et al., 2013). Five tapers were used, which resulted in a frequency resolution of 2.4 Hz. The multi-taper PLV estimate minimizes spectral leakage (i.e., reduces mixing of information between far-away frequencies) for any given spectral resolution, and is calculated from the Fourier representations of two signals $X(f)$ and $Y(f)$ (representing target-speech envelope and EEG response, respectively) as follows:

$$C_{kn}(f) = exp[\mathrm{j}(\angle X_{kn}(f) - \angle Y_{kn}(f))] \tag{3.1}$$

$$PLV_{XY}(f) = \frac{1}{K_{tapers} N_{epochs}} \sum_{k=1}^{K_{tapers}} \left| \sum_{n=1}^{N_{epochs}} C_{kn}(f) \right| \tag{3.2}$$

Here, k indexes the taper, n indexes the epoch, and f is modulation frequency.

In addition to the long-term PLV measure described above, we also computed a short-term multi-resolution PLV for modulation frequencies above 7 Hz to account for any modulation masking release that may occur in short time windows. Multi-resolution analyses have been shown to predict intelligibility better than long-term analyses in the case of fluctuating maskers (Jørgensen et al., 2013). A Morlet wavelet was used to compute the EEG and speech spectra in short time windows using seven cycles at each frequency bin (which resulted in a frequency resolution that monotonically decreased with increasing bin center frequency). The window length is inversely proportional to the wavelet center frequency; thus, the number of windows also varied according to frequency (with fewer windows at lower frequencies, and more windows at higher frequencies). Given that each target sentence was a little over

2 s and that each wavelet had seven cycles, the multi-resolution analysis was restricted to 7 Hz and above in order for at least two non-overlapping windows to be resolvable. The multi-resolution PLV is calculated from the Fourier representations of two signals $X(f)$ and $Y(f)$ (representing target-speech envelope and EEG response, respectively) as follows:

$$C_{mn}(f) = exp[\mathrm{j}(\angle X_{mn}(f) - \angle Y_{mn}(f))] \tag{3.3}$$

$${}^{mr}PLV_{XY}(f) = \frac{1}{M_{win}(f)N_{epochs}} \left| \sum_{m=1}^{M_{win}(f)} \sum_{n=1}^{N_{epochs}} C_{mn}(f) \right| \tag{3.4}$$

Here, m indexes the window, n indexes the epoch, and f is modulation frequency.

The long-term PLV spectra were averaged within octave-wide modulation bands, spaced half an octave apart. In the case of the multi-resolution PLV computation, we used a similar half-octave spacing when defining the wavelet center frequencies. The binned long-term and multi-resolution PLV spectra thus obtained were z-scored with respect to corresponding null distributions of zero phase locking, which were obtained by pairing EEG trials with mismatching speech trials as described in Section 3.2.8. The z-scores from the long-term and multi-resolution analyses were thresholded at zero, and then summed at each frequency bin. Then, to obtain a summary metric of neural envelope coding *ENVneural*, the average PLV over all modulation frequency bins was computed after weighting the z-scores in the bins to compensate for the 1/f power transfer function that is characteristic of the SNR of EEG measurements (Buzsáki et al., 2012; Roß et al., 2000; Stinstra & Peters, 1998). Specifically, the z-score in each frequency bin was weighted by a factor proportional to the square root of the bin center frequency, then the weighted-average z-score across bins was computed. In this way, a separate *ENVneural* metric was quantified for each experimental condition. Note that although different carrier frequency bands and modulation frequencies likely differ in their perceptual importance (Drullman et al., 1994; Kryter, 1962), our *ENVneural* metric does not assign any importance weighting to them. This is because of the possibility that the physiological computations that contribute to our EEG measurements implicitly incorporate such weighting. Figure 3.2 illustrates the steps used to quantify *ENVneural*.

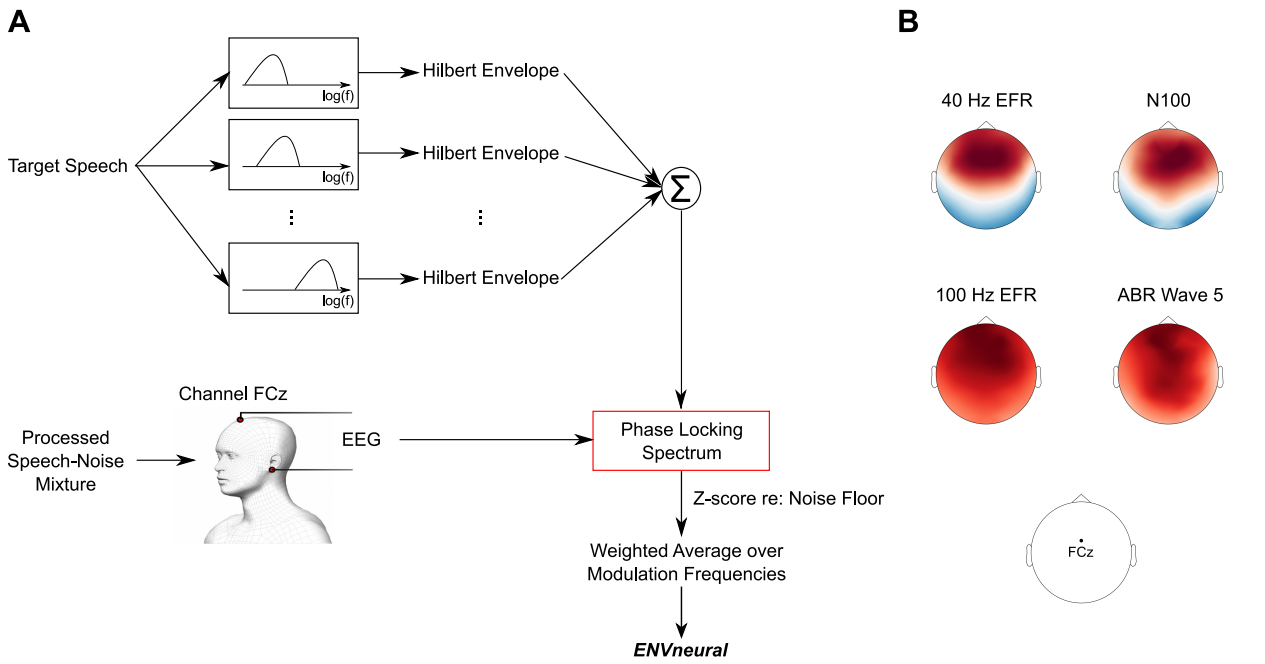

Figure 3.2. : **Quantifying the fidelity of target-speech envelope encoding with EEG. Panel A** illustrates the steps used to quantify target-envelope coding. Target-speech envelopes were extracted using a bank of 10 gammatone filters simulating cochlear processing, with roughly log-spaced center frequencies spanning 100–8500 Hz. The envelope at the output of each filter was extracted using the Hilbert transform, and the results summed across all filters to obtain one overall temporal envelope for each target speech sentence. The fidelity of neural envelope coding of target speech relative to that of background noise (i.e., neural SNR in the envelope domain) was quantified for each experimental condition by computing the phase-locking spectrum between the EEG response in channel FCz and the target-speech envelope across the different trials of that condition (see Equations 3.1, 3.2, 3.3, and 3.4). The resulting phase-locking spectra were z-scored with respect to an appropriate null distribution of zero phase locking. To obtain a summary metric of neural envelope coding *ENVneural*, the average z-score over all modulation frequencies was computed by weighting the frequencies to compensate for the 1/f transfer function that is observed in EEG measurements. **Panel B** shows that responses from auditory cortex and brainstem are strongest in EEG channel FCz. Data shown are from three different subjects who did not participate in the main experiment, but underwent the same screening protocols as the subjects in the main study. Established paradigms for envelope-following responses (EFRs) and onset-evoked potentials (N100 and ABR wave 5) were used to elicit responses from the auditory cortex and brainstem (Picton, 2010). The scalp maps obtained from these responses were normalized such that the amplitudes across channels within each map add to one. The red and blue colors in a scalp map indicate opposite polarities, and the color saturation indicates the normalized amplitude. These scalp maps were used to select the sensor location (FCz) used for all analyses and results presented in this report.

3.2.7 Testing the hypothesis that the fidelity of target-envelope coding in the brain predicts intelligibility

The hypothesis that the fidelity (i.e., SNR) with which envelopes of target speech are coded in the brain relative to background noise predicts intelligibility (in a quantitative, statistical sense) was tested using a rigorous two-step approach. In the calibration step, a logistic/sigmoid function was used to map the EEG-based *ENVneural* measurements to perceptual intelligibility for speech in stationary noise. This mapping revealed a monotonic relationship between *ENVneural* and intelligibility across the three SiSSN conditions (see Fig. 3.4B). A crucial test of envelope-based predictions is whether a mapping between *ENVneural* and perceptual intelligibility derived from one type of background noise can be used to estimate intelligibility for novel backgrounds and linear and non-linear distortions applied to the input sounds. In the next step, we predicted intelligibility for speech presented in various novel, realistic backgrounds and distortions from EEG *ENVneural* measurements and by using the mapping created with stationary noise. The following conditions were tested in the prediction step: SiB at 4 dB SNR, SiB at -2 dB SNR, SiB at 6 dB SNR subjected to reverberation, SiB at 4 dB SNR subjected to 64-channel envelope vocoding, and SiB at -6 dB SNR subjected to non-linear denoising (ITFS). Figure 3.3 illustrates the calibration and prediction steps that were used to test the hypothesis.

3.2.8 Statistical analysis

The distributions for the PLV metric (one for the long-term analysis and another separately for the multi-resolution approach) under the null hypothesis of zero phase locking were obtained using a non-parametric shuffling procedure (Le Van Quyen et al., 2001). Each realization from either null distribution was obtained by following the same computations used to obtain the actual PLV measures, but by pairing EEG response epochs randomly with mismatching speech epochs. That is, when computing the PLV between the speech signal and the EEG signal, the order of epochs for one of the two signals was randomly permuted. This procedure was repeated with 16 distinct randomizations for each experimental condition. Samples were pooled across the 16 randomizations, and across all eight conditions, to yield a

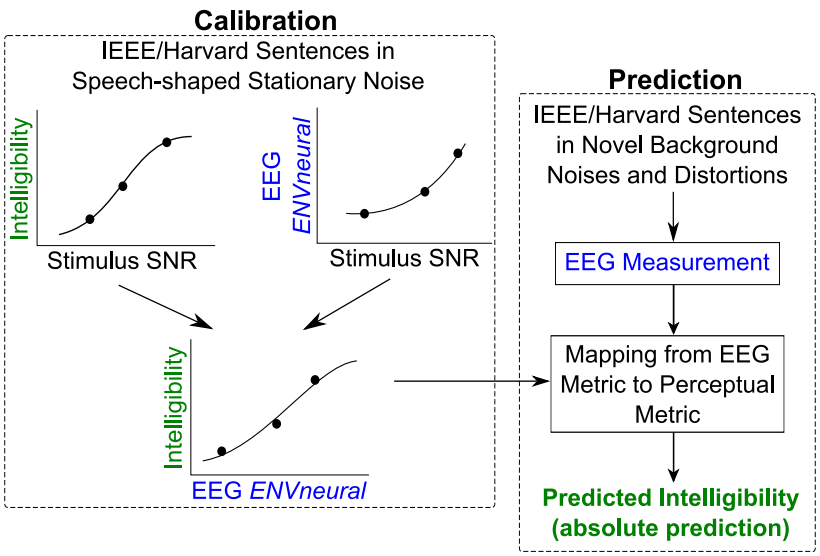

Figure 3.3. : **Our rigorous two-step approach to test the hypothesis that the fidelity of neural envelope coding of target speech relative to background noise predicts speech intelligibility (in a quantitative, statistical sense).** The first step is a calibration step, where a logistic/sigmoid function was used to map an EEG-based target envelope-coding metric *ENVneural* to perceptual intelligibility for speech in stationary noise. In the second step, we used this mapping to blindly predict speech intelligibility in various completely novel realistic background noises and distortions only from EEG-based *ENVneural* measurements.

total of 128 realizations from each null distribution. This procedure ensured that the data used in the computation of the null distributions had the same statistical properties as the original speech and EEG signals.

To test the hypothesis that the fidelity of neural envelope coding of target speech relative to that of background noise (i.e., neural SNR in the envelope domain) predicts intelligibility, we computed the Pearson correlation between our EEG-based intelligibility predictions and the actual intelligibility measurements. The p-value for the correlation was derived using Fisher's approximation (Fisher, 1921).

The noise floor parameters used for computing the z-scores shown in Figure 3.8 were derived as described in Viswanathan et al., 2019.

3.2.9 Software accessibility

Stimulus presentation was controlled using custom MATLAB (The MathWorks, Inc., Natick, MA) routines. EEG data preprocessing was performed using the open-source software

tools MNE-Python (Gramfort et al., 2014) and SNAPsoftware (Bharadwaj, 2018). All further analyses were performed using custom software in Python (Python Software Foundation, www.python.org) and MATLAB. Copies of all custom code can be obtained from the authors.

3.3 Results

3.3.1 Neural envelope-domain SNR in target encoding predicts speech intelligibility over a variety of realistic listening conditions novel to the predictive model

Figure 3.4 shows results from the calibration step of our two-step approach to test the hypothesis that speech intelligibility can be predicted from the fidelity (i.e., SNR) with which target-speech envelopes are encoded in the brain relative to background noise. As described in Section 3.2, the fidelity of target-envelope coding in the brain was estimated from the phase-locking spectrum (computed using the phase-locking value; PLV) between the EEG response and the target-speech envelope. Panel A shows target phase-locking (PLV) spectra for three speech in speech-shaped stationary noise (SiSSN) conditions, which correspond to different acoustic SNRs. Comparing the areas under the PLV spectra for -2 dB SNR, -5 dB SNR, and -8 dB SNR shows that the strength of neural envelope coding of the target monotonically decreases with increasing noise. A summary metric of target-envelope coding, *ENVneural*, was derived separately for each condition by pooling the PLV across modulation frequencies (see Section 3.2). Panel B illustrates the monotonic relationship between *ENVneural* and perceptual intelligibility measurements across the different SNRs of SiSSN. We fit this relationship with a sigmoid/logistic function (shown in the figure) to map *ENVneural* to perceptual intelligibility.

The mapping created in the calibration step was used to predict intelligibility for speech in novel realistic background noises and with different distortions (i.e., conditions not used in calibration), purely from EEG measurements. Figure 3.5 compares predictions to measured intelligibility for the novel conditions. A total of five novel conditions were tested: speech in four-talker babble (SiB) at SNRs of 4 dB and -2 dB, SiB at 6 dB SNR subjected to reverberation, SiB at 4 dB SNR subjected to 64-channel envelope vocoding, and SiB at -6 dB SNR subjected to non-linear denoising (using ideal time-frequency segregation; ITFS).

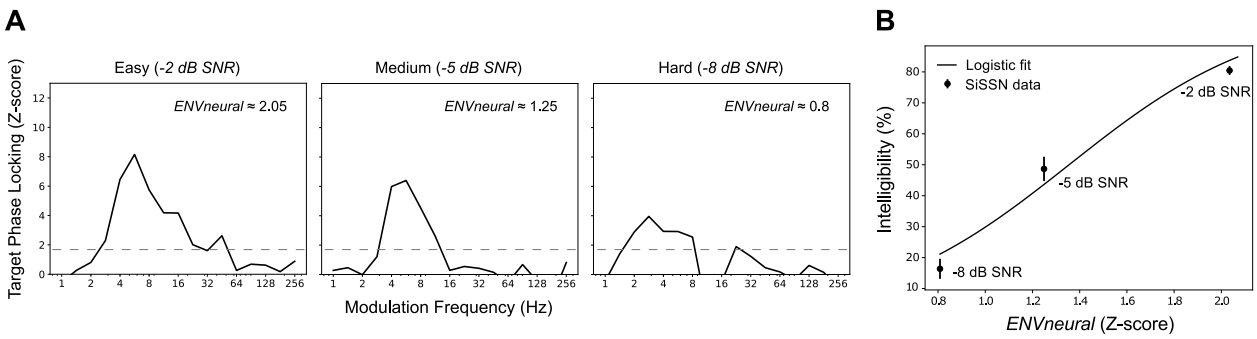

Figure 3.4. : **The calibration step: Stationary noise was used to create a mapping between our EEG-based target envelope-coding metric *ENVneural* and perceptual intelligibility.** Shown are results from the calibration step of our two-step approach to test the hypothesis that speech intelligibility can be predicted from the fidelity (i.e., SNR) with which target-speech envelopes are coded in the brain relative to background noise. Target-envelope coding fidelity was estimated by computing the phase-locking (PLV) spectrum between the EEG response and the target-speech envelope. **Panel A** shows target PLV spectra (z-scored with respect to a null distribution that is common across conditions) for three SNRs of speech in speech-shaped stationary noise (SiSSN). The dashed lines indicate z = 1.64, i.e., the 95th percentile of the noise floor distribution. Neural envelope coding of the target monotonically decreases with increasing noise (compare the areas under the PLV spectra for -2 dB SNR, -5 dB SNR, and -8 dB SNR). A summary metric of target-envelope coding (i.e., *ENVneural*) was derived separately for each condition by pooling the PLV across modulation frequencies. **Panel B** shows *ENVneural* versus intelligibility measurements (mean and standard error across subjects). The monotonic relationship between *ENVneural* and measured intelligibility across the three SNRs of SiSSN allowed us to fit a sigmoid/logistic function mapping *ENVneural* to intelligibility, as shown, which can then be used for predicting intelligibility from measured *ENVneural* for novel conditions.

Predictions match measured performance closely ($R^2 = 0.93$, $p = 0.004$), suggesting that envelope coding of the target (relative to the background) in the central auditory system predicts intelligibility. Note that the measurement noise (i.e., background EEG activity unrelated to the target or masker) would be constant across our comparisons. Hence, the variation in *ENVneural* across conditions should primarily reflect the fidelity of target-envelope coding relative to the masker's internal representation (i.e., the neural modulation-domain SNR). In light of this, the result that *ENVneural* predicts intelligibility across a range of novel realistic conditions provides neurophysiological evidence for perceptual modulation masking.

3.3.2 The modulation frequencies that contribute to the overall *ENVneural* metric, which predicts intelligibility, depend strongly on the envelope spectrum of the masker

Figure 3.6 shows target phase-locking (PLV) spectra for two distinct listening conditions: SiSSN, and SiB, as well as modulation spectra for the speech-shaped stationary noise and four-

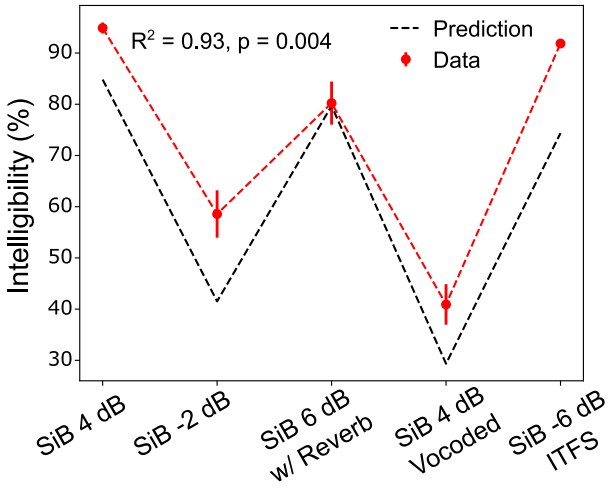

Figure 3.5. : **EEG-based target-envelope coding fidelity predicts intelligibility for a variety of realistic conditions not used in calibration.** The mapping created using stationary noise (Fig. 3.4B) was used to predict intelligibility for speech in completely novel realistic background noises and with various distortions, purely from EEG measurements. A total of five novel conditions were tested: speech in four-talker babble (SiB) at SNRs of 4 dB and -2 dB, SiB at 6 dB SNR subjected to reverberation, SiB at 4 dB SNR subjected to 64-channel envelope vocoding, and SiB at -6 dB SNR subjected to non-linear denoising (using ITFS). Shown are our intelligibility predictions versus actual measurements (mean and standard error across subjects) for these conditions. Predictions match measured performance closely ($R^2 = 0.93$, $p = 0.004$), suggesting that neural envelope coding of target speech (relative to the background) in the central auditory system predicts intelligibility. Since the measurement noise (i.e., background EEG activity unrelated to the target or masker) would be constant across our comparisons, the variation in *ENVneural* across conditions should primarily reflect the fidelity of target-envelope coding relative to the masker's internal representation (i.e., the neural modulation-domain SNR). In light of this, the result shown provides neurophysiological evidence for perceptual modulation masking.

talker babble maskers. The modulation spectra for the maskers were generated by computing the multi-tapered spectral estimates (with five tapers, resulting in a frequency resolution of 2.4 Hz, and 72 trials) of the 2.5-s-long temporal envelope (summed across cochlear bands) of those maskers. Note that the procedure used to generate the masker envelopes was the same as that used to obtain target-speech envelopes for the PLV computation (see Section 3.2). Comparing the modulation spectrum of speech-shaped stationary noise to the target PLV spectrum for the -2 dB SNR SiSSN condition, we find that speech-shaped stationary noise degrades the representation of high-frequency target modulations more (and low-frequency modulations less), in line with the fact that there is greater power for high-frequency than for low-frequency modulation in stationary noise. On the other hand, comparing the modulation spectrum of four-talker babble to the target PLV spectrum for the 4 dB SNR SiB condition, we see that four-talker babble degrades the representation of low-frequency target modulations more (and

high-frequency modulations less). This is consistent with the fact that there is greater power for low-frequency rather than high-frequency modulation in babble. These results show that the spectral profile of EEG-based target-envelope coding fidelity (i.e., the neural envelope-domain SNR in target-speech encoding) is shaped by the masker's modulation spectrum. This result, in combination with our finding that EEG-based target-envelope coding predicts intelligibility, provides further neurophysiological evidence for perceptual modulation masking. These results also suggest that the modulation frequencies that contribute most to speech intelligibility in realistic listening conditions could lie anywhere in the full continuum from slow prosodic fluctuations to fast pitch-range fluctuations. Previous studies that examined electrophysiological responses to speech in background noise, and how those relate to speech perception, focused on either the cortical tracking of low-frequency envelopes (Ding & Simon, 2014), or on the subcortical tracking of envelope fluctuations in the pitch range (Bidelman, 2017; B. Shinn-Cunningham et al., 2017). Our findings thus suggest that the prominent use of stationary noise in the previous cortical speech-tracking literature may have been a contributing factor to their focus on low-frequency speech envelopes, i.e., in the so-called "Delta" and "Theta" ranges.

3.3.3 EEG-based envelope coding fidelity and intelligibility are shaped not just by peripheral envelopes, but also by TFS

Comparing the SiB at 4 dB SNR (intact) condition with the 64-channel envelope-vocoded SiB at 4 dB SNR in Figure 3.7, we find that intelligibility and target-envelope coding fidelity in central auditory neurons are both significantly degraded in the vocoded condition. Note, however, that the envelopes at the cochlear level are very similar before and after vocoding (see Section 3.2), due to the relatively large number of channels (i.e., 64) used in the vocoding process. Despite this, intelligibility is far worse for the vocoded condition compared to the intact condition, demonstrating that the integrity of peripheral envelope cues alone cannot account for speech intelligibility. Importantly, the neural representation of the target envelope in these conditions mirrors these behavioral differences. Thus, the central representation of target envelopes is shaped by factors other than just peripheral envelopes, such as fine-structure-aided segregation mechanisms and selective-attention mechanisms that

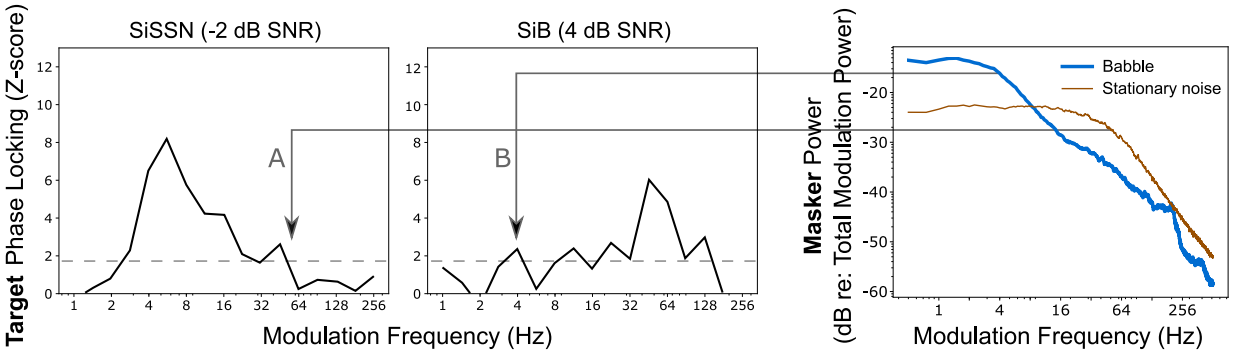

Figure 3.6. : **The modulation frequencies that contribute to the overall *ENVneural* metric, which predicts intelligibility, depend strongly on the envelope spectrum of the masker.** The target phase-locking (PLV) spectra shown are z-scored with respect to a null distribution that is common across conditions. The dashed lines indicate z = 1.64, i.e., the 95th percentile of the noise floor distribution. Comparing the modulation spectrum of speech-shaped stationary noise (rightmost panel) to the target PLV spectrum for the -2 dB SNR SiSSN condition (A), we find that speech-shaped stationary noise degrades the representation of high-frequency target modulations more (and low-frequency modulations less), in line with stationary noise containing relatively more high-frequency modulation power. In contrast, comparing the modulation spectrum of four-talker babble (rightmost panel) to the target PLV spectrum for the 4 dB SNR SiB condition (B), we show that four-talker babble degrades the representation of low-frequency target modulations more (and high-frequency modulations less), consistent with babble containing relatively more low-frequency modulation power. These results show that the spectral profile of EEG-based target-envelope coding fidelity (i.e., the neural envelope-domain SNR in target-speech encoding) is shaped by the masker's modulation spectrum. This result, in combination with our finding that EEG-based target-envelope coding predicts intelligibility, provides further neurophysiological evidence for perceptual modulation masking. These results also suggest that the modulation frequencies that contribute most to speech intelligibility in everyday listening could lie anywhere in the full continuum from slow prosodic fluctuations to fast pitch-range fluctuations.

operate on the segregated representations of target and masker. For example, perceptual cues such as pitch and timbre can aid segregation and selective attention (Darwin, 1997; Micheyl & Oxenham, 2010; B. Shinn-Cunningham, 2008), but these attributes rely upon stimulus TFS (Smith et al., 2002). When segregation cues are ambiguous, selective attention is impaired, as demonstrated by experiments that engineered conflicting cues (Bressler et al., 2014; B. Shinn-Cunningham, 2008). The notion that fine-structure cues work together with envelopes in facilitating segregation is consistent with previous psychophysical studies showing that broadband stimuli produce greater pitch-based masking release compared to low-pass or high-pass speech (A. J. Oxenham & Simonson, 2009).

Many previous studies show that attentional focus, manipulated through subject instruction, can alter central neural envelope coding (e.g., Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). Figure 3.8 illustrates this for a previous study from our lab (reanalysis of

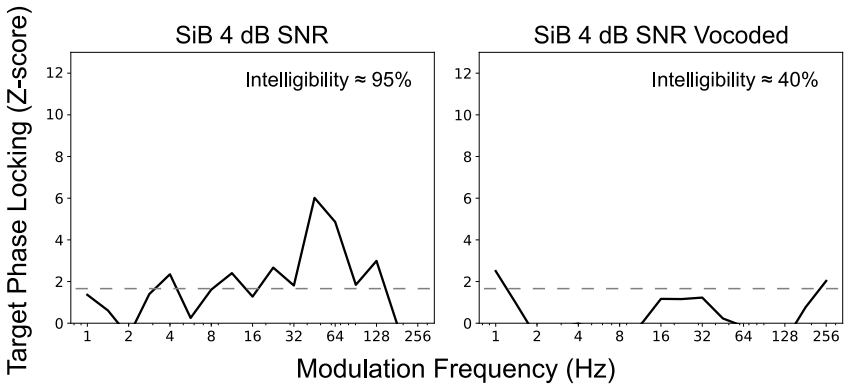

Figure 3.7. : **EEG-based envelope coding fidelity and intelligibility are shaped not just by peripheral envelopes, but also by TFS.** Comparing the target phase-locking (PLV) spectra (z-scored with respect to a null distribution that is common across conditions) for intact and vocoded SiB at 4 dB SNR shows that 64-channel envelope vocoding significantly degrades envelope coding of the target relative to the background in central auditory neurons, even though the envelopes at the cochlear level are very similar before and after vocoding. Concomitantly, intelligibility is far worse for the vocoded condition compared to the intact condition, demonstrating that the integrity of peripheral envelope cues alone cannot account for speech intelligibility. This result shows that central neural envelope coding and intelligibility are shaped by factors other than just peripheral envelopes, such as stimulus TFS, which supports source segregation and selective attention. Note that the dashed lines indicate z = 1.64, i.e., the 95th percentile of the noise floor distribution.

data from Viswanathan et al., 2019). Phase locking (averaged over 10 cochlear bands with center frequencies spanning 100–8500 Hz) between the input speech envelope and EEG response depends directly on what speech a listener attends. For the same input speech stream, the speech envelope of a stream is represented more strongly in the brain when that speech is attended to, rather than when it is ignored. Thus, central neural envelope coding is shaped by not just peripheral envelopes, but also fine-structure-dependent segregation and selective-attention effects. However, no model of speech intelligibility accounts for this fine-structure contribution.

3.3.4 Results support an integrative conceptual model of speech intelligibility

To summarize, our results show that the strength of neural tracking of the target envelope relative to that of the background provides a neural correlate of perceptual interference from a competing sound. Specifically, the ultimate strength of the central auditory system's encoding of the envelope of target speech relative to other interfering sounds predicts speech intelligibility in a variety of real-world listening conditions. Moreover, we find that the

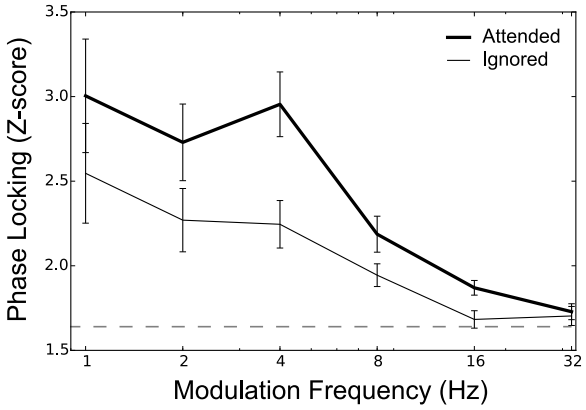

Figure 3.8. : **For the same input speech stream, attentional manipulations (via experimental design) alter central neural envelope coding (data reanalyzed from Viswanathan et al., 2019).** Subjects were presented with a mixture of two running speech streams, one to be attended to and the other ignored. Selective-attention-dependent phase locking was computed between the input speech envelope and EEG response, and averaged over 10 cochlear bands with center frequencies spanning 100–8500 Hz. The data shown are the mean and standard errors of phase locking across 10 subjects. The dashed line indicates z = 1.64, i.e., the 95th percentile of the noise floor distribution. Speech envelopes are represented more strongly in the brain when speech is attended to, versus when the same speech is ignored.

modulation frequencies that contribute to our overall *ENVneural* metric, which predicts intelligibility, depend strongly on the envelope spectrum of the masker and the scene acoustics. Note that modulation-frequency-specific effects can arise from within-channel masking where the masker contains elements that share the same carrier and modulation frequency as some target elements (Jørgensen & Dau, 2011), or from cross-channel interference where masker elements that are coherently modulated with target elements interfere with target coding and perception (Apoux & Bacon, 2008). Our EEG-based metric does not distinguish between these distinct forms of temporal-coherence-based effects. Rather, our results provide evidence that some combination of the two shapes scene analysis and speech perception in noise. Our results also provide direct neural evidence that TFS cues affect how well neural responses in the central auditory system encode the envelope of target speech, likely by aiding in successful source segregation (Darwin, 1997; Micheyl & Oxenham, 2010; A. J. Oxenham & Simonson, 2009) and selective attention (which can operate on the internal representation of segregated target and masker objects to boost the neural representation of the target relative to the masker; Ding & Simon, 2012; J. A. O'Sullivan et al., 2014; Viswanathan et al., 2019). Taken together, our neurophysiological results support the theory that scene analysis and attentive

selection of target speech are influenced by both modulation masking and TFS, consistent with the broader temporal coherence theory. These ideas motivate our conceptual model of speech intelligibility (Fig. 3.9), which consolidates these elements into a single framework.

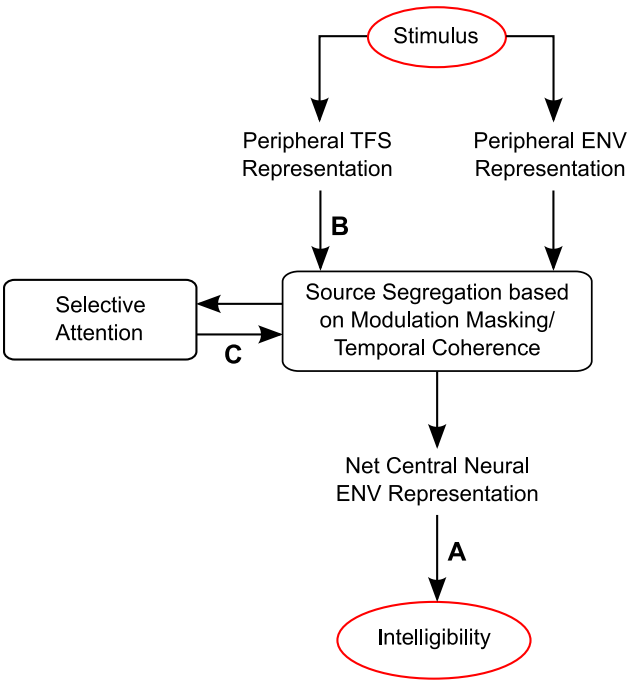

Figure 3.9. : **Results support an integrative conceptual model of speech intelligibility.** Taken together, our results support this integrative conceptual model of speech intelligibility, in that they clarify what internal representation is predictive of speech intelligibility and how that representation is related to the acoustics of the auditory scene and cognitive variables. Our results show that the strength of the net envelope (ENV) coding of target speech relative to other interfering sounds in the central auditory system predicts intelligibility in a variety of real-world listening conditions (arrow A). The modulation frequencies that contribute to these EEG-based intelligibility predictions depend strongly on the envelope spectrum of the masker and the scene acoustics. TFS cues (arrow B) also affect how well neural responses in the central auditory system encode the envelope of target speech, likely by aiding in source segregation (Darwin, 1997; Micheyl & Oxenham, 2010; A. J. Oxenham & Simonson, 2009). Selective attention can then operate effectively on the distinct representations of segregated target and masker objects (arrow C), to boost the neural representation of the target relative to the masker (Ding & Simon, 2012; J. A. O'Sullivan et al., 2014; Viswanathan et al., 2019). Taken together, our results support the theory that scene analysis and attentive selection of target speech are influenced by both modulation masking and TFS, consistent with the broader temporal coherence theory.

3.4 Discussion

The present study systematically examined how neural encoding of target speech in the central auditory system varied as characteristics of the scene acoustics and background noise

were manipulated, and how these neural metrics are related to speech intelligibility. Our results provide support for the temporal coherence theory of scene analysis (Elhilali et al., 2009) in that (i) our EEG-based target-envelope coding metric, which predicts intelligibility, is strongly influenced by the envelopes in background noise in a modulation-frequency-specific manner, and (ii) the availability of intact TFS enhances target-envelope coding.

A key result here is that the neural envelope-domain SNR in target encoding predicts intelligibility (in a quantitative, statistical sense) for a range of strategically chosen real-world conditions that are completely novel to the prediction model. Furthermore, the set of target-envelope frequencies that contribute to our EEG-based intelligibility prediction depends strongly on the envelope frequencies contained in the background sounds. These results together suggest that modulation masking may be fundamentally important for speech perception in noise, thus validating previous behavioral studies (Bacon & Grantham, 1989; Stone & Moore, 2014) and current speech-intelligibility models (Dubbelboer & Houtgast, 2008; Relaño-Iborra et al., 2016) with neurophysiological evidence. Note, however, that our results do not directly provide evidence of neural modulation filter banks (Jørgensen et al., 2013; Relaño-Iborra et al., 2016). Another mechanism by which modulation masking could occur is through temporal-coherence-based binding across a distributed assembly of neurons (Eckhorn et al., 1990; Eggermont, 2006). Through this mechanism, those envelope and fine-structure frequencies of the target that are temporally coherent with components of the masker may get bound together (i.e., a failure of source segregation), which in turn can lead to degraded target representation and perceptual modulation masking at those specific frequencies. Indeed, there is evidence that the redundancy in temporal pitch information across low-frequency resolved harmonics and high-frequency envelopes is more effective in facilitating masking release than what is obtained from either of them individually (A. J. Oxenham & Simonson, 2009). Our findings underscore the need for further research into the neural circuit-level computations that support such complex integration of various temporal cues during active listening.

Previous psychophysical studies with carefully processed speech stimuli in quiet (Elliott & Theunissen, 2009; Shannon et al., 1995; Smith et al., 2002) and the success of envelope-based cochlear implants in quiet backgrounds (B. S. Wilson & Dorman, 2008) suggest that

envelope coding is fundamental for speech perception. However, a more general and rigorous test of this hypothesis requires an examination of whether or not envelope coding predicts intelligibility for the average listener over a range of realistic listening conditions not used by the predictive model. Some prior studies compared individual variations in envelope coding to intelligibility; these used just one type of masker, such as stationary noise (e.g., Ding & Simon, 2013; Vanthornhout et al., 2018) or a multi-talker interferer (e.g., Bharadwaj et al., 2015). In contrast, we were able to predict intelligibility in a variety of novel ecologically relevant conditions from just average neural metrics, learning the prediction model from the independent stationary-noise condition. Despite EEG measurement noise or any errors introduced due to variability in intelligibility measurements in the calibration step, our EEG-based predictions closely track ($R^2 = 0.93$, $p = 0.004$) the overall pattern in measured intelligibility across conditions (Fig. 3.5C). This is in fact stronger evidence that neural envelope coding is a correlate of speech intelligibility than being able to correlate individual differences in neural coding and behavior, both because pooling across subjects (who differ in performance) adds noise to the metrics we computed, and more importantly because correlated individual differences in EEG and behavior could easily arise from extraneous factors such as motivation, attention, level of arousal, etc. that are unrelated to envelope coding (Bharadwaj et al., 2019).

Another fundamental insight from the present study is that central neural envelope coding depends not only on envelopes conveyed by the inner ear, but also on the TFS. Although this result was reported by Ding et al., 2014, they used 4- and 8-channel envelope vocoding to degrade the TFS; this broadband vocoding is in contrast to the high-resolution (64-channel) envelope vocoding that we used here. As demonstrated in Section 3.2, low-resolution vocoding introduces spurious envelopes (not present in the original stimuli) during cochlear filtering of the noise carrier used in vocoding (Gilbert & Lorenzi, 2006). These spurious envelopes introduced within individual frequency channels are large enough to degrade neural envelope coding in a way that is easily perceptible (Swaminathan & Heinz, 2012), and could account for the reduced cortical target-envelope coding they observed (Fig. 3.1). Previous behavioral work (Dorman et al., 1998; Qin & Oxenham, 2003) also shows that increasing the number of noise-vocoding channels beyond eight considerably improves

speech intelligibility in noise, despite the fact that the TFS is uninformative regardless of the number of channels used in vocoding. Together, these results demonstrate that it is necessary to use high-resolution vocoding, as we do here, to unambiguously attribute effects to TFS cues rather than spurious envelopes. Our 64-channel vocoding procedure leaves place coding and cochlear-level envelopes largely intact (Fig. 3.1), not only at filters with center frequencies matching the vocoder sub-bands, but also at filters that are midway between adjacent sub-bands. Thus, it is unlikely that peripheral envelope distortion can account for degraded central neural envelope coding and intelligibility in the present study. These neurophysiological results are consistent with previous behavioral studies showing that fine-structure cues aid in scene segregation and selective processing of target speech (Darwin, 1997; A. J. Oxenham & Simonson, 2009; B. Shinn-Cunningham, 2008). The present study also points to a need for more sophisticated speech-intelligibility models that account for the various scene-analysis mechanisms in play to better predict performance across a wider range of conditions (including vocoded speech-in-noise; Steinmetzger et al., 2019).

Our EEG-based two-step approach can be used to test and refine speech-intelligibility models. A major strength of this approach is that it intrinsically factors in listener attributes (e.g., hearing-loss profile, language experience) and listening state (e.g., focus of attention), in addition to purely stimulus-dependent aspects of coding. How different factors contribute to speech perception can be systematically investigated by characterizing how much each factor contributes to the neural response and how the respective contributions are weighted to best predict intelligibility across various conditions. For instance, here we studied how an acoustic aspect of the stimulus (temporal envelope) is coded in the central auditory system by deriving EEG metrics from scalp locations that strongly reflect auditory cortex and brainstem contributions. In addition, higher-order stimulus features, such as phonemic (categorical) processing (Di Liberto, Crosse, et al., 2018) and semantic composition (Brodbeck et al., 2018) may be studied in future experiments, perhaps by extending our analyses to other brain regions (Di Liberto, Lalor, et al., 2018; Du et al., 2014; Kim et al., 2020). Similarly, by studying individuals with different peripheral pathophysiologies, the effects of various forms of hearing loss on neural coding and intelligibility can also be characterized (Rallapalli & Heinz, 2016; Swaminathan & Heinz, 2011).

One limitation of our approach is that stimulus-related responses in the EEG can be captured and separated from background brain activity only by virtue of their temporal signature. If certain features are encoded through abstract rate-based representations or through different activation profiles within a spatially distributed organization of receptive fields, our approach cannot readily account for them. For example, cortical neurons represent temporal envelopes not only through phase locking, but also through rate-based tuning (X. Wang et al., 2008). Furthermore, place/spectral cues are important for speech recognition (Boothroyd et al., 1996; Elhilali et al., 2003; Shannon et al., 1998), whereas EEG measurements are not place specific but instead reflect population neural activity. One consequence of this fact is that our metrics cannot distinguish between within-channel modulation masking where the masker contains elements that share the same carrier and modulation frequency as some target elements (Jørgensen & Dau, 2011), and cross-channel modulation interference where masker elements that are coherently modulated with target elements interfere with target coding and perception (Apoux & Bacon, 2008). Future EEG studies should attempt to delineate cross-channel versus within-channel effects in scene analysis and speech perception, perhaps by employing frequency-separated target speech and masking sounds. Despite these issues, we find that neural encoding of temporal envelopes can account for much of the intelligibility variations seen across the stimulus conditions tested in this study. This may be because: (i) although EEG signals cannot be readily used to decode the perceived phonemes, they can adequately capture the overall fidelity with which envelopes are coded despite the lack of tonotopic specificity, and (ii) at slow modulation frequencies, temporal coding may be a prominent mechanism in the cortex (X. Wang et al., 2008), and at faster modulation frequencies (e.g., in the pitch range), our metric also includes a small contribution from subcortical portions of the auditory pathway where the coding of envelopes is largely temporal (P. Joris et al., 2004).

3.5 Acknowledgments

This work was sponsored by grants from the National Institutes of Health [F31DC017381 (to V.V.), R01DC009838 (to M.G.H.), R01DC015989 (to H.M.B.), and 9605702, R01DC013825 (to B.G.S.-C.)] and from Action on Hearing Loss [G72 (to M.G.H.)].

4. SPEECH CATEGORIZATION REVEALS THE ROLE OF EARLY-STAGE TEMPORAL-COHERENCE PROCESSING IN AUDITORY SCENE ANALYSIS

Abstract

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 predicted performance was 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 utilized 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 (TFS) 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.

4.1 Introduction

An accumulating body of evidence suggests that temporal-coherence processing is important for multisensory scene analysis (Elhilali et al., 2009; Singer & Gray, 1995). In audition, a rich psychophysical literature on grouping (Darwin, 1997), comodulation masking release (CMR; Schooneveldt & Moore, 1987) and cross-channel interference (Apoux & Bacon, 2008), and pitch-based masking release (A. J. Oxenham & Simonson, 2009) support 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 separate different sources in an acoustic mixture. This theory may help explain how we perform speech separation in a multi-source environment (Krishnan et al., 2014), as speech naturally has common temporal fluctuations across different channels, particularly in the syllabic (0–5 Hz), phonemic (5–64 Hz), and periodicity (i.e., pitch; 64–300 Hz) ranges (Crouzet & Ainsworth, 2001; Swaminathan & Heinz, 2011).

Prior studies on the neural bases of temporal-coherence processing mostly focused on cortical contributions (Elhilali et al., 2009; J. A. O'Sullivan et al., 2015; Teki et al., 2013). 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 (Meddis et al., 2002; Pressnitzer et al., 2001). Moreover, attention, which operates on segregated auditory objects (B. Shinn-Cunningham, 2008), affects responses in the primary 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 theory that speech understanding in noise may be shaped by aspects of temporal-coherence processing that exist in early auditory areas.

While previous studies of temporal-coherence processing mostly used non-speech stimuli (e.g., Elhilali et al., 2009; J. A. O'Sullivan et al., 2015; Teki et al., 2013), 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 in order 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/vowel confusion matrices (Miller & Nicely, 1955), have been widely used in the speech acoustics and cue-weighting literatures, and can indeed provide deeper insight into underlying mechanisms if utilized 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 & Grantham, 1989; Stone & 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 (i.e., when an error was made, what consonant was reported instead of the consonant presented) 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.

4.2 Materials and Methods

4.2.1 Stimulus generation

The stimuli used in the present study draw from and expand on the Materials and Methods previously described in Viswanathan, Shinn-Cunningham, et al., 2021. 20 consonants from the STeVI corpus (Sensimetrics Corporation, Malden, MA) were used. The consonants were /b/, /ʧ/, /d/, /ð/, /f/, /g/, /ʤ/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /ʃ/, /t/, /θ/, /v/, /z/, and /ʒ/. The consonants were presented in CV (consonant-vowel) 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 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 comprising 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 100–7800 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 intact SiB. In accordance with the procedure described in Viswanathan, Shinn-Cunningham, et al., 2021, we retained the cochlear-level envelopes during vocoding but replaced the stimulus temporal fine structure (TFS) with a noise carrier. We verified that the vocoding procedure did not significantly change envelopes at the cochlear level, as described in Viswanathan, Shinn-Cunningham, et al., 2021.

Table 4.1 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 Section 4.2.3). The root mean square (RMS) 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.

4.2.2 Participants

Full details of participant recruitment and screening are provided in Viswanathan, Shinn-Cunningham, et al., 2021, and are only briefly reviewed here. Anonymous subjects were

Table 4.1. : **Rationale for the different stimulus conditions included in this study.** The different listening conditions were chosen to span a range of modulation masking spectral profiles and temporal fine structure (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.

No.	Stimulus condition	Rationale for inclusion in study
1	Speech in Quiet (SiQuiet)	Used as a control condition
2	Speech in Speech-shaped Stationary Noise (SiSSN) at -8 dB SNR	Widely used in the literature; used for calibration of prediction model
3	Speech in Babble (SiB) at -8 dB SNR	Simulates ecologically relevant cocktail-party listening
4	Speech in a masker with only DC modulations (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, Bharadwaj, Shinn-Cunningham, & Heinz, 2021)
5	SiB at 0 dB SNR subjected to 64-channel envelope vocoding	Used to compare performance across models that consider TFS and those that do not (since TFS can influence scene analysis and can convey consonant voicing information in noise; Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021; Viswanathan, Shinn-Cunningham, et al., 2021)

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 neurological 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: (i) 18–55 years old, (ii) self-reported no hearing loss, neurological disorders, or persistent tinnitus, (iii) born and residing in US/Canada, and native speaker of North American English, (iv) experienced Prolific subject, and (v) 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).

4.2.3 Experimental design

The online consonant identification experiment was previously described in Viswanathan, Shinn-Cunningham, et al., 2021. 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 had three parts: (i) Headphone/earphone checks, (ii) Demonstration ("Demo"), and (iii) 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 use was checked using the procedures outlined in Mok et al., 2021. They were then asked to set their computer volume to 10–20% of the full volume, following 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." In order to ensure that all subjects understood and were able to perform the task, only those subjects who scored $\geq 85\%$ in the Demo's Speech in Quiet control condition were selected for the Test stage.

Subjects were given similar instructions in the Test stage as 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 as to whether or not their response was correct to promote engagement with the task. However, subjects were not told what consonant was presented to avoid over-training to the acoustics of how each consonant sounded across the different conditions; the only exception to this rule was in the first sub-part of the Demo where subjects heard all consonants in quiet in sequential order.

We used 50 subjects per talker (subject overlap between talkers was not controlled); with four talkers, this yielded 200 subject-talker pairs, or samples. Separate studies were posted on Prolific.co for the different talkers; thus, when a subject performed a particular study, they would be presented with the speech stimuli for one specific talker consistently over all trials. 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.

4.2.4 Data preprocessing

Only samples with intelligibility scores $\geq 85\%$ for the Speech in Quiet 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.

4.2.5 Quantifying confusion matrices from perceptual measurements

The 20 English consonants used in this study were assigned the phonetic features described in Table 4.2. 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.

Table 4.2. : Phonetic features of the 20 English consonants used in this study.

Consonant	Voicing	Manner of articulation (MOA)	Place of articulation (POA)
/b/	Voiced	Stop	Bilabial
/ʧ/	Unvoiced	Affricative	Palatal
/d/	Voiced	Stop	Alveolar
/ð/	Voiced	Fricative	Dental
/f/	Unvoiced	Fricative	Labiodental
/g/	Voiced	Stop	Velar
/ʤ/	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

In order to test whether there are significant differences in confusion patterns between SiSSN, SiB, SiDCmod, and Vocoded SiB, 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 (due to one condition being inherently easier at a particular SNR). Since overall intelligibility was similar across conditions to start with (Fig. 4.5), 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 Section 4.2.8) to extract statistically significant differences (shown in Fig. 4.6).

4.2.6 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. 30 cochlear filters with characteristic frequencies (CFs) equally spaced on an ERB-number scale (Glasberg & Moore, 1990) between 125 Hz and 8 kHz 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; however, with suppression, the Glasberg and Moore, 1990 tuning is effectively obtained for our broad-band stimuli (Heinz et al., 2002; A. J. Oxenham & 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. 200 stimulus repetitions were used to derive peri-stimulus time histograms (PSTHs) from model auditory-nerve outputs. The model was simulated for the full duration of each stimulus (versus 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 so as 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 manually segmented out from the model PSTHs before being input into the scene analysis models. To do this, we calculated the time segment corresponding to when the target consonant was presented 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). 128 gammatone filters were used for this purpose, with center frequencies between 100–8000 Hz and equally spaced on an ERB-number scale (Glasberg & 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.

4.2.7 Scene analysis modeling to predict consonant confusions

In order 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 Section 4.2.6) corresponding to the different consonants, conditions, and talkers were filtered within a 1-ERB bandwidth (Glasberg & 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 & 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 band-pass 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 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 1; 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 4.1. 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 4.2 shows schematics of both the within- and across-channel models.

To verify that the CMR circuit model (Fig. 4.1) produced physiological correlates of CMR similar to those reported by Pressnitzer et al., 2001, we used the same complex stimuli

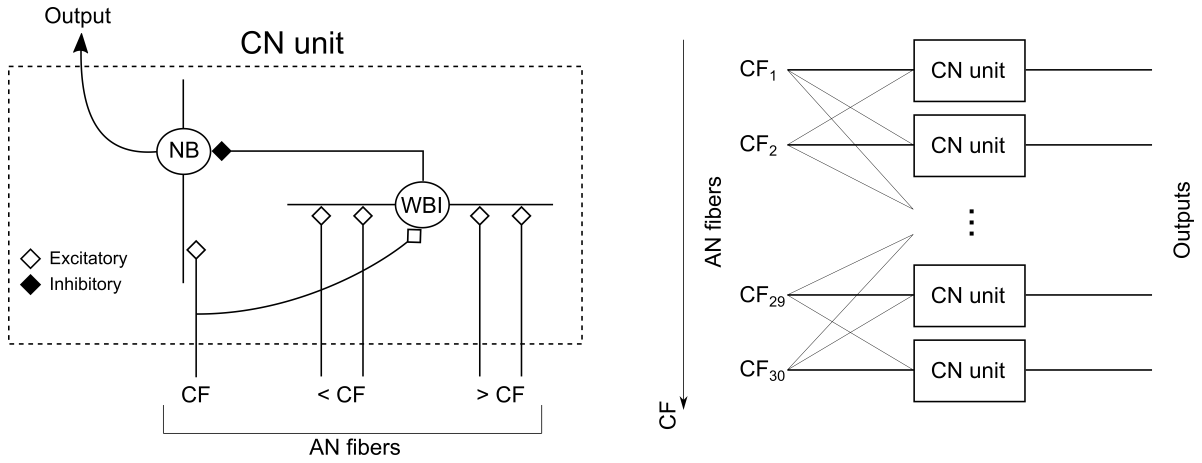

Figure 4.1. : **Comodulation masking release (CMR) circuit based on wideband inhibition in the cochlear nucleus.** This physiologically plausible circuit was proposed by Pressnitzer et al., 2001 to model CMR effects seen in the cochlear nucleus (CN). CN units at different characteristic frequencies (CFs) 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 two octaves below to one 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, i.e., they used AN peri-stimulus time histograms (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 so as 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. Note that the model parameter corresponding to the EI 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.

that they used (Fig. 4.3). 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 Hz, -600 Hz, -400 Hz, 400 Hz, 600 Hz, 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.

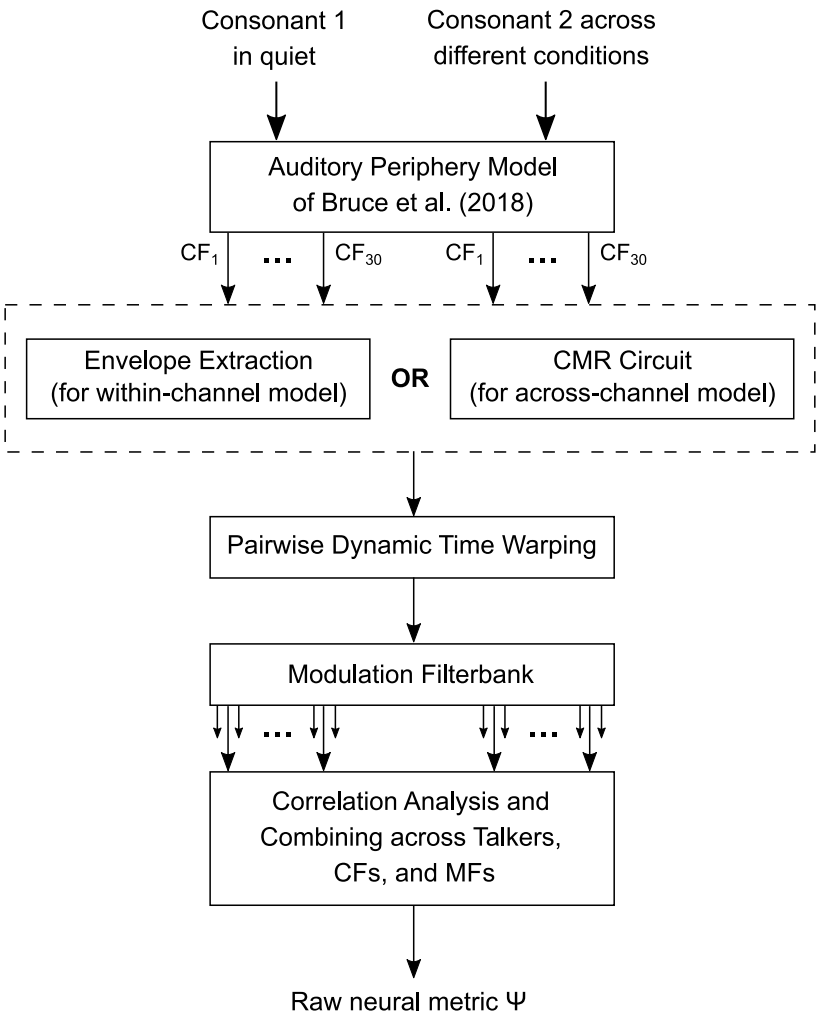

Figure 4.2. : **Schematic of the within- and across-channel scene analysis models.** The speech stimuli were input into the Bruce et al., 2018 model, which simulated a normal auditory periphery with 30 cochlear filters having characteristic frequencies (CFs) equally spaced on an ERB-number scale (Glasberg & 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 & Moore, 1990) to extract band-specific envelopes; however, for the across-channel model, the results were instead input to the CMR circuit model (Fig. 4.1). 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 & 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 1-Hz cutoff in parallel with eight band-pass filters with octave spacing, a quality factor of 1, and center frequencies between 2–256 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 1, which yielded a condition-specific "neural consonant confusion matrix."

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.

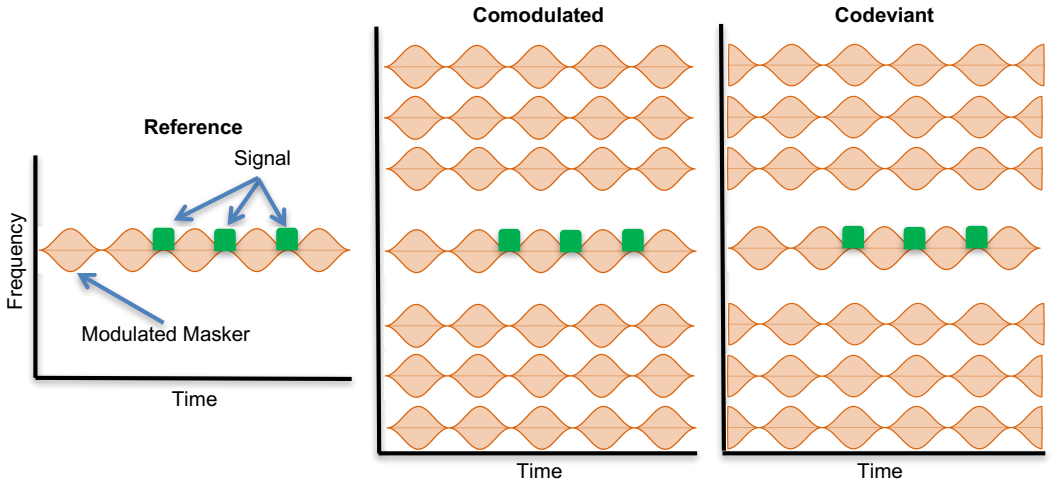

Figure 4.3. : **Stimuli used to validate the CMR circuit model.** The stimuli used were from Pressnitzer et al., 2001, and consisted of a target signal in a 10-Hz 100% sinusoidally amplitude-modulated (SAM) tonal complex masker. The masker differed depending on the experimental condition. In the Reference condition, the masker was a 1.1 kHz-carrier SAM tone (referred to as the on-frequency component or 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 Hz, -600 Hz, -400 Hz, 400 Hz, 600 Hz, 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 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).

The rate-level function at the output of the CMR circuit model (Fig. 4.4D) closely matches physiological data for chopper units in the ventral cochlear nucleus (Winter & 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 4.4A. 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 to the corresponding null distribution to estimate a separate d' for each SCR and condition (Fig. 4.4B). 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 d' 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. 4.4A and 4.4B with Figs. 2 and 6A from Pressnitzer et al., 2001, respectively). 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 and followed the same low-pass trend as human perceptual data (Carlyon et al., 1989) (Fig. 4.4C).

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,

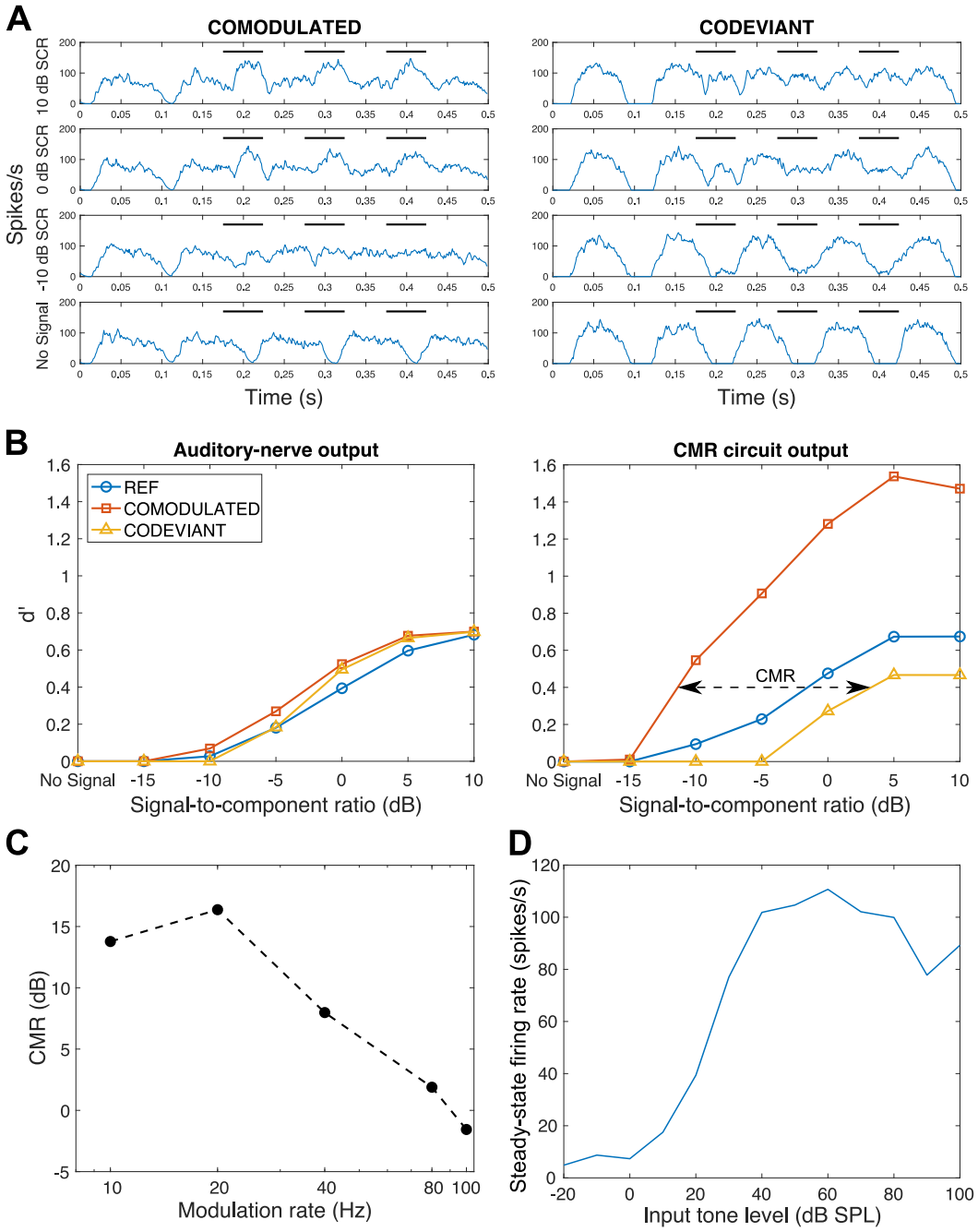

Figure 4.4. : **CMR circuit model validation. Panel A** shows PSTH outputs from the CMR circuit model at 1.1 kHz CF for the stimuli in Figure 4.3. Results are shown separately for the Comodulated and Codeviant conditions, and at different SCRs. The black horizontal bars indicate the time points corresponding to when the target signal was presented. **Panel B** summarizes the results from Panel A by 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, i.e., a CMR effect. This is not seen at the level of the auditory nerve. **Panel C** shows the variation in the CMR obtained from the circuit model as a function of modulation rate. **Panel D** shows the pure-tone rate-level function (i.e., mean steady-state firing rate versus input tone level) for the CMR circuit model.

POA, and MOA categories. The prediction accuracy for the different models is reported in Section 4.3.

4.2.8 Statistical analysis

Permutation testing (Nichols & Holmes, 2002) with multiple-comparisons correction at 5% false-discovery rate (FDR; Benjamini & Hochberg, 1995) was used to extract significant differences in the SiSSN, SiB, SiDCmod, and Vocoded SiB consonant confusion matrices quantified in Section 4.2.5. The null distributions for permutation testing were obtained using a non-parametric shuffling procedure, which 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 difference in the confusion matrices between 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 4.3 and 4.4) 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 employed 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 against the corresponding null distributions to estimate (uncorrected) p-values. To adjust for multiple testing, an FDR

procedure (Benjamini & Hochberg, 1995) was employed. Table 4.5 indicates whether each test met criteria for statistical significance under an FDR threshold of 5%.

4.2.9 Software accessibility

Subjects were directed from Prolific to the SNAPlab online psychoacoustics infrastructure (https://snaplabonline.com; Mok et al., 2021) to perform the study. Offline data analyses were performed using custom software in Python (Python Software Foundation, https://www.python.org) and MATLAB (The MathWorks, Inc., Natick, MA). Copies of all custom code can be obtained from the authors.

4.3 Results

Figure 4.5 shows speech intelligibility measurements from the online consonant identification study. Approximately equal overall intelligibility was achieved for SiSSN, SiDCmod, SiB, and Vocoded SiB due to our careful choice of SNRs for these conditions based on piloting (see Table 4.1). 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.

Given that all psychophysical data were collected online, data quality was verified by comparing results for SiSSN with previous lab-based findings; the analyses performed and the results are described in Mok et al., 2021 and Viswanathan, Shinn-Cunningham, et al., 2021.

The identification data collected in the Test stage of the online experiment were used to construct a consonant confusion matrix for each condition (see Section 4.2.5). Then, statistically significant differences in these matrices across conditions were extracted (Section 4.2.8). Results (Fig. 4.6) show significant differences in the confusion patterns across (i) conditions with different masker modulation statistics, and (ii) stimuli with intact versus degraded TFS information. Computational modeling was then used to predict these differences across conditions to test specific theories of scene analysis (Section 4.2.7).

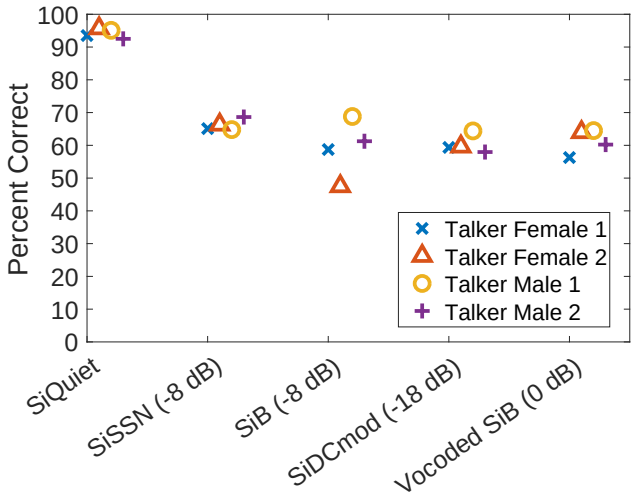

Figure 4.5. : **Overall intelligibility measured in the online consonant identification study for different conditions and talkers.** Approximately equal overall intelligibility was achieved for SiSSN, SiDCmod, SiB, and Vocoded SiB (N=191).

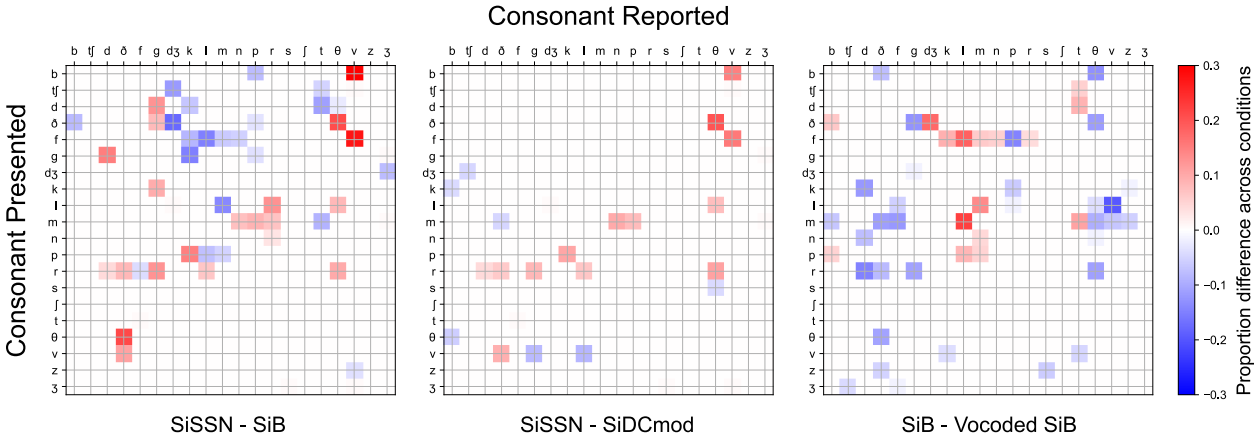

Figure 4.6. : **Measured consonant confusion-matrix differences across conditions (pooled over samples; N=191).** The first two columns represent differences across maskers with different modulation spectra, whereas the third column shows the difference across stimuli with intact versus degraded TFS 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 (Section 4.2.5) 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 (due to one condition being inherently easier at a particular SNR).

Figure 4.7 shows results from the calibration step of testing the within- and across-channel models of scene analysis. In this step, a separate logistic/sigmoid function was fit for each model to map neural confusion matrix entries for the SiSSN condition to corresponding perceptual measurements.

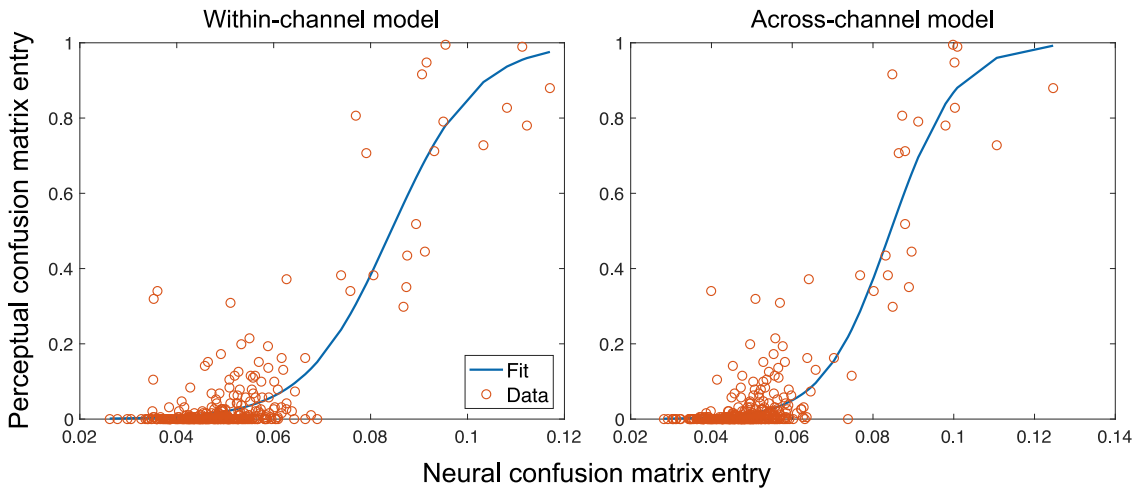

Figure 4.7. : **Calibration result for the within- and across-channel models of scene analysis.** Shown are the model-specific sigmoid/logistic functions that were fit to map neural confusion matrix entries for the SiSSN condition to corresponding perceptual measurements. This model-specific mapping was used to predict perceptual consonant confusion matrices from neural confusion matrices for unseen conditions.

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 (Figs. 4.9, 4.10, and 4.11). Finally, model predictions were compared to perceptual measurements for the different confusion matrix entries across the voicing, POA, and MOA categories. The results are shown in Figure 4.8 for SiB, SiDCmod, and Vocoded SiB. 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-diagonals). But overall, across all entries for SiQuiet, both models predicted diagonal entries 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 (shown in Fig. 4.8) and are given in Tables 4.3 and 4.4 for the within- and across-channel models, respectively. Since the range of confusion matrix entries spanned three orders of magnitude, all comparisons were performed with log-transformed values. The

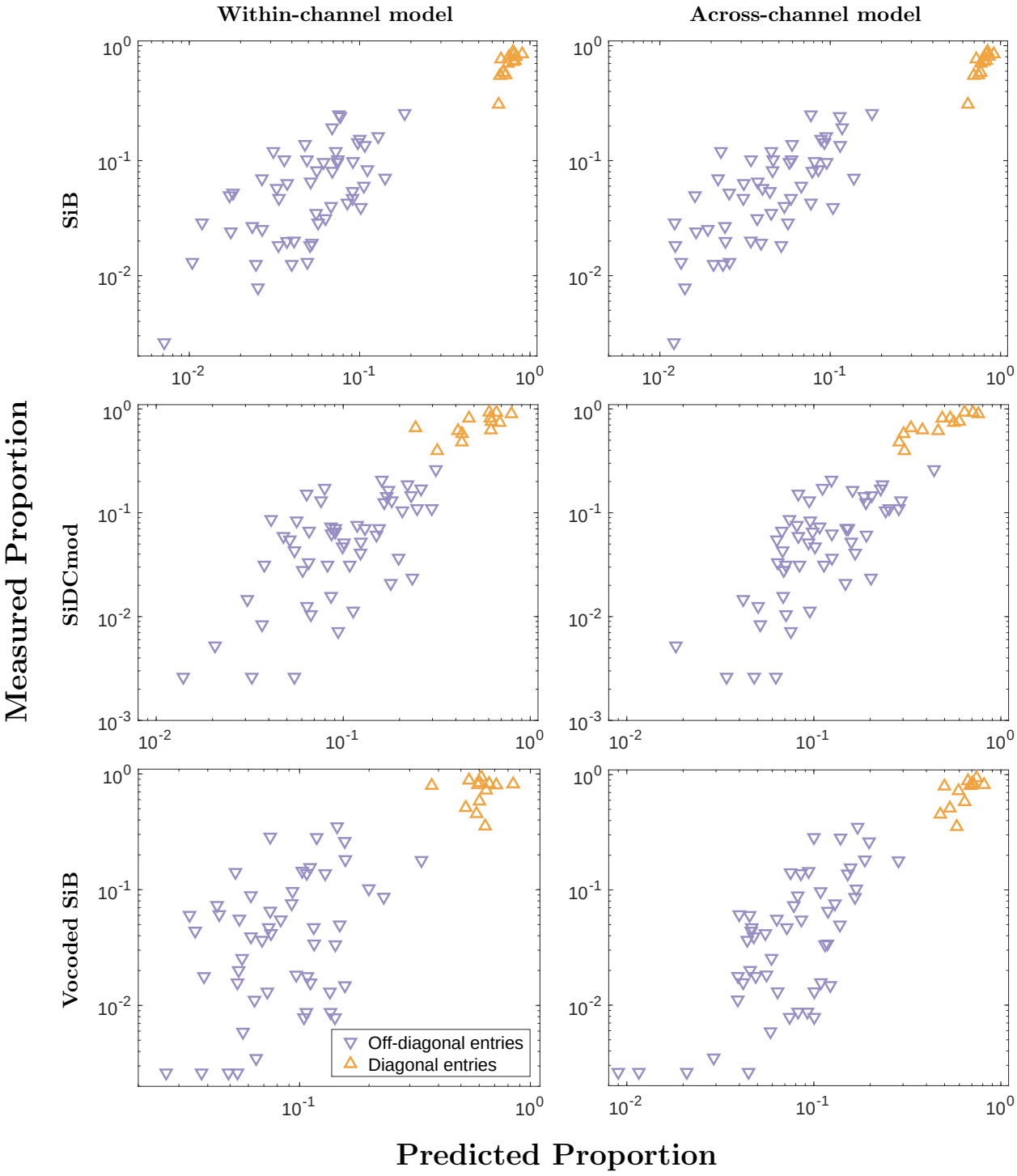

Figure 4.8. : **Within- and across-channel model predictions versus measured confusion matrix entries for the unseen conditions.** Diagonal entries correspond to intelligibility measurements for the different consonant phonetic categories (transmission scores for voicing, POA, and MOA), and off-diagonal entries correspond to true confusions. 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. These trends are quantified in Tables 4.3, 4.4, and 4.5.

correlations were statistically significant across all non-vocoded conditions for the within-channel model, and across all conditions for the across-channel model (see Section 4.2.8 for statistical analysis details). The strong correlation of the within-channel model predictions with perceptual data in the non-vocoded conditions (where TFS cues are preserved) provides independent evidence that speech understanding is strongly influenced by modulation masking when TFS cues are available (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021); 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 4.5; for analysis details, see Section 4.2.8). Thus, a simple physiologically plausible model of across-channel cochlear nucleus processing that shows CMR (Fig. 4.4) 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. 4.1) where temporally coherent pieces of the target and masker, even across distinct cochlear channels, coherently drive the wideband inhibitor (WBI), thereby enhancing outputs of the narrowband (NB) unit (which is inhibited by the WBI) 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 4.5 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 intact SiB (and other non-vocoded 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; A. J. Oxenham & 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.

Table 4.3. : **Pearson correlation coefficients between within-channel model predictions and perceptual measurements.** 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.

Condition	**Diagonal entries**		**Off-diagonal entries**		**All entries**	
	Correlation	p-value	Correlation	p-value	Correlation	p-value
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} ***
SiSSN	83%	0.0002 ***	67%	10^{-8} ***	87%	10^{-21} ***

Table 4.4. : **Pearson correlation coefficients between across-channel model predictions and perceptual measurements.** 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.

Condition	**Diagonal entries**		**Off-diagonal entries**		**All entries**	
	Correlation	p-value	Correlation	p-value	Correlation	p-value
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} ***
SiSSN	89%	10^{-5} ***	81%	10^{-13} ***	92%	10^{-27} ***

Table 4.5. : Improvement in prediction accuracy offered by the across-channel model compared to the within-channel model. The across-channel model showed improved correlations between model predictions and perceptual measurements for all of the unseen conditions, with the largest improvement apparent for Vocoded SiB.

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

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 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 due to discarding TFS, we re-ran the within-channel model by retaining the full auditory-nerve output (results not shown). We found that the predictions from the modified within-channel model were not significantly better than 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. 4.4C), 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.

4.4 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 or not 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 (i) purely within-channel modulation masking (in line with current speech-intelligibility models; Relaño-Iborra et al., 2016), and (ii) 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. 4.4) 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 4.5), 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 (Apoux & Bacon, 2008; Darwin, 1997; Schooneveldt & Moore, 1987). 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 as 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 4.3). One explanation for this is that because pitch-based masking release is poorer in the vocoded condition due to degraded TFS information (A. J. Oxenham & 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 4.5), 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; J. A. O'Sullivan et al., 2015; Teki et al., 2013). Note that the CMR circuit model does not perform pitch-range temporal-coherence processing and no CMR effect was seen at high modulation rates (Fig. 4.4C), 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 (A. J. Oxenham et al., 2004; A. J. Oxenham & Simonson, 2009; Shamma & Klein, 2000).

One limitation of the periphery model 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 & 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 (B. Shinn-Cunningham et al., 2013). Although a debate exists regarding the spatio-temporal properties of different periphery models in cochlear responses (Verhulst et al., 2015), 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 & 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.

4.5 Acknowledgments

This research was supported by grants from the National Institutes of Health [F31DC017381 (to V.V.) and R01DC009838 (to M.G.H.)] and Office of Naval Research [ONR N00014-20-12709 (to B.G.S.-C.)]. We thank Hari Bharadwaj for access to online psychoacoustics infrastructure (https://snaplabonline.com; Mok et al., 2021). We also thank Andrew Sivaprakasam, François Deloche, Hari Bharadwaj, and Ravinderjit Singh for valuable feedback on an earlier version of this chapter.

4.6 Supplementary Information

For completeness, the full set of model-predicted and measured perceptual confusion matrices are shown in Figures 4.9, 4.10, and 4.11 for the voicing, POA, and MOA categories, respectively. 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).

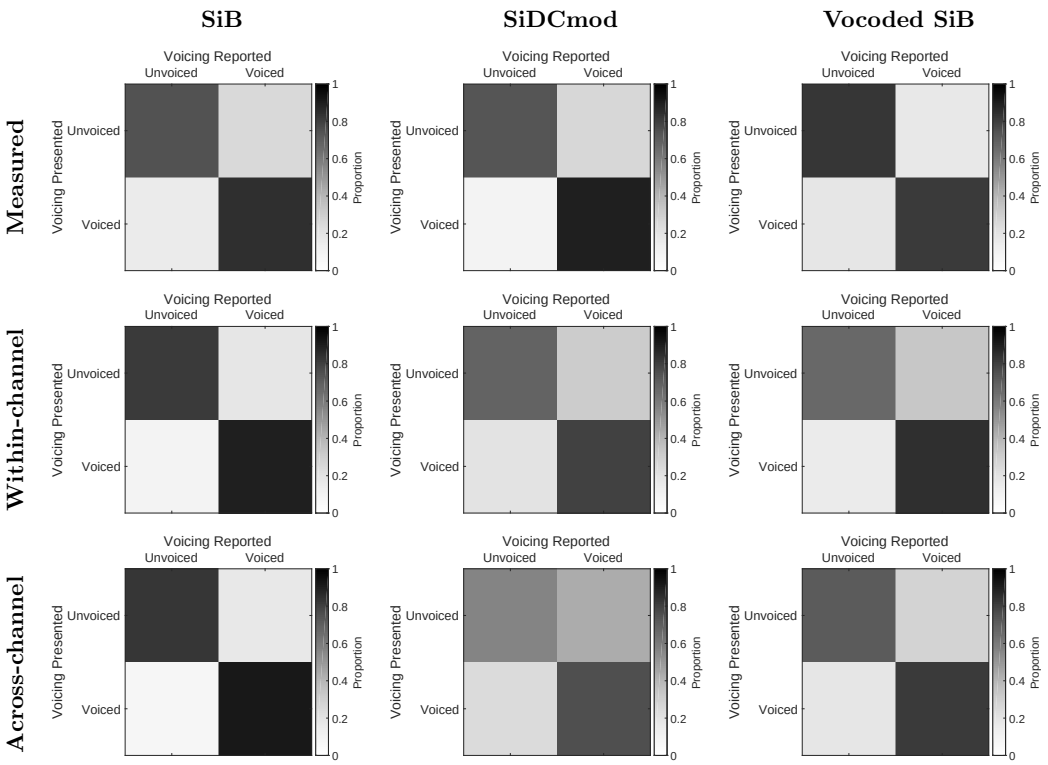

Figure 4.9. : Full set of measured and model-predicted voicing confusion matrices.

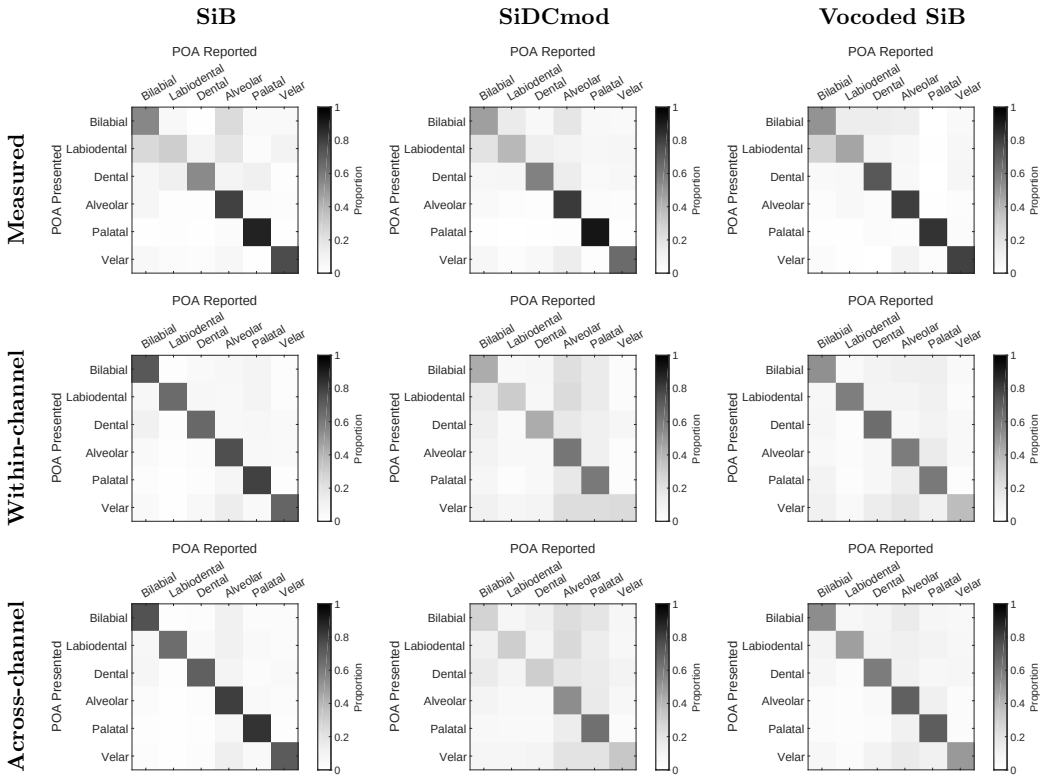

Figure 4.10. : Full set of measured and model-predicted POA confusion matrices.

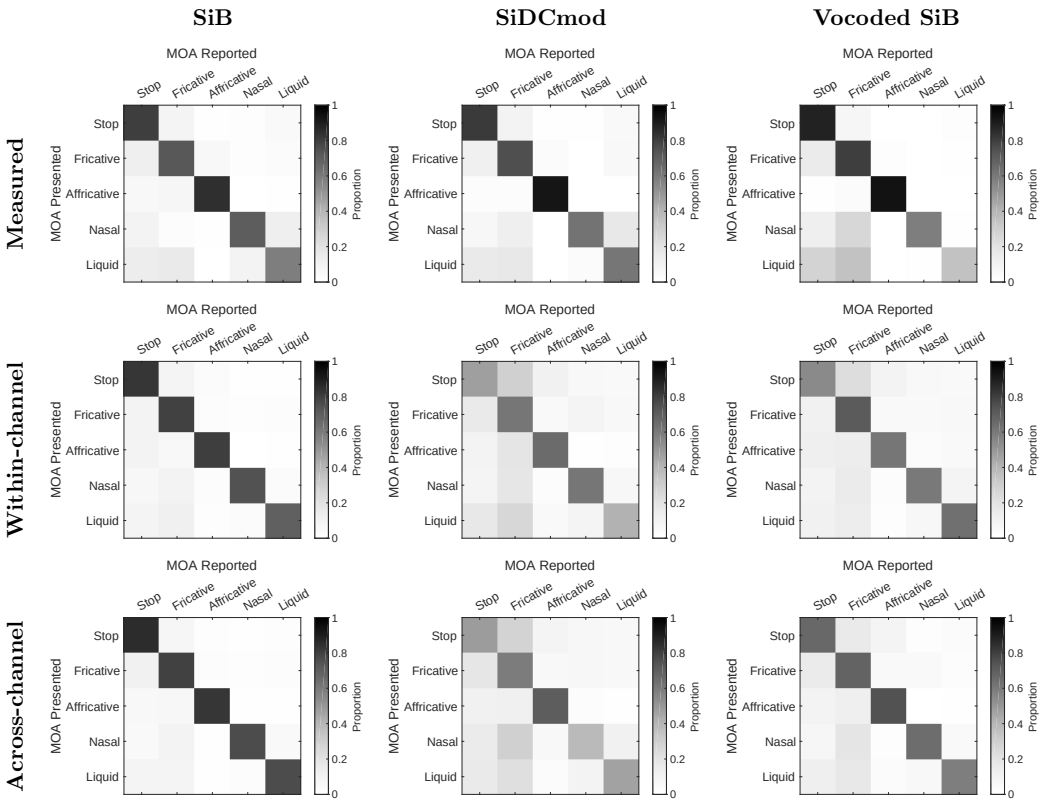

Figure 4.11. : Full set of measured and model-predicted MOA confusion matrices.

5. TEMPORAL FINE STRUCTURE INFLUENCES VOICING CONFUSIONS FOR CONSONANT IDENTIFICATION IN MULTI-TALKER BABBLE

Abstract[1]

To understand the mechanisms of speech perception in everyday listening environments, it is important to elucidate the relative contributions of different acoustics cues in transmitting phonetic content. Previous studies suggest that the envelope of speech in different frequency bands conveys most speech content, while the temporal fine structure (TFS) can aid in segregating target speech from background noise. However, the role of TFS in conveying phonetic content beyond what envelopes convey for intact speech in complex acoustic scenes is poorly understood. The present study addressed this question using online psychophysical experiments to measure the identification of consonants in multi-talker babble for intelligibility-matched intact and 64-channel envelope-vocoded stimuli. Consonant confusion patterns revealed that listeners had a greater tendency in the vocoded (versus intact) condition to be biased towards reporting that they heard an unvoiced consonant, despite envelope and place cues being largely preserved. This result was replicated when babble instances were varied across independent experiments, suggesting that TFS conveys voicing information beyond what is conveyed by envelopes for intact speech in babble. Given that multi-talker babble is a masker that is ubiquitous in everyday environments, this finding has implications for the design of assistive listening devices such as cochlear implants.

5.1 Introduction

Any acoustic signal can be decomposed into a slowly varying amplitude envelope, or temporal modulation, and a fast-varying temporal fine structure (TFS) (Hilbert, 1906). The cochlea decomposes sound input into a multi-channel representation organized by frequency, where each channel encodes the signal content in a relatively narrow band of frequencies around a different carrier frequency. The envelope and TFS information in each channel are then

[1]↑This chapter was published in bioRxiv (Viswanathan, Shinn-Cunningham, et al., 2021).

conveyed to the central nervous system through the ascending auditory pathway (Johnson, 1980; P. X. Joris & Yin, 1992). Elucidating the relative contributions of envelope and TFS cues to speech perception in everyday listening environments is important not just from a basic science perspective, but also for translation to clinical technologies such as cochlear implants.

Psychophysical studies suggest that speech content in quiet can be largely conveyed by envelopes (Shannon et al., 1995). Psychophysical (Bacon & Grantham, 1989; Stone & Moore, 2014), modeling (Dubbelboer & Houtgast, 2008; Relaño-Iborra et al., 2016), and electroencephalography (EEG) (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021) studies support the theory that in the presence of background noise, modulation masking of envelopes of target speech by distracting masker envelopes predicts speech intelligibility across diverse listening conditions. However, in addition to this contribution of envelopes to intelligibility, TFS may also play a role, especially in noisy listening environments (Hopkins & Moore, 2010; Lorenzi et al., 2006).

Psychophysical studies suggest that cues conveyed by TFS (e.g., fundamental frequency; B. C. Moore et al., 2006) can support perceptual scene segregation or unmasking (Darwin, 1997; A. J. Oxenham & Simonson, 2009). Moreover, EEG studies raised the possibility that the neural representation of the attended speech in a sound mixture is sensitive to the spectro-temporal details of the acoustic scene (Ding et al., 2014; Rimmele et al., 2015). By using high-resolution vocoding to alter TFS cues without introducing spurious envelopes, Viswanathan, Bharadwaj, Shinn-Cunningham, and Heinz, 2021 showed that TFS cues per se can influence the coding of attended-speech envelopes in the brain, and that this neural envelope coding in turn predicts intelligibility across a range of backgrounds and distortions. Despite the extensive prior literature on TFS and speech intelligibility, whether TFS can contribute to speech-in-noise perception beyond supporting masking release, i.e., whether TFS can directly convey phonetic content when envelopes are available, is poorly understood. As an analogy to help clarify this gap, consider the role of spatial cues. Spatial cues can provide masking release even though they do not carry any phonetic content. The analogous question here is whether TFS plays a similar role for speech perception in noise in that it only aids in

unmasking, or if TFS can also convey speech content when redundant intact envelope cues are available.

Previous behavioral studies that used TFS-vocoded speech (i.e., where the TFS or phase information in different frequency channels is retained but the envelope information is degraded; e.g., Ardoint & Lorenzi, 2010; Sheft et al., 2008) showed that TFS can convey certain phonetic features with relatively high levels of information reception by means other than envelope reconstruction (i.e., the recovery of degraded speech envelopes at the output of cochlear filters; Gilbert & Lorenzi, 2006; Heinz & Swaminathan, 2009). However, while these studies examined the role of TFS when envelope cues were degraded, they did not address the question of whether or not TFS cues are used for intact speech that has preserved envelope cues.

Another limitation of previous studies that investigated the role of TFS in conveying speech content is that they used masking conditions that were not ecologically realistic. While some used speech in quiet (Ardoint & Lorenzi, 2010; S. Rosen, 1992; Sheft et al., 2008), others presented speech in stationary noise (Gnansia et al., 2009; Swaminathan & Heinz, 2012). Ecologically relevant maskers such as multi-talker babble—a common source of interference in everyday cocktail-party listening—have not been utilized to study this problem. The spectro-temporal characteristics of multi-talker babble (envelope and TFS cues) are similar to what may be encountered in realistic scenarios and a better match to competing speech (albeit without semantic and linguistic content). Thus, multi-talker babble is an important masker to use when studying the role of TFS in speech understanding.

The present study addressed these gaps using online envelope-vocoding experiments designed to probe directly the role of TFS in conveying consonant information beyond what envelopes convey for intact speech (i.e., with redundant envelope cues) in realistic masking environments. Multi-talker babble was used as an ecologically relevant masker. Consonant confusion patterns (Miller & Nicely, 1955) were analyzed, grouping consonants into categories based upon the features of voicing, place of articulation (POA), and manner of articulation (MOA). Confusion patterns were compared between intact and 64-channel envelope-vocoded stimuli for consonants presented in multi-talker babble and separately in quiet (as a control). 64-channel envelope vocoding largely preserves cochlear-level envelopes (Viswanathan,

Bharadwaj, Shinn-Cunningham, & Heinz, 2021), allowing us to study the role of the original TFS in conveying speech content beyond what is conveyed by the intact envelopes. Since TFS plays a role in masking release, vocoding at the same signal-to-noise ratio (SNR) as for intact stimuli produces considerably lower intelligibility. Here this intelligibility drop is mitigated by using a higher SNR for vocoded stimuli so that overall intelligibility was matched for intact and vocoded conditions. By matching intelligibility in this manner, differences in confusion patterns across conditions could be attributed to changes in consonant categorization and category errors rather than differences in overall error counts. Moreover, equalizing intelligibility also maximizes the statistical power for detecting differences in the pattern of confusions. Finally, given that consonants are transient sounds, whether or not effects were robust to changes in the local statistics of the masker were also examined by testing whether results were replicated when the specific instances (i.e., realizations) of multi-talker babble varied across experiments.

The current study tested the hypothesis that TFS does not convey speech content beyond what is conveyed by envelopes for intact speech (i.e., the classic view that envelopes convey speech content and that TFS conveys other attributes like pitch and aids source segregation). As a result, it was expected that once intelligibility was matched across conditions, confusion patterns would be the same for intact and envelope-vocoded stimuli corresponding to speech in (i) babble, and (ii) quiet. The experiments used to test this hypothesis and the results and their implications are described below.

5.2 Materials and Methods

5.2.1 Stimulus generation

20 consonants from the STeVI corpus (Sensimetrics Corporation, Malden, MA) were used. The consonants were /b/, /ʧ/, /d/, /ð/, /f/, /g/, /ʤ/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /ʃ/, /t/, /θ/, /v/, /z/, and /ʒ/. The consonants were presented in CV (consonant-vowel) 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

carrier phrase: "You will mark /CV/ please" (i.e., in natural running speech). Stimuli were created for five experimental conditions:

1. **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 (by applying a filter with a transfer function equal to the ratio of the two spectra). To create each SiB stimulus, a babble sample was randomly selected from a list comprising 72 different four-talker babble maskers obtained from the QuickSIN corpus (Killion et al., 2004).

2. **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 intact SiB. The vocoding process retained the cochlear-level envelopes, but replaced the stimulus fine structure with a noise carrier, in accordance with the procedure described in Qin and Oxenham, 2003. The 64 frequency channels were contiguous with their center frequencies equally spaced on an ERB-number scale (Glasberg & Moore, 1990) between 80 Hz and 6000 Hz. This resulted in roughly two channels per ERB, which ensured that for any given channel, there was one additional channel on each side within 1 ERB. This helps to mitigate spurious envelope recovery on the slopes of cochlear filters, which in turn allows for TFS effects to be better isolated (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021). The envelope in each channel was extracted by using a sixth-order butterworth band-pass filter to extract the component of the intact stimulus in that channel, followed by half-wave rectification and low-pass filtering using a second-order butterworth filter with a cut-off 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.

3. **Speech in Quiet (SiQuiet):** Speech in quiet was used as a control condition.

4. **Vocoded Speech in Quiet (Vocoded SiQuiet):** SiQuiet subjected to 64-channel envelope vocoding (using the same procedure as for Vocoded SiB) was used to examine whether TFS conveys speech content beyond what envelopes convey for intact speech in quiet.

5. **Speech in Speech-shaped Stationary Noise (SiSSN):** Speech was added to stationary Gaussian noise at -8 dB SNR. Similar to what was done for SiB, 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. The SiSSN condition was used for online data quality checking, given that lab-based confusion data were available for this condition (Phatak & Allen, 2007).

Prior to the main consonant identification study, a behavioral pilot study (with three subjects who did not participate in the actual online experiments) was used to determine appropriate SNRs for the different experimental conditions. The SNRs for the intact and vocoded SiB conditions were chosen to give intelligibility of roughly 60%, so that a sufficient number of confusions would be obtained for data analysis.

To verify that the vocoding procedure did not significantly change envelopes at the cochlear level, the envelopes at the output of 128 filters were extracted (using a similar procedure as in the actual vocoding process) both before and after vocoding for SiQuiet and SiB at 0 dB SNR, and for each of the different consonants and talkers. The use of 128 filters allowed us to compare envelopes for both on-band filters (i.e., filters whose center frequencies matched those of the sub-bands of the vocoder), and off-band filters (i.e., filters whose center frequencies were halfway between adjacent vocoder sub-bands on the ERB-number scale). The average correlation coefficient between envelopes before and after vocoding (across the different stimuli and cochlear filters, and after adjusting for any vocoder group delays) was about 0.9 (Fig. 5.1). This suggests that the 64-channel envelope-vocoding procedure left the within-band cochlear-level envelopes largely intact. Thus, although intrinsic envelope fluctuations conveyed by the noise carrier used in vocoding may mask crucial speech-envelope cues in some cases (Kates, 2011), this issue is mitigated by using high-resolution vocoding as

was done in the current study. This high-resolution vocoding allowed us to unambiguously attribute vocoding effects to TFS cues rather than any spurious envelopes (not present in the original stimuli) that can be introduced within individual frequency bands during cochlear filtering of the noise carrier used in vocoding when low-resolution vocoding is performed (Gilbert & Lorenzi, 2006; Swaminathan & Heinz, 2012; Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021).

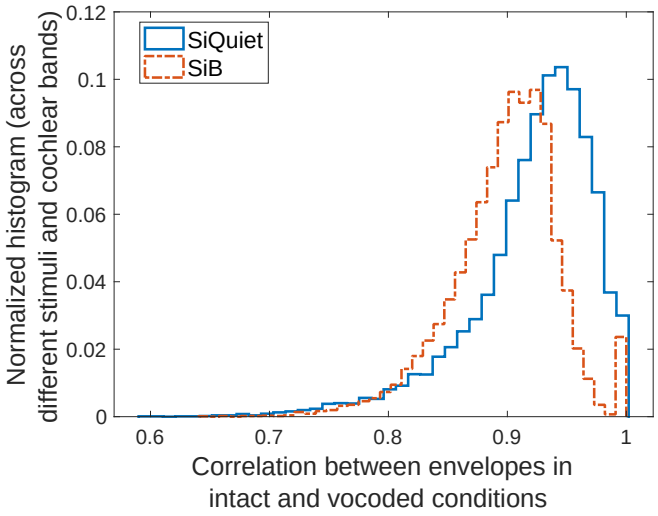

Figure 5.1. : **64-channel envelope vocoding largely preserves the envelopes within individual cochlear bands.** Shown are the normalized histogram of the group-delay-adjusted correlation between the envelope for intact speech in quiet (SiQuiet) and 64-channel vocoded SiQuiet (i), and that for intact speech in babble (SiB) and 64-channel vocoded SiB (ii). The histograms are across the different consonants and talkers, as well as across 128 different cochlear bands equally spaced on an ERB-number scale from 80-6000 Hz. The average correlation between envelopes before and after vocoding was about 0.9.

The stimulus used for online volume adjustment was separately generated, and consisted of running speech mixed with four-talker babble. The speech and babble samples were both obtained from the QuickSIN corpus (Killion et al., 2004); these were repeated over time to obtain a total stimulus duration of ∼20 s (to give subjects adequate time to adjust their computer volume with the instructions described in Section 5.2.3). The volume adjustment stimulus was designed to have a root mean square (RMS) value that corresponded to 75% of the dB difference between the softest and loudest stimuli in the study. This ensured that once subjects had adjusted their computer volume, the stimuli used in the main consonant identification tasks were never too loud for subjects, even at adverse SNRs.

5.2.2 Participants

Data were collected online from anonymous subjects recruited using Prolific.co. The subject pool was restricted using a screening method developed by Mok et al., 2021. The screening method contained three parts: (i) a core survey that was used to restrict subjects based on age to 18–55 years (to exclude significant age-related hearing loss), whether or not they were US/Canada residents, US/Canada born, and native speakers of North American English (because North American speech stimuli were used), history of hearing and neurological diagnoses if any, and whether or not they had persistent tinnitus, (ii) headphone/earphone checks, and (iii) a speech-in-babble-based hearing screening. Subjects who passed the screening were invited to participate in the consonant identification study, and when they returned, headphone/earphone checks were performed again. All subjects had completed at least 40 previous studies on Prolific and had $> 90\%$ of them approved (Prolific allows researchers to reject participant submissions if there is clear evidence of non-compliance with instructions or poor attention). These procedures were validated in previous work, where they were shown to successfully select participants for near-normal hearing status, attentive engagement, and stereo headphone use (Mok et al., 2021). Subjects provided informed consent in accordance with remote testing protocols approved by the Purdue University Institutional Review Board (IRB).

5.2.3 Experimental design

Three nearly identical consonant-identification experiments were conducted to assess the replicability of any main effect of TFS. The experiments were designed with the goal of contrasting intact and vocoded conditions (i.e., stimuli with original and disrupted TFS), while roving the levels of all other experimental variables (i.e., consonants, talkers, conditions, and masker instances). Thus, each experiment presented, in random order, one stimulus repetition for each of the 20 consonants across all four talkers and all five experimental conditions. Within a given experiment, in creating each intact or vocoded SiB stimulus, babble instances (i.e., realizations) were randomly chosen from a list comprising 72 different four-talker babble maskers (see Section 5.2.1); thus, the babble instances that were used for

a particular consonant and talker were not the same between the intact and vocoded SiB conditions. To test whether the main effects of fine structure generalized when the babble instances used were varied across experiments, a different random pairing of masker instances was used across consonants, talkers, and conditions in Experiment 2 compared to Experiment 1. Experiment 3 used, as a sanity check while testing replication of effects, the same stimuli as Experiment 2. Thus, the only difference in the stimuli between the experiments was in the particular instance of babble that was paired with a particular consonant, talker, and SiB condition (intact, and vocoded). As observed by Zaar and Dau, 2015, when effects are instance-specific, different realizations of the same masker random process can contribute significantly larger variability to consonant identification measurements than across-listener variability. Thus, our study design of varying babble instances across the three experiments helped to disambiguate any effects of vocoding from masker-instance effects.

25 subjects per talker were used (subject overlap between talkers was not controlled) in each of the three experiments. With four talkers, this yielded 100 subject-talker pairs, or samples, per experiment. Separate studies were posted on Prolific.co for the different talkers; thus, when a subject performed a particular study, they would be presented with the speech stimuli for one specific talker consistently over all trials. There was no overlap between experiments in the particular set of 100 samples that were used, i.e., samples were independent across experiments. Within each experiment, 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.

Subjects performed the tasks using their personal computers and headphones/earphones. Our online infrastructure included checks to prevent the use of mobile devices. Each of the three experiments had three parts: (i) Headphone/earphone checks, (ii) Demonstration ("Demo"), and (iii) Test (which was the main stage of the experiment). 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. They were then asked to set their computer volume to 10–20% of the full volume, following 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 paradigm of Mok et al., 2021 was used for headphone/earphone checks. In this paradigm, subjects first performed the task described by Woods et al., 2017. While the Woods et al., 2017 task can distinguish between listening with a pair of free-field speakers versus using stereo headphones/earphones, it cannot detect the use of a single free-field speaker or a mono headphone/earphone. Thus, the Woods et al., 2017 task was supplemented with a second task where the target cues were purely binaural in nature, thereby allowing us to test if headphones/earphones were used in both ears. The second task was a 3-interval 3-alternative forced-choice task where the target interval contained white noise with interaural correlation fluctuating at 20 Hz, while the dummy intervals contained white noise with a constant interaural correlation. Subjects were asked to detect the interval with the most flutter or fluctuation. Only those subjects who scored greater than 65% in each of these two tasks were allowed to proceed to the next (Demo) stage of the experiment. This two-task paradigm to verify stereo headphone/earphone use was validated in Mok et al., 2021.

In the Demo stage, subjects performed a short training task designed to familiarize them 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 at the end of the trial, they would be given a set of options for *something*, and that they would have to click on the corresponding option. Consonants were first presented in quiet, and in sequential order starting with /b/ and ending with /ʒ/. This order was matched in the consonant options shown on the screen at the end of each trial. After the stimulus ended in each trial, subjects were asked to click on the consonant they heard. After subjects had heard all consonants sequentially 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." Only subjects who scored $\geq 85\%$ in the Demo's Speech in Quiet control condition were selected for the Test stage, so as to ensure that all subjects understood and were able to perform the task.

In the Test stage, subjects were given similar instructions as in the Demo, but told to expect trials with background noise from the beginning (rather than midway through the task as in the Demo). In both Demo and Test, the background noise (babble or stationary noise), when present, started 1 s before the target speech and continued for the entire duration of the trial. In both Demo and Test, to promote engagement with the task, subjects received feedback after every trial as to whether or not their response was correct. Subjects were not told what consonant was presented, to avoid over-training to the acoustics of how each consonant sounded across the different conditions, except for the first sub-part of the Demo, where subjects heard all consonants in quiet in sequential order.

5.2.4 Data preprocessing

Only samples (i.e., subject-talker pairs) with intelligibility scores $\geq 85\%$ for the Speech in Quiet 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.

5.2.5 Quantifying confusion matrices

The 20 English consonants used in this study were assigned the phonetic features described in Table 5.1. The identification data collected in the Test stage of each experiment were used to construct consonant confusion matrices (pooled over samples) separately for each condition. Overall intelligibility was normalized to 60% for intact and vocoded SiB, and to 90% for intact and vocoded SiQuiet by scaling the confusion matrices such that the sum of the diagonal entries was the desired intelligibility. Matching intelligibility in this manner allowed for differences in confusion patterns across conditions to be attributed to changes in consonant categorization and category errors rather than differences in overall error counts (due to one condition being inherently easier at a particular SNR). Furthermore, equalizing intelligibility also maximizes the statistical power for detecting differences in the pattern of confusions. The resulting confusion matrices (Fig. 5.10) were used to construct voicing, POA, and MOA confusion matrices by pooling over all consonants. In order to test our hypothesis that voicing,

POA, and MOA confusion patterns would be the same for intact and envelope-vocoded speech in babble (after matching intelligibility), the difference between intelligibility-matched intact and vocoded SiB confusion matrices was computed. Confusion-matrix differences were then compared with appropriate null distributions of zero differences (see Section 5.2.6) to extract statistically significant differences (shown in Figs. 5.6, 5.7, and 5.8). A similar procedure was used to test whether TFS conveys phonetic content beyond what is conveyed by envelopes for intact speech in quiet, but by pooling data across all three experiments when constructing confusion matrices for intact and vocoded SiQuiet (versus examining effects separately for each experiment, as was done for intact and vocoded SiB). This data pooling across experiments was performed to improve statistical power because of the relatively high overall intelligibility in quiet.

Table 5.1. : Phonetic features of the 20 English consonants used in this study.

Consonant	Voicing	Manner of articulation (MOA)	Place of articulation (POA)	Binary POA
/b/	Voiced	Stop	Bilabial	Front
/ʧ/	Unvoiced	Affricative	Palatal	Back
/d/	Voiced	Stop	Alveolar	Back
/ð/	Voiced	Fricative	Dental	Front
/f/	Unvoiced	Fricative	Labiodental	Front
/g/	Voiced	Stop	Velar	Back
/ʤ/	Voiced	Affricative	Palatal	Back
/k/	Unvoiced	Stop	Velar	Back
/l/	Voiced	Liquid	Alveolar	Back
/m/	Voiced	Nasal	Bilabial	Front
/n/	Voiced	Nasal	Alveolar	Back
/p/	Unvoiced	Stop	Bilabial	Front
/r/	Voiced	Liquid	Palatal	Back
/s/	Unvoiced	Fricative	Alveolar	Back
/ʃ/	Unvoiced	Fricative	Palatal	Back
/t/	Unvoiced	Stop	Alveolar	Back
/θ/	Unvoiced	Fricative	Dental	Front
/v/	Voiced	Fricative	Labiodental	Front
/z/	Voiced	Fricative	Alveolar	Back
/ʒ/	Voiced	Fricative	Palatal	Back

5.2.6 Statistical analysis

To examine the role of TFS in conveying speech content, the difference in the voicing, POA, and MOA confusion matrices between intact and vocoded conditions was computed, separately for speech in babble and speech in quiet. Permutation testing (Nichols & Holmes, 2002) with multiple-comparisons correction at 5% false-discovery rate (FDR; Benjamini &

Hochberg, 1995) was used to extract significant differences in the confusion patterns. The null distributions for permutation testing were obtained using a non-parametric shuffling procedure, which ensured that the data used in the computation of the null distributions had the same statistical properties as the measured confusion data. Separate null distributions were generated for speech in babble and speech in quiet, and for the different phonetic categories. Each realization from each null distribution was obtained by following the same computations used to obtain the actual "intact - vocoded" confusion matrices, but with random shuffling of intact versus vocoded condition labels corresponding to the measurements. This procedure was repeated with 10,000 distinct randomizations for each null distribution.

To quantify the degree to which statistically significant "intact - vocoded" confusion differences were replicated across the three experiments, simple Pearson correlation was used and the p-value for the correlation was derived using Fisher's approximation (Fisher, 1921). Although the entries of each difference matrix are not strictly independent (which can cause p-values to be underestimated), this p-value approximation was considered adequate given that the individual p-value estimates were not near conventional significance criteria (i.e., were orders of magnitude above or below 0.05).

5.2.7 Signal-detection theoretic analysis

A signal-detection theoretic analysis (Green & Swets, 1966) was used to calculate the bias, i.e., the shift in the classification boundary, in the average subject's percept of voicing for target speech in babble relative to an unbiased ideal observer (i.e., a classifier that optimally uses the acoustics to arrive at a speech-category decision) (see Fig. 5.2). The extent to which this bias was altered by vocoding was then quantified. This analysis was motivated by the finding that vocoding had a significant and replicable effect on voicing confusions for speech in babble across the three experiments in our study.

Let us define the null and alternative hypotheses for the voicing categorization performed by listeners. Let $\mathcal{H}0$ be the null hypothesis that an unvoiced consonant was presented, and let $\mathcal{H}1$ be the alternative hypothesis that a voiced consonant was presented. Let FA be the probability of a false alarm, and HR be the hit rate. The FA and HR values for each

experiment and condition were obtained from the voicing confusion matrix (pooled over samples and consonants) corresponding to that experiment and condition.

The cutoff C (or decision boundary) for the average subject's perceptual decision on whether or not to reject $\mathcal{H}0$, d', and listener bias B (expressed as a percentage relative to an unbiased ideal observer's cutoff) were calculated separately for each experiment and condition (intact versus vocoded SiB) as:

$$C = \phi(1 - FA), \tag{5.1}$$

$$d' = \phi(1 - FA) - \phi(1 - HR), \tag{5.2}$$

and

$$B = \frac{(C - d'/2) \times 100}{d'/2}, \tag{5.3}$$

where ϕ is the inverse of the standard normal cumulative distribution.

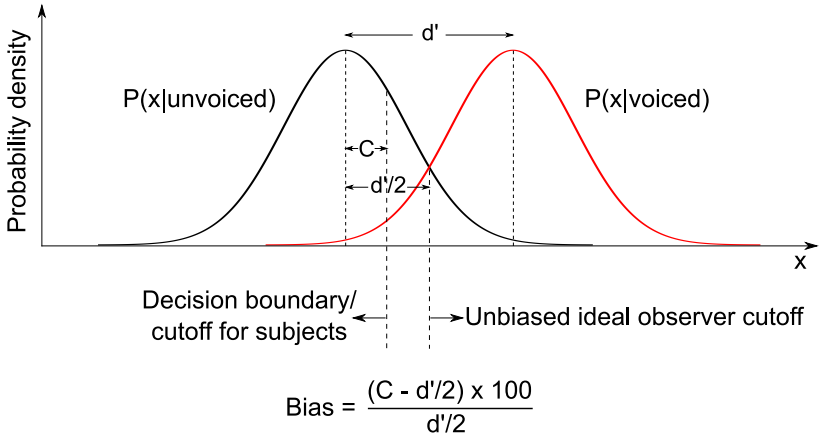

Figure 5.2. : **Illustration of a decision-theoretic quantification of speech categorization bias.** x denotes the internal decision variable. Bias is quantified as the percent shift in the average listener's cutoff (or decision boundary) relative to an unbiased ideal observer's cutoff. The cutoff values for the average listener and the ideal observer were estimated from the false-alarm and hit rates in the data.

The change in the listener bias between the intact and vocoded SiB conditions was derived as:

$$B_{vocoded} - B_{intact}, \tag{5.4}$$

where $B_{vocoded}$ and B_{intact} are the biases in the vocoded and intact SiB conditions, respectively.

5.2.8 Software accessibility

Subjects were directed from Prolific to the SNAPlab online psychoacoustics infrastructure (https://snaplabonline.com; Mok et al., 2021) to perform the study. Offline data analyses were performed using custom software in Python (Python Software Foundation, https://www.python.org) and MATLAB (The MathWorks, Inc., Natick, MA). Copies of all custom code can be obtained from the authors.

5.3 Results

Figure 5.3 shows intelligibility scores for all conditions and experiments. Approximately equal overall intelligibility was achieved for intact and vocoded SiB due to our choice of SNRs for these conditions, based on extensive piloting. This allowed small differences in intelligibility to be normalized without loss of statistical power. Overall intelligibility was normalized to 60% for intact and vocoded SiB, and to 90% for intact and vocoded SiQuiet, respectively (as described in Section 5.2.5), before examining the effects of vocoding on voicing, POA, and MOA confusion patterns.

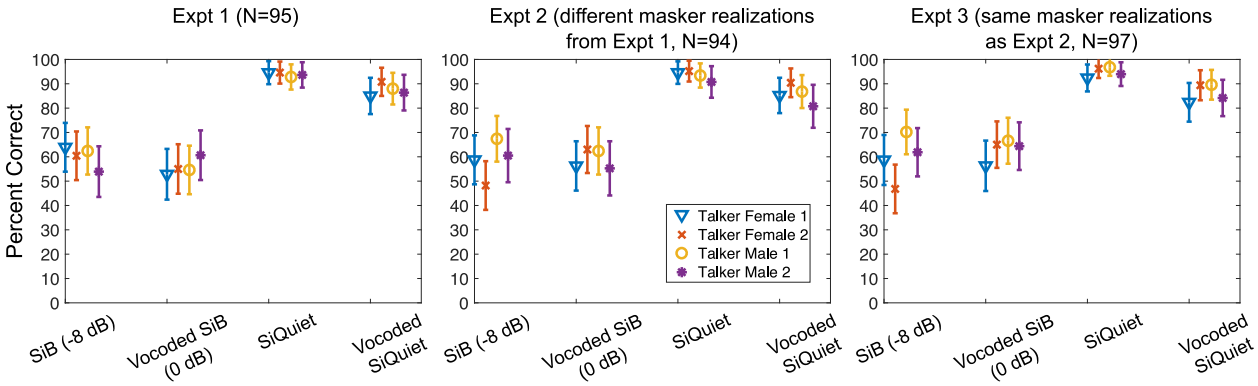

Figure 5.3. : Overall intelligibility (mean and standard error) measured in the online consonant identification experiments for the different conditions and talkers. Approximately equal overall intelligibility was achieved across intact and vocoded SiB, and across intact and vocoded SiQuiet.

Given that our data were collected online, a few different data quality checks were conducted. The first of these 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. As shown in Figure 5.4, percent errors in our data fall outside the distributions expected from random confusions. This result suggests that the error patterns in our data have a non-random structure, which supports the validity of our online-collected data. Moreover, there are small differences in the percent errors for voicing, place, and manner between intact and vocoded SiB, and also between intact and vocoded SiQuiet. These differences were further investigated by quantifying full consonant confusion matrices for the voicing, place, and manner categories and examining the differences in these matrices across intact and vocoded conditions (Figs. 5.6, 5.7, 5.8, and 5.9). This allowed us to obtain a richer characterization of the error patterns in consonant categorization (i.e., when an error was made, what consonant was reported instead of the consonant presented, and what proportion of trials was the alternative reported) compared to the percent error scores shown in Figure 5.4.

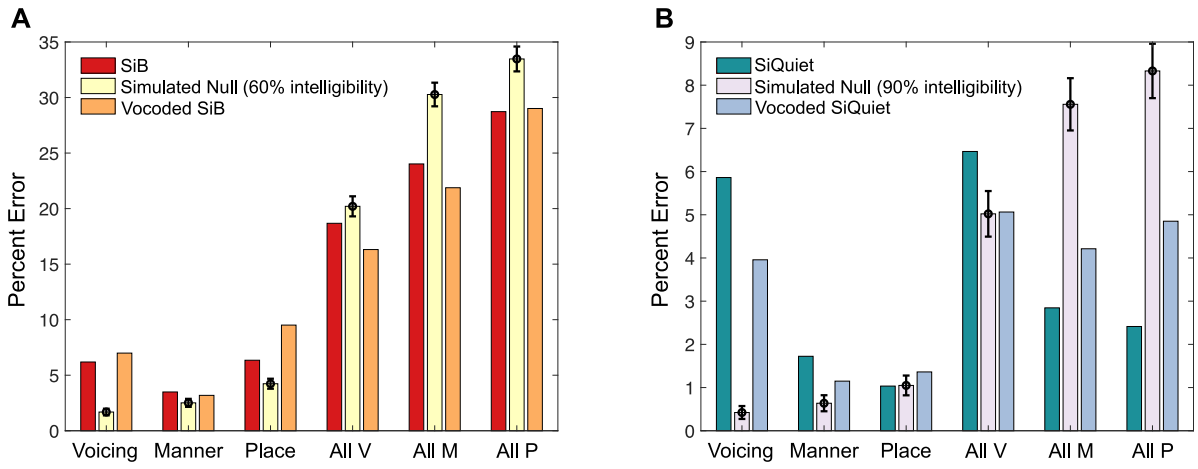

Figure 5.4. : Percent errors (mean and standard deviation from Experiment 1) for each phonetic category for intact and vocoded SiB (Panel A), and intact and vocoded SiQuiet (Panel B). The labels "Voicing", "Manner", and "Place" correspond to when the consonant reported differed from the consonant presented only in voicing, manner of articulation (MOA), or place of articulation (POA), respectively. "All V", "All M", and "All P" correspond to when the consonant reported differed from the consonant presented in at least voicing, MOA, or POA, respectively (e.g., "All V" includes the following types of errors: (i) voicing only, (ii) voicing and MOA simultaneously, (iii) voicing and POA simultaneously, and (iv) voicing, MOA, and POA simultaneously). The expected distribution of errors under the null hypothesis of random confusions was generated separately for Panels A and B, and with 1000 realizations each. Each realization of each null distribution was produced by generating a Bernouilli trial with "success" probability = 60% for Panel A, or 90% for Panel B, followed by uniform-random selection of a different consonant from what was presented if the trial outcome was "failure".

To further test data quality, consonant confusions for the SiSSN condition were compared with previous lab-based findings, since speech-shaped stationary noise is a commonly used masker in the phoneme confusions literature. 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. Our online-collected data for SiSSN (Fig. 5.5A) closely replicate that key trend for the groups they identified, after matching the SNR they used. Moreover, using a hierarchical clustering analysis (Ward Jr, 1963) of the consonant confusion matrix (pooled over samples) for SiSSN, perceptual "clusters" (i.e., sets where one consonant is confused most with another in the same set) were identified (shown as a dendrogram plot in Fig. 5.5B). The clusters identified here closely replicate the lab-based clustering results of Phatak and Allen, 2007, further supporting the validity of our online data.

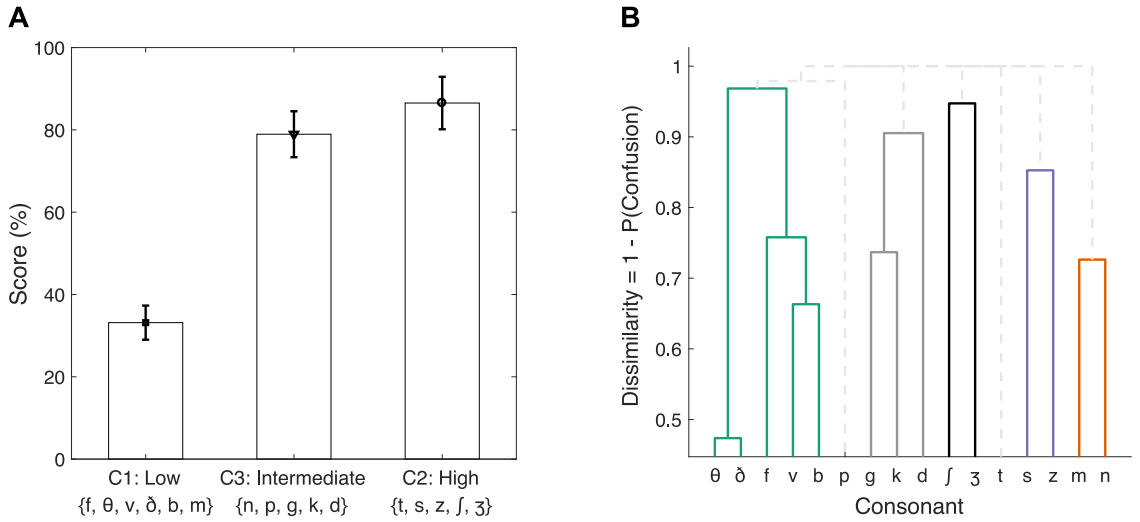

Figure 5.5. : Consonant groups (Panel A) and confusion clusters (Panel B) for the speech in speech-shaped stationary noise (SiSSN) data from Experiment 1. Panel A shows recognition scores for our SiSSN data for the three groups of consonants, "C1", "C2", and "C3" that Phatak and Allen, 2007 identified as having low, high, and intermediate recognition scores, respectively in speech-shaped noise (for a given overall intelligibility). Panel B shows the perceptual "clusters" (visualized as a dendrogram plot) identified with our SiSSN data. Each cluster is a set where one consonant is confused most with another in the same set. Clusters with greater than 3% probability of confusion share a color. For example, /θ/ and /ð/ form a cluster because they are more confused with each other than with the other consonants; moreover, while /θ/ and /ð/ are less confused with the cluster comprising /f/, /v/, and /b/ than with each other, they are even less confused with all the remaining consonants.

After verifying data quality, the hypothesis that confusion patterns would be the same for intelligibility-matched intact and envelope-vocoded speech in babble was tested. Figure 5.6 shows the results for voicing confusions. Vocoding altered the voicing percept for speech in babble by changing subject bias relative to an ideal observer. In particular, there was a greater tendency in the vocoded (versus intact) condition for the subject to be biased towards reporting an unvoiced consonant despite envelope and place cues being largely preserved. A detection-theoretic analysis (see Section 5.2.7) was used to quantify the decision boundary for the average subject's perceptual decision on whether or not to reject the null hypothesis that an unvoiced consonant was presented. The bias or shift in this boundary relative to an unbiased ideal observer was then quantified and compared between intact and vocoded conditions. Intact-to-vocoded bias changes were found to be about 40%, 24%, and 19% in Experiments 1, 2, and 3, respectively. Thus, the result that vocoding biases voicing percept towards unvoiced consonants is replicated across Experiments 1–3, supporting the idea that this bias effect is robust and generalizes across different babble instances. Note that the bias change between the intact and vocoded SiB conditions was observed even though the percent correct scores for the unvoiced and voiced categories were similar across these conditions (i.e., the diagonal entries in Fig. 5.6 are zero after statistical testing; for the precise number of errors, see Fig. 5.4). That is, while there were not a significantly different number of voicing errors after vocoding, the errors in the vocoded condition were biased towards reporting an unvoiced consonant even when a voiced consonant was presented. The errors in the intact condition were biased in the opposite direction, causing the total number of errors to be similar across the two conditions. This result suggests that the original TFS conveys important voicing information even when envelope cues are intact, since degrading the TFS led to a greater bias towards the percept of unvoiced consonants. This result also demonstrates that independent insight can be gained into the role of TFS cues from analyzing error patterns in consonant categorization rather than just examining transmission scores for the different phonetic categories.

Figures 5.7 and 5.8 show the results from testing our hypothesis for POA and MOA confusions. Although significant differences were found in the POA and MOA confusion patterns between intact and vocoded SiB, the results were not consistent across Experiments

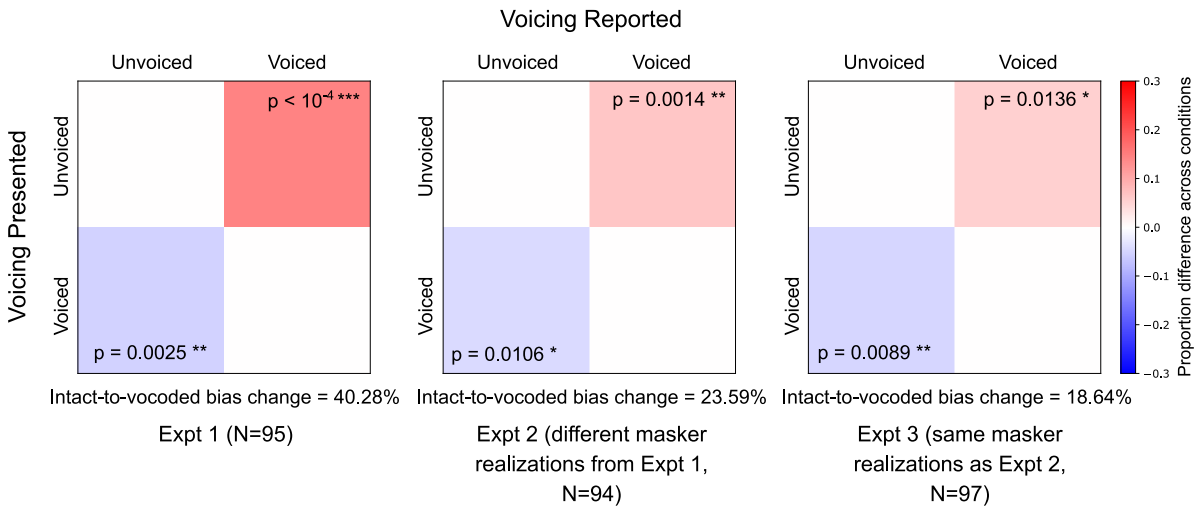

Figure 5.6. : Voicing confusion-matrix differences (pooled over consonants and samples) between intact and vocoded SiB conditions (SiB - Vocoded SiB). Overall intelligibility was matched at 60% before computing the differences across conditions. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR). Uncorrected p-values are also indicated for the individual matrix entries.

1 and 2, which used different instances of babble ($R^2 = 2 \times 10^{-6}$, $p = 0.99$ for POA, and $R^2 = 0.03$, $p = 0.44$ for MOA). The results were replicated only when the stimuli were kept constant, between Experiments 2 and 3 ($R^2 = 0.85$, $p = 3.77 \times 10^{-13}$ for POA, and $R^2 = 0.94$, $p = 1.44 \times 10^{-12}$ for MOA). Note that the differences in POA and MOA confusions between intact and vocoded SiB could be due to either TFS or masker-instance effects; our goal behind using different masker instances across Experiments 1 and 2 was to extract those effects that are not instance-specific and rather due to a true effect of TFS. However, because the confusion-matrix differences for POA and MOA were not replicated across different masker instances, it is not possible to disambiguate between these two effects here. Nevertheless, the fact that these results did not generalize across different babble instances suggests that any effects of TFS on POA and MOA reception are weak when compared to differences across different samples of babble.

To test whether TFS conveys phonetic content beyond what is conveyed by envelopes for intact speech in quiet, the effect of vocoding on consonant confusion patterns for the SiQuiet condition was examined. The results (Fig. 5.9) indicate no significant effects of degrading TFS on either voicing, POA, or MOA confusions in quiet.

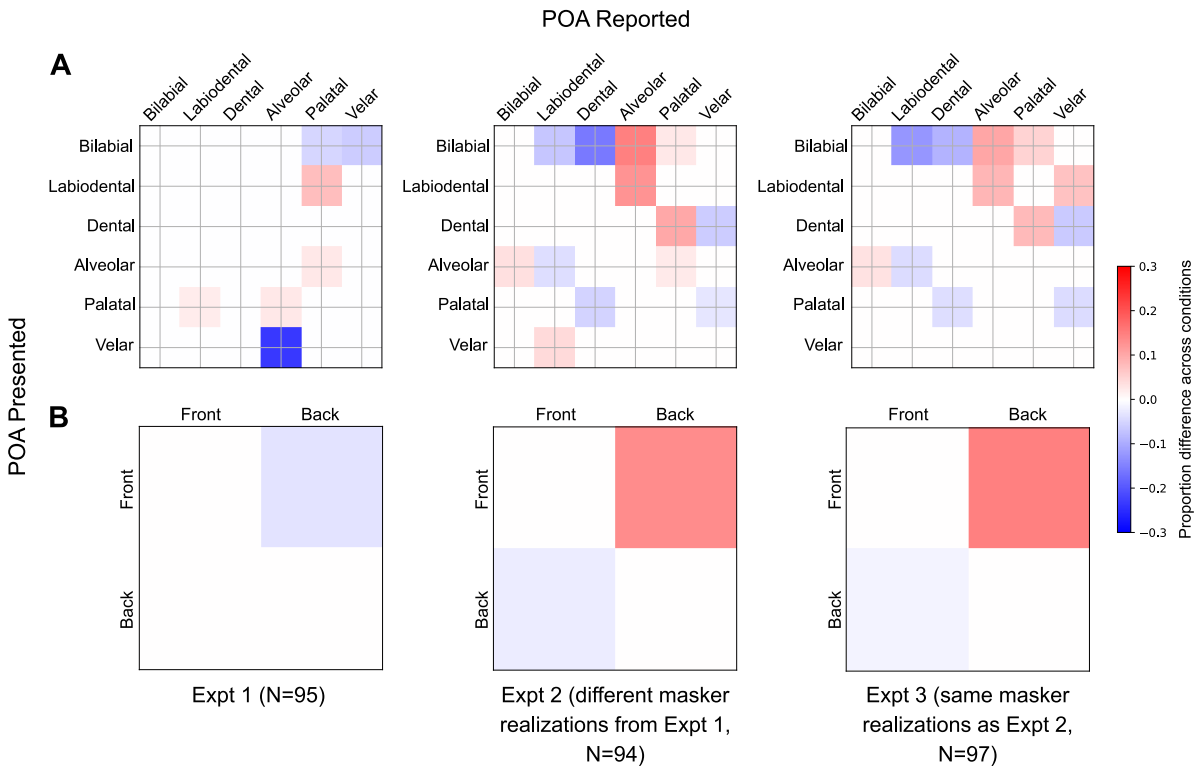

Figure 5.7. : **POA confusion-matrix differences (pooled over consonants and samples) between intact and vocoded SiB (SiB - Vocoded SiB).** Overall intelligibility was matched at 60% before computing the differences across conditions. Panel A shows full (5x5) matrices, whereas Panel B shows simplified (binary) matrices after collapsing over front versus back places of articulation. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR).

5.4 Discussion

The present study examined the influence of TFS on consonant confusion patterns by degrading TFS using high-resolution vocoding while controlling intelligibility to match that for intact stimuli. The results suggest that TFS is used to extract voicing content for intact speech in babble (i.e., even when redundant envelope cues are available). Moreover, this finding generalized across different babble instances. However, there were no significant vocoding effects on consonant confusions in quiet even after pooling data across all experiments; instead, overall intelligibility for Vocoded SiQuiet was ∼90%.

The finding that TFS conveys voicing information beyond what is conveyed by envelopes for intact speech in babble is previously unreported to the best of our knowledge. This result deviates from the commonly held view that envelopes convey most speech content (Shannon

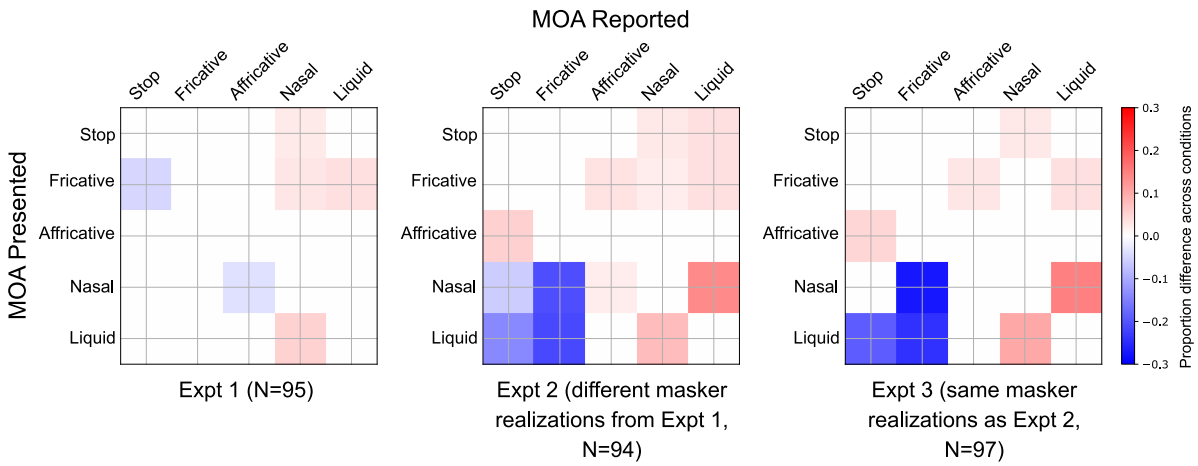

Figure 5.8. : **MOA confusion-matrix differences (pooled over consonants and samples) between intact and vocoded SiB (SiB - Vocoded SiB).** Overall intelligibility was matched at 60% before computing the differences across conditions. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR).

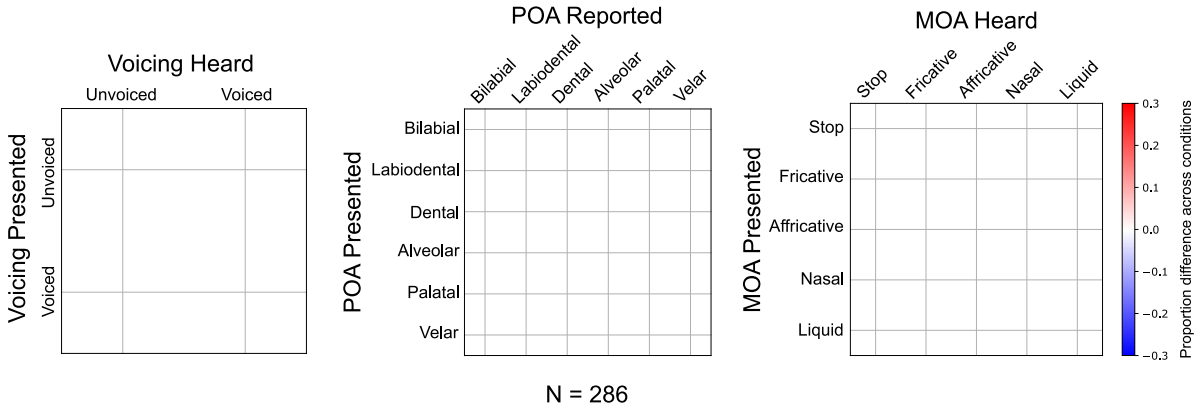

Figure 5.9. : **Voicing, POA, and MOA confusion-matrix differences (pooled across all experiments, consonants, and samples) between intact and vocoded speech in quiet (SiQuiet - Vocoded SiQuiet).** Overall intelligibility was matched at 90% before computing the differences across conditions. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR).

et al., 1995). Several acoustic cues have been implicated in the categorization of consonant voicing, such as voice onset time (VOT), fundamental frequency at the onset of voicing (onset F0), and the relative amplitude of any aspiration noise in the period between the burst release and the onset of voicing (Francis et al., 2008). Of these, VOT appears to be the dominant cue in quiet (Francis et al., 2008). However, listeners shift reliance to onset F0 when the VOT is ambiguous in the presence of noise (Holt et al., 2018; Winn et al., 2013). Our finding that vocoding alters the voicing percept in noise, but not quiet, is consistent with this result

from the cue-weighting literature, and can be attributed to impaired F0 cues resulting from TFS degradation in the vocoded (versus intact) SiB condition. Indeed, voiced sounds (unlike unvoiced) have quasi-periodic acoustic energy reflecting the quasi-periodic vibrations of the vocal folds; this periodicity has a fundamental frequency (F0) that is perceived as pitch (S. Rosen, 1992). Our finding that TFS is used to extract voicing content for intact speech in babble is consistent with the view that the pitch of complex sounds (with resolved harmonics) is coded either via TFS (Meddis & O'Mard, 1997; B. C. Moore et al., 2006), or a combination of TFS and tonotopic place (A. J. Oxenham et al., 2004; Shamma & Klein, 2000). Indeed, psychophysical studies have found that melody perception (B. C. Moore & Rosen, 1979) and F0 discrimination (Bernstein & Oxenham, 2006; Houtsma & Smurzynski, 1990) are both better when conveyed by low-frequency resolved harmonics where the auditory nerve can robustly phase lock to the TFS (Johnson, 1980; Verschooten et al., 2015). Our results from directly manipulating TFS cues also corroborate previous correlational work relating model auditory-nerve TFS coding and voicing reception in noise (Swaminathan & Heinz, 2012). Other previous studies have suggested that low-frequency speech information is important for voicing transmission (Li & Loizou, 2008), but the experimental manipulations they used altered multiple cues including low-frequency place cues, slower envelopes, and possibly the masking of more basal regions by upward spread; this makes it difficult to unambiguously attribute their results to the role of TFS. In contrast, by isolating TFS manipulations in the present study, these limitations were overcome.

The current study found a strong babble-instance effect on POA and MOA confusion patterns. The effects of vocoding on these confusion patterns were not replicated when babble instances differed between Experiments 1 and 2, but were replicated when instances were fixed across Experiments 2 and 3. One explanation for the differences in confusion patterns across varying babble instances is that even though the average masker modulation spectrum was kept constant (the envelope of babble is dominated by low modulation frequencies; Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021), there can be small variations in the modulation spectrum of the babble masker across instances within any given short time window. This, in turn, can cause variations in modulation masking across instances due to the relatively short duration of each consonant. Although not directly tested, hints

of such effects of short-term envelope statistics were also found in Phatak and Grant, 2012, where alterations of masker modulations produced less predictable effects on consonants than vowels. In the present study, masker-instance effects on consonant perception were explicitly measured and confirmed. The role of short-term masker statistics should be further examined in future studies, perhaps using computational modeling to predict instance effects on consonant confusions from variations in modulation masking across short masker instances. Indeed, psychoacoustic literature on speech-in-noise perception (Bacon & Grantham, 1989; Stone & Moore, 2014), neurophysiological studies using EEG (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021), and and the success of current speech intelligibility models (Dubbelboer & Houtgast, 2008; Relaño-Iborra et al., 2016) show that modulation masking (i.e., masking of the internal representation of temporal modulations in the target by distracting fluctuations from the background) is a key contributor to speech perception in noise.

The fact that no significant vocoding effects on consonant confusions in quiet were found, even after pooling data across experiments, is consistent with previous behavioral studies that suggested that speech content in quiet is mostly conveyed by envelopes (Elliott & Theunissen, 2009; Shannon et al., 1995), and with the success of envelope-based cochlear implants in quiet backgrounds (B. S. Wilson & Dorman, 2008). However, our finding that voicing cues are degraded in vocoded (versus intact) SiB has implications for current cochlear implants that do not appear to be able to provide usable TFS cues (Heng et al., 2011; Magnusson, 2011), because babble is a masker that is ubiquitous in everyday listening environments. Indeed, multi-talker babble, which has modulations spanning the range of modulations in the target speech, is a more ecological masker than either stationary noise (which has predominantly high-, but not low-frequency modulations as are present in speech) or even narrow-band syllabic-range AM modulations imposed on stationary noise (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021), as were used in previous studies (Gnansia et al., 2009; Holt et al., 2018; Swaminathan & Heinz, 2012; Winn et al., 2013). In addition to our finding here that TFS can convey important voicing cues, there is evidence from previous studies that TFS can also aid in source segregation (Darwin, 1997; Micheyl & Oxenham, 2010; A. J. Oxenham & Simonson, 2009), which can lead to stronger representation of attended-speech envelopes

in the brain that predicts intelligibility (Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021). The effect of TFS on segregation is reflected in the present study too, where the SNR for vocoded SiB had to be increased by 8 dB relative to intact SiB in order to match their respective overall intelligibility values. Taken together, these results suggest that patients with cochlear implants may benefit from improvements that allow these implants to provide usable TFS cues for speech recognition in everyday listening environments with multiple talkers or sound sources (Heng et al., 2011; Magnusson, 2011). This finding should be further examined in future studies using clinical populations.

One limitation of the current study is the use of isolated CV syllables (e.g., /ba/) rather than words commonly used in the English language (e.g., bat) to measure consonant categorization. However, the use of CV syllables allowed us to easily standardize the context across the different consonants (i.e., the vowel used was always /a/, and it always occurred after the consonant), thereby eliminating any confounds between the consonant used and condition effects (i.e., the effect of vocoding). One issue with standardizing context in this manner is that the effect of TFS may depend on the specific context used (i.e., C/a/). Thus, future work should explore whether such interaction effects exist. That being said, the C/a/ syllables were not presented in complete isolation; instead, a carrier phrase was used to help guide the listeners' attention in a manner similar to natural running speech.

5.5 Acknowledgments

This research was supported by grants from the National Institutes of Health [F31DC017381 (to V.V.), R01DC009838 (to M.G.H.), and R01DC015988 (to B.G.S.-C.)] and Office of Naval Research [ONR N00014-20-12709 (to B.G.S.-C.)]. The authors would like to thank Hari Bharadwaj for access to online psychoacoustics infrastructure (https://snaplabonline.com; Mok et al., 2021). We also thank Agudemu Borjigan, Andrew Sivaprakasam, François Deloche, Hari Bharadwaj, Ivy Schweinzger, Ravinderjit Singh, and Satyabrata Parida for valuable feedback on an earlier version of this chapter. Finally, we thank Christian Lorenzi, Brian Moore, and an anonymous reviewer for their insightful and helpful comments.

5.6 Supplementary Information

For completeness, the raw confusion matrices for all conditions and experiments are shown in Figure 5.10.

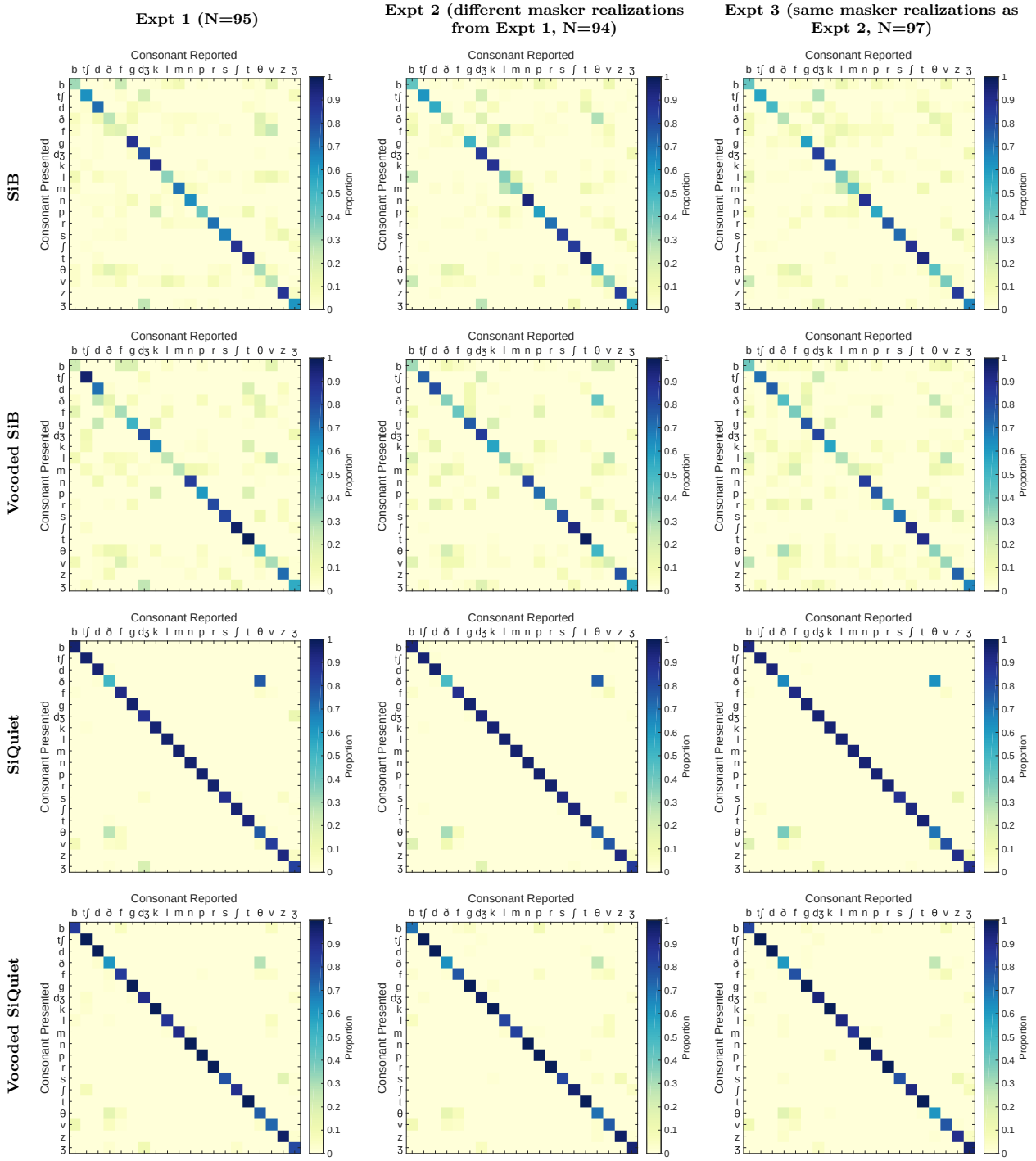

Figure 5.10. : **Raw confusion matrices for all conditions and experiments (pooled over samples).** Overall intelligibility was 60% for the SiB and Vocoded SiB conditions, and 90% for the SiQuiet and Vocoded SiQuiet conditions.

6. CONCLUSIONS

This dissertation addresses fundamental questions about the neurophysiological mechanisms supporting speech intelligibility in everyday listening environments with interfering sound sources and distortions. Examples of such environments include crowded streets/rooms with multiple talkers or other sources of noise, reverberant rooms and stairwells, and non-linear distortion produced by certain speech-processing algorithms (e.g., wind-noise reduction; Launer et al., 2016) in cell phones, hearing aids, and cochlear implants. In particular, this dissertation investigated the mechanisms underlying scene segregation and selective attention in such everyday environments, and the roles of different acoustic cues in supporting these processes and in conveying phonetic content. The main findings of this dissertation and the gaps they address are summarized in this concluding chapter. Additionally, the significance of these findings is briefly discussed, and some future research directions are proposed.

6.1 Summary of Main Findings

6.1.1 Cortical signatures of speech-on-speech selective attention

The neural mechanisms that underlie selective attention to speech in the presence of other competing talkers—an ability that is critical for everyday communication—are poorly understood. Results from noninvasive electrophysiology show that low-frequency cortical responses preferentially track the envelopes of attended speech in a mixture of sources (Ding & Simon, 2012; J. A. O'Sullivan et al., 2014). In contrast, invasive studies show that the power fluctuations in the high-gamma band preferentially phase lock to attended speech more than ignored speech (Golumbic et al., 2013; Mesgarani & Chang, 2012). However, no prior noninvasive studies had reported how the full complement of canonical brain oscillations (Buzsáki & Draguhn, 2004) track speech sounds in a mixture of competing sources, when attention is selectively directed to one source stream. We addressed this gap by using electroencephalography (EEG) to systematically study the attention-dependent relationships between input speech envelopes and the neural response in different canonical frequency bands (Chapter 2). Consistent with previous literature, we found that brain

rhythms in the low-frequency delta and theta bands (corresponding to the prosodic and syllabic rates, respectively) show more phase synchrony with speech envelopes when speech is attended versus ignored (Ding & Simon, 2012; J. A. O'Sullivan et al., 2015). Additionally, we found that the slow power fluctuations of the gamma band selectively track the low-frequency envelopes of attended speech, a result that had previously been reported only with invasive recordings (Golumbic et al., 2013; Mesgarani & Chang, 2012). This result supports the view that gamma activity plays an important role in the underlying physiological computations that support selective listening (Ribary, 2005; Tallon-Baudry & Bertrand, 1999; X.-J. Wang, 2010) and demonstrates that non-invasive EEG can be used to measure these effects. Our results also showed individual differences in the overall magnitude of attentional enhancement of speech-envelope coding, which suggests that difficulty listening in noise can stem from purely top-down factors such as attention. Finally, we also found individual differences in the speech and EEG features (i.e., channels and frequency bands) that are most informative about attentional focus.

6.1.2 Neurophysiological mechanisms of scene segregation

Speech, like any acoustic signal, can be decomposed into a slowly varying temporal envelope (or modulation) and rapidly varying temporal fine structure (TFS). A leading hypothesis in the field is that the signal-to-noise ratio (SNR) in the internal representation of envelopes of a target speech source (relative to interfering sounds) predicts speech intelligibility (Bacon & Grantham, 1989; Dubbelboer & Houtgast, 2008; Relaño-Iborra et al., 2016; Stone & Moore, 2014). We tested this hypothesis using EEG and simultaneous speech-intelligibility measurements over a range of realistic background noises and distortions (Chapter 3). Our results provide neurophysiological evidence for modulation masking in that they showed that EEG-based target-speech envelope encoding fidelity is shaped by masker envelopes in a modulation-frequency-specific manner, and that this net target-envelope coding predicts intelligibility across all tested conditions. We also found that the modulation frequencies that contribute most to speech intelligibility in everyday listening could lie anywhere in the full continuum from slow prosodic fluctuations to fast pitch-range fluctuations. Another

important finding in this study is that envelope coding in central auditory neurons is shaped not only by input envelopes, but also by the TFS. Indeed, when there are competing sound sources, TFS cues can help segregate a target speech source from distracting sounds (Darwin, 1997; A. J. Oxenham & Simonson, 2009), which in turn facilitates attentional boosting of target-speech envelopes (Viswanathan et al., 2019). Our results thus elucidate the acoustic cues and scene-analysis mechanisms that shape the neural processing of a target speech sound to predict speech intelligibility. Our results are consistent with the theory that temporal coherence of sound elements across envelopes and/or TFS influences scene analysis and attentive selection (Elhilali et al., 2009). Through this mechanism, those envelope and TFS frequencies of the target that are temporally coherent with components of the masker may get bound together (i.e., a failure of source segregation), which in turn can lead to degraded target representation and perceptual modulation masking at those specific frequencies. A conceptual model of speech intelligibility that integrates these ideas was proposed.

6.1.3 Computational modeling of speech categorization to test fundamental theories of auditory scene analysis

Temporal coherence of sound fluctuations across different frequency channels is thought to be important for auditory scene analysis (Apoux & Bacon, 2008; Darwin, 1997; Elhilali et al., 2009; Schooneveldt & Moore, 1987). Prior studies on the neural bases of temporal-coherence processing mostly focused on cortical contributions (Elhilali et al., 2009; J. A. O'Sullivan et al., 2015; Teki et al., 2013). However, results from single-unit measurements and modeling suggest that physiological correlates of comodulation masking release (CMR)—a temporal-coherence-based phenomenon—may be present as early as brainstem (Meddis et al., 2002; Pressnitzer et al., 2001). Prior studies of temporal-coherence processing also mostly used simple non-speech stimuli. Thus, the theory that speech understanding in noise may be shaped by aspects of temporal-coherence processing that exist in early auditory areas had not been previously tested. To address this gap, we used a combination of computational modeling and a psychophysical consonant-identification experiment (Chapter 4). We constructed separate models for (i) within-channel modulation masking (Relaño-Iborra et al., 2016), and (ii) across-channel temporal-coherence processing mirroring the computations that exist in the

cochlear nucleus (i.e., the first auditory region that can support cross-channel processing over a wide frequency range) (Pressnitzer et al., 2001) combined with within-channel modulation masking. Crucially, we predicted confusion patterns in consonant categorization using each of these models. Consonant confusions provide a richer characterization of error patterns in speech categorization compared to percent-correct scores; thus, the use of confusion data provides independent constraints on our understanding of scene-analysis mechanisms beyond what overall intelligibility can provide. Despite this, confusion patterns had not been utilized in prior studies of scene analysis. Here, by comparing model predictions to measured consonant confusions across different listening conditions, we found that across-channel temporal-coherence processing strongly influences consonant categorization across diverse noises and distortions, and that physiological computations that exist as early as the cochlear nucleus can contribute significantly to temporal-coherence-based scene analysis.

6.1.4 Roles of different acoustic cues in conveying speech content

Behavioral experiments in quiet backgrounds suggest a fundamental dichotomy in speech perception, with envelopes conveying most speech content and TFS conveying other sound attributes such as fundamental frequency (F0) (B. C. Moore et al., 2006; Shannon et al., 1995). However, TFS can also influence speech intelligibility in noise, in particular, by supporting scene segregation (Darwin, 1997; A. J. Oxenham & Simonson, 2009; Viswanathan, Bharadwaj, Shinn-Cunningham, & Heinz, 2021). However, whether TFS can contribute to speech-in-noise perception beyond supporting masking release, i.e., whether TFS can directly convey phonetic content, was poorly understood. Some prior studies examined this role of TFS when envelope cues were degraded, but did not address whether TFS cues are used for intact speech that has preserved envelope cues (Ardoint & Lorenzi, 2010; Sheft et al., 2008). Furthermore, prior studies used masking conditions that were not ecologically realistic (Ardoint & Lorenzi, 2010; Gnansia et al., 2009; S. Rosen, 1992; Sheft et al., 2008; Swaminathan & Heinz, 2012); maskers such as multi-talker babble, which has spectro-temporal characteristics similar to what may be encountered in realistic scenarios, had not been utilized to study this problem. To address these gaps, we measured consonant confusions for intact and envelope-vocoded speech in

ecologically relevant multi-talker babble (Chapter 5). We found that degrading TFS cues while controlling overall performance biased subjects towards the percept of unvoiced consonants, despite envelope and place cues being largely preserved. This result suggests that TFS is used to extract consonant voicing even when redundant envelope cues are available. This finding was replicated when the babble instances were varied across independent experiments, suggesting that the effects were robust to changes in the local statistics of the masker. This result deviates from the commonly held view that envelopes convey most speech content. Indeed, even though VOT appears to be the dominant cue for voicing in quiet (Francis et al., 2008), listeners shift reliance to fundamental frequency at the onset of voicing (onset F0) when VOT is ambiguous in the presence of noise (Holt et al., 2018; Winn et al., 2013). Our results are consistent with these previous findings from the cue-weighting literature, and suggest that TFS conveys voicing in multi-talker environments by contributing to the percept (Meddis & O'Mard, 1997; B. C. Moore et al., 2006) and discrimination (Bernstein & Oxenham, 2006; Houtsma & Smurzynski, 1990) of F0.

6.2 Significance

The findings of this dissertation can inform clinical applications, models of auditory scene analysis, attention, and speech intelligibility, and audio technologies, as described below.

6.2.1 Implications for clinical diagnostics, individualized interventions, and assistive listening devices

Difficulty understanding speech in noise can be caused by any of multiple factors. Although peripheral hearing impairments (including sensorineural loss, conductive loss, synaptopathy, and presbycusis) were traditionally thought to be the dominant contributor to hearing difficulties in noisy environments, deficits in top-down cognitive function such as selective attention (Dai et al., 2018; B. Shinn-Cunningham & Best, 2008) and working memory (Lunner, 2003) can also contribute. Indeed, the results in Chapter 2 show that even those with clinically normal hearing exhibit individual differences in the overall strength of selective attention to speech, and suggest that EEG measurements from an easier speech-based selective attention

task may be used to quantify the top-down attentional contribution to individual differences in speech intelligibility in adverse listening conditions (Choi et al., 2014). Moreover, the results in Chapter 3 establish that the internal representation of envelopes is fundamentally important for intelligibility across diverse realistic listening conditions, thereby lending support to envelope-coding metrics as assays of suprathreshold hearing health (Bharadwaj et al., 2015).

Characterizing, for each individual, the factors affecting speech intelligibility (e.g., audibility, temporal envelope coding, and attention) would allow for more effective, targeted, individualized interventions to improve communication. For example, sensorineural hearing loss can be treated with hearing aids; alternatively, cochlear implants may be used if the hearing loss is severe to profound. On the other hand, those with poorer selective attention, either from the natural variation experienced by clinically normal-hearing individuals (Chapter 2), or due to peripheral hearing damage (Dai et al., 2018; B. Shinn-Cunningham & Best, 2008), may benefit from cognitive aural training. Indeed, approaches such as neurofeedback training of auditory selective attention show promise in enhancing speech-in-noise perception (Kim et al., 2021). EEG-guided hearing aids of the future represent another approach to mitigate selective attention issues (Fiedler et al., 2017; Fuglsang et al., 2017; J. O'Sullivan et al., 2017; Van Eyndhoven et al., 2017). Unlike traditional hearing aids that are used to improve audibility in patients with peripheral hearing loss, future EEG-guided hearing aids aim to mitigate attention deficits by performing attention decoding and selective amplification of the target sound source. Our graph-theoretic approach (Chapter 2) can be used identify the speech and EEG features (channels and frequency bands) that are most informative about an individual's attentional focus in such attention-guided hearing aids of the future. More generally, this approach can be used to identify optimal feature sets in any individualized brain-computer interface (BCI) that requires a compact feature set.

Our consistent finding that TFS influences target-speech intelligibility in the presence of interfering sounds (Chapters 3 and 5) has implications for modern cochlear implants that are used to treat severe to profound sensorineural hearing loss. This is because cochlear implants do not appear to be able to provide usable TFS cues (Heng et al., 2011; Magnusson, 2011). Degradation to TFS (while preserving place and envelope cues) can reduce speech intelligibility by as much as $\sim$55% in ecologically relevant multi-talker environments (Chapter 3), and an $\sim$8

dB SNR increase is needed for TFS-degraded stimuli relative to intact stimuli in order to match their respective overall intelligibility values (Chapter 5). This is because TFS cues can support source segregation (Darwin, 1997; A. J. Oxenham & Simonson, 2009), thereby influencing the encoding of attended speech in the central auditory system (Chapter 3). In addition to this role of TFS, it also conveys speech content (in particular, voicing) for intact speech (i.e., with preserved envelope cues) in ecologically relevant multi-talker environments (Chapter 5). In particular, there is a greater tendency in the TFS-degraded (versus intact) condition to be biased towards reporting an unvoiced consonant as being heard, despite envelope and place cues being largely preserved. Taken together, these findings have implications for cochlear implants and other assistive listening devices.

6.2.2 Implications for models of auditory scene analysis, attention, and speech intelligibility

The auditory periphery can convey information to the central nervous system through both place and temporal coding (B. C. Moore, 2012). Classic models of speech intelligibility, i.e., the Articulation Index (AI; ANSI, 1969) and Speech Intelligibility Index (SII; ANSI, 1997), emphasized audibility and the SNR of signal components at different cochlear places. However, these approaches fail in many complex listening conditions (Kryter, 1962; Steeneken & Houtgast, 1980). More recent models consider temporal aspects of coding. Since temporal modulations convey important information about speech content (Shannon et al., 1995), the Speech Transmission Index (STI; Steeneken & Houtgast, 1980) was developed to predict speech intelligibility from how much temporal modulations of clean speech in different audio frequency bands are degraded in the presence of noise and/or distortion. However, the STI does not explicitly consider noise modulations, which may interfere with speech intelligibility, and fails for non-linear distortions, including those typical of hearing-aid processing and cell-phone denoising algorithms (Jørgensen & Dau, 2011).

In keeping with behavioral studies that suggested that modulation masking may be a key contributor to speech understanding in noise (Bacon & Grantham, 1989; Stone & Moore, 2014), current intelligibility models are based on the fidelity (i.e., SNR) of the internal representation of temporal modulations in the target relative to those from the background

(which contains inherent distracting fluctuations), and have been successful over a wide range of listening conditions (Dubbelboer & Houtgast, 2008; Jørgensen et al., 2013; Relaño-Iborra et al., 2016). The study presented in Chapter 3 was the first, to the best of our knowledge, to validate these models with neurophysiological evidence. Specifically, we showed that the neural envelope-domain SNR in target-speech encoding, which is shaped by masker modulations, predicts intelligibility over a range of realistic interfering sounds and linear and non-linear distortions that are unseen by the predictive model. The findings of this dissertation, however, also suggest that TFS can influence speech intelligibility. In particular, we found that TFS can convey important speech content in everyday listening situations even when intact envelope cues are available (Chapter 5), and that TFS influences the internal representation of attended-speech envelopes that predicts intelligibility (Chapter 3). However, no models consider these roles of TFS in speech intelligibility.

Our finding that both modulation masking and TFS shape target-speech envelope encoding in the brain, which predicts intelligibility, is in line with the temporal coherence theory, thought important in both auditory (Elhilali et al., 2009) and visual (Singer & Gray, 1995) scene analysis. Accordingly, we proposed that envelope and/or TFS components that fluctuate coherently are perceptually grouped together, and that masker elements that are temporally coherent with target speech perceptually interfere even when they are in a different frequency channel from the target. Chapter 3 describes our conceptual model of speech intelligibility that integrates these ideas as well as the role of selective attention. In Chapter 4, we directly tested the role of across-channel temporal-coherence processing in consonant categorization using physiologically plausible computational modeling of within-channel modulation masking and temporal-coherence-based across-channel modulation interference. Our modeling results suggest that temporal-coherence processing shapes confusion patterns in speech categorization across diverse listening conditions and that this processing may start as early as brainstem.

Our psychophysical, neurophysiological, and modeling results across Chapters 2, 3, 4, and 5 can inform future modeling studies of auditory scene analysis, attention, and speech intelligibility. Crucially, our series of experiments also helps bridge the speech-intelligibility modeling literature with the separate literature on binding, source segregation, and attention,

to provide integrative insight into how the acoustic cues that are often considered for predicting speech intelligibility contribute to scene analysis and target selection.

6.2.3 Implications for other audio technologies

Intelligibility models form the basis for evaluating speech-denoising strategies in audio technologies such as cell phones, cochlear implants, and hearing aids (D. Wang & Chen, 2018). However the models most commonly used for this purpose (e.g., short-time objective intelligibility or STOI; Taal et al., 2011) have known limitations. These limitations are addressed by more recent intelligibility models based on within-channel modulation masking (Jørgensen et al., 2013; Relaño-Iborra et al., 2016) as well as the models proposed in this dissertation (Chapters 3 and 4). Our proposed models are in fact more accurate than these recent models because we account for across-channel interference in addition to modeling within-channel effects. Indeed, the addition of across-channel temporal-coherence processing to our model in Chapter 4 significantly improved predictions of confusion patterns in speech categorization across all tested conditions (including vocoded stimuli; Steinmetzger et al., 2019). In addition to their use in evaluating speech processing strategies, accurate, physiologically realistic models of scene analysis and intelligibility can also guide source separation algorithms in audio technologies (i.e., computational auditory scene analysis; D. Wang and Brown, 2006) as an alternative to black-box deep-learning-based approaches. A further application of our proposed intelligibility models is to evaluate room acoustics designs (Houtgast & Steeneken, 1985).

6.3 Future Work

There are several avenues for future research into the mechanisms of speech intelligibility that go beyond the topics investigated in this dissertation and the specific approaches used here. One such avenue is to integratively study scene-analysis mechanisms beyond temporal-coherence processing (e.g., streaming mechanisms based on spatial location or frequency separation; Bregman, 1990) within the modeling framework proposed in this dissertation. Another direction is to investigate the relative contributions of higher-order cognitive processes

such as working memory (Füllgrabe & Rosen, 2016), categorical perception and short-term plasticity from changing sound statistics (Holt et al., 2018), predictive coding (Davis & Johnsrude, 2007; Hickok & Poeppel, 2007), context (acoustic, lexical, semantic, and linguistic) processing (McClelland et al., 2006; Stilp, 2020), and multisensory integration (Lee et al., 2019) to speech understanding. In pursuing either of these directions, multi-pronged approaches that use a combination of electrophysiology, behavioral, and modeling experiments to combine findings across different stages of input processing along the auditory pathway are especially useful to obtain integrative insight into the mechanisms of speech perception.

Although this dissertation used exclusively noninvasive human experiments, deep mechanistic insight can in fact be obtained from invasive neurophysiology using either clinical populations or animal subjects. Approaches such as stereoelectroencephalography (sEEG) and electrocorticography (ECoG) are routinely and safely used in epilepsy patients undergoing neurosurgery planning (Katz & Abel, 2019); these approaches afford greater spatial specificity and SNR than scalp EEG for studying neural correlates of multisensory perception using these populations (Herff et al., 2020). Animal neurophysiology is particularly useful for manipulating different internal structures and processes in order to understand the specific conditions under which the system "breaks"; thus, this approach allows for causal relationships to be investigated at various stages of input processing. Moreover, when used in conjunction with carefully designed psychophysics, invasive neurophysiology has even greater potential to inform theories and computational models of sensory processing, especially at the cellular and network scales (Meddis et al., 2010). In cases where the experimental approach is limited to noninvasive measurements (e.g., in non-clinical human populations), spatial resolution in neurophysiology may be improved without sacrificing temporal resolution by using high-density EEG and magnetoencephalography (MEG) recordings along with source localization techniques (Hämäläinen et al., 1993).

Future human studies in this area can also benefit from the use of intervention (versus correlational) study designs. Indeed, in addition to their use for investigating the efficacy of treatments in clinical populations, intervention designs can also be used in basic-science studies on the general population to infer causal relationships between different aspects of sensory processing (e.g., working memory, attention, and executive control) and speech intelligibility.

Results from intervention studies such as audiomotor closed-loop training (Whitton et al., 2017), neurofeedback training for selective attention (Kim et al., 2021), working memory training (Wayne et al., 2016), and neurostimulation (Deng et al., 2019) are also informative about what strategies work to improve speech comprehension in everyday listening. These studies may hence inform future aural training, either to augment treatments like hearing aids and cochlear implants for peripheral hearing damage (Zeng et al., 2011), or to treat central auditory processing disorders such as deficits in selective attention or working memory (e.g., in older adults and certain clinical populations) (Musiek & Chermak, 2013).

Along the same lines, another future direction is to study clinical populations in order to understand the effects that different kinds of peripheral hearing damage (B. C. Moore, 2007) and central auditory processing disorders (Musiek & Chermak, 2013) have on speech recognition in noise. Studying impaired systems can provide insight into the aspects of healthy function that matter for speech intelligibility in normal systems, and can also guide diagnostics and strategies to treat disordered function. For example, pathological neural encoding of sound can result from disorders such as synaptopathy (Bharadwaj et al., 2014), distorted tonotopy (Henry et al., 2016), and age-related changes in central auditory processing (Frisina, 2010), all of which have implications for everyday speech recognition.

REFERENCES

ANSI. (1969). S3.5-1969. *American National Standard Methods for the Calculation of the Articulation Index, Acoustical Society of America, New York*.

ANSI. (1997). S3.5-1997. *Methods for Calculation of the Speech Intelligibility Index, American National Standards Institute, New York*.

Apoux, F., & Bacon, S. P. (2008). Selectivity of modulation interference for consonant identification in normal-hearing listeners. *J Acoust Soc Am*, *123*(3), 1665–1672.

Ardoint, M., & Lorenzi, C. (2010). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hear Res*, *260*(1-2), 89–95.

Bacon, S. P., & Grantham, D. W. (1989). Modulation masking: Effects of modulation frequency, depth, and phase. *J Acoust Soc Am*, *85*(6), 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*, 289–300.

Bernstein, J. G., & Oxenham, A. J. (2006). The relationship between frequency selectivity and pitch discrimination: Effects of stimulus level. *J Acoust Soc Am*, *120*(6), 3916–3928.

Bharadwaj, H. M. (2018). SNAPsoftware/ANLffr: Software tools for electrophysiology from the Systems Neuroscience of Auditory Perception Lab. Available at https://github.com/SNAPsoftware/ANLffr. https://doi.org/10.5281/zenodo.1490918

Bharadwaj, H. M., Mai, A. R., Simpson, J. M., Choi, I., Heinz, M. G., & Shinn-Cunningham, B. G. (2019). Non-invasive assays of cochlear synaptopathy–candidates and considerations. *Neurosci*, *407*, 53–66.

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015). Individual differences reveal correlates of hidden hearing deficits. *J Neurosci*, *35*(5), 2161–2172.

Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2014). Rapid acquisition of auditory subcortical steady-state responses using multichannel recordings. *Clin Neurophysiol*, *125*(9), 1878–1888.

Bharadwaj, H. M., Verhulst, S., Shaheen, L., Liberman, M. C., & Shinn-Cunningham, B. G. (2014). Cochlear neuropathy and the coding of supra-threshold sound. *Front Syst Neurosci*, *8*, 26.

Bidelman, G. M. (2017). Communicating in challenging environments: Noise and reverberation. *The frequency-following response* (pp. 193–224). Springer.

Bokil, H., Purpura, K., Schoffelen, J.-M., Thomson, D., & Mitra, P. (2007). Comparing spectra and coherences for groups of unequal size. *J Neurosci Meth*, *159*(2), 337–345.

Boothroyd, A., Mulhearn, B., Gong, J., & Ostroff, J. (1996). Effects of spectral smearing on phoneme and word recognition. *J Acoust Soc Am*, *100*(3), 1807–1818.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *J Acoust Soc Am*, *84*(1), 101–114.

Börgers, C., Epstein, S., & Kopell, N. J. (2008). Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *Proc Natl Acad Sci USA*, *105*(46), 18023–18028.

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

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychol Res*, *78*(3), 349–360.

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *Neuroimage*, *172*, 162–174.

Bruce, I. C., Erfani, Y., & Zilany, M. S. (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.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J Acoust Soc Am*, *120*(6), 4007–4018.

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes. *Nat Rev Neurosci*, *13*(6), 407.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*(5679), 1926–1929.

Cannon, J., McCarthy, M. M., Lee, S., Lee, J., Börgers, C., Whittington, M. A., & Kopell, N. (2014). Neurosystems: Brain rhythms and cognitive processing. *Eur J Neurosci*, *39*(5), 705–719.

Cardin, J. A., Carlén, M., Meletis, K., Knoblich, U., Zhang, F., Deisseroth, K., Tsai, L.-H., & Moore, C. I. (2009). Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature*, *459*(7247), 663.

Carlyon, R. P., Buus, S., & Florentine, M. (1989). Comodulation masking release for three types of modulator as a function of modulation rate. *Hear Res*, *42*(1), 37–45.

Carney, L. H., Li, T., & McDonough, J. M. (2015). Speech coding in the brain: Representation of vowel formants by midbrain neurons tuned to sound fluctuations. *Eneuro*, *2*(4).

Chermak, G. D., & Musiek, F. E. (1997). *Central auditory processing disorders: New perspectives*. San Diego: Singular publishing group.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am*, *25*, 975–979.

Choi, I., Wang, L., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Individual differences in attentional modulation of cortical responses correlate with selective attention performance. *Hear Res*, *314*, 10–19.

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

Dai, L., Best, V., & Shinn-Cunningham, B. G. (2018). Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. *Proc Natl Acad Sci USA*, *115*(14), E3286–E3295.

Darwin, C. J. (1997). Auditory grouping. *Trends Cogn Sci*, *1*(9), 327–333.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hear Res*, *229*(1-2), 132–147.

Delgutte, B., & Kiang, N. Y. (1984). Speech coding in the auditory nerve: I. vowel-like sounds. *J Acoust Soc Am*, *75*(3), 866–878.

Deng, Y., Reinhart, R. M., Choi, I., & Shinn-Cunningham, B. G. (2019). Causal links between parietal alpha activity and spatial auditory attention. *Elife*, *8*, e51184.

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. *Eneuro*, *5*(2).

Di Liberto, G. M., Lalor, E. C., & Millman, R. E. (2018). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *Neuroimage*, *166*, 247–258.

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage*, *88*, 41–46.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neurosci Biobehav Rev*, *81*, 181–187.

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci USA*, *109*(29), 11854–11859.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J Neurosci*, *33*(13), 5728–5735.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Front Hum Neurosci*, *8*, 311.

Dobie, R. A., & Wilson, M. J. (1989). Analysis of auditory evoked potentials by magnitude-squared coherence. *Ear Hear*, *10*(1), 2–13.

Dobie, R. A., & Wilson, M. J. (1994). Objective detection of 40 Hz auditory evoked potentials: phase coherence vs. magnitude-squared coherence. *Electroencephalogr Clin Neurophysiol - Evoked Potentials Section*, *92*(5), 405–413.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, *85*, 761–768.

Dorman, M. F., Loizou, P. C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *J Acoust Soc Am*, *104*(6), 3583–3585.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am*, *95*(2), 1053–1064.

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proc Natl Acad Sci USA*, *111*(19), 7126–7131.

Dubbelboer, F., & Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *J Acoust Soc Am*, *124*(6), 3937–3946.

Dubno, J. R., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *J Acoust Soc Am*, *69*(1), 249–261.

Eckhorn, R., Reitboeck, H. J., Arndt, M., & Dicke, P. (1990). Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex. *Neural Comput*, *2*(3), 293–307.

Eggermont, J. J. (2006). Properties of correlated neural activity clusters in cat auditory cortex resemble those of neural assemblies. *Journal Neurophysiol*, *96*(2), 746–764.

Elhilali, M., Chi, T., & Shamma, S. A. (2003). A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech Commun*, *41*(2-3), 331–348.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, *61*(2), 317–329.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput Biol*, *5*(3), e1000302.

Engel, A. K., & Fries, P. (2010). Beta-band oscillations?signalling the status quo? *Curr Opin Neurobiol*, *20*(2), 156–165.

Ewert, S. D., & Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *J Acoust Soc Am*, *108*(3), 1181–1196.

Faust, K. (1997). Centrality in affiliation networks. *Soc Networks*, *19*(2), 157–191.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., & Obleser, J. (2017). Single-channel in-ear-eeg detects the focus of auditory attention to concurrent tone streams and mixed speech. *J Neural Eng*, *14*(3), 036020.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 1–32.

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in english. *J Acoust Soc Am*, *124*(2), 1234–1251.

Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, *291*(5508), 1560–1563.

Frisina, R. D. (2010). Aging changes in the central auditory system. *The Oxford handbook of auditory science: the auditory brain*, *2*, 418–438.

Fuglsang, S. A., Dau, T., & Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage*, *156*, 435–444.

Füllgrabe, C., & Rosen, S. (2016). Investigating the role of working memory in speech-in-noise identification for listeners with normal hearing. *Physiology, psychoacoustics and cognition in normal and impaired hearing* (pp. 29–36). Springer, Cham.

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2012). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Front Human Neurosci*, *6*.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–126.

Gilbert, G., & Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *J Acoust Soc Am*, *119*(4), 2438–2444.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat Neurosci*, *15*(4), 511–517.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res*, *47*(1), 103–138.

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*(6), 4023–4033.

Goh, K.-I., & Choi, I.-G. (2012). Exploring the human diseasome: The human disease network. *Brief Funct Genomics*, els032.

Gold, B., & Morgan, N. (2002). Vocoders. *Speech and audio signal processing: Processing and perception of speech and music* (pp. 431–447). John Wiley & Sons, Singapore.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*(5), 980–991.

Gorga, M. P., Neely, S. T., Bergman, B. M., Beauchaine, K. L., Kaminski, J. R., Peters, J., Schulte, L., & Jesteadt, W. (1993). A comparison of transient-evoked and distortion product otoacoustic emissions in normal-hearing and hearing-impaired subjects. *J Acoust Soc Am*, *94*(5), 2639–2648.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front Neurosci*, *7*, 267.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446–460.

Grant, K. W., & Seitz, P. F. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *J Acoust Soc Am*, *107*(2), 1000–1011.

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron*, *24*(1), 31–47.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *J Acoust Soc Am*, *87*(6), 2592–2605.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol*, *11*(12).

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev Mod Phys*, *65*(2), 413.

Hannan, E. J. (1970). Inference about spectra. *Multiple time series* (pp. 245–324). John Wiley & Sons, Hoboken, NJ.

Heinz, M. G., Colburn, H. S., & Carney, L. H. (2002). Quantifying the implications of nonlinear cochlear tuning for auditory-filter estimates. *J Acoust Soc Am*, *111*(2), 996–1011.

Heinz, M. G., & Swaminathan, J. (2009). Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *J Assoc Res Otolaryngol*, *10*(3), 407–423.

Heng, J., Cantarero, G., Elhilali, M., & Limb, C. J. (2011). Impaired perception of temporal fine structure and musical timbre in cochlear implant users. *Hear Res*, *280*(1-2), 192–200.

Henry, K. S., Kale, S., & Heinz, M. G. (2016). Distorted tonotopic coding of temporal envelope and fine structure with noise-induced hearing loss. *J Neurosci*, *36*(7), 2227–2237.

Herff, C., Krusienski, D. J., & Kubben, P. (2020). The potential of stereotactic-eeg for brain-computer interfaces: Current progress and future directions. *Front Neurosci*, *14*, 123.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nat Rev Neurosci*, *8*(5), 393–402.

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, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*(4108), 177–180.

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hear Res*, *366*, 50–64.

Hopkins, K., & Moore, B. (2010). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *J Acoust Soc Am*, *127*(3), 1595–1608.

Houtgast, T., & Steeneken, H. J. (1985). A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am*, *77*(3), 1069–1077.

Houtsma, A. J., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *J Acoust Soc Am*, *87*(1), 304–310.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J Acoust Soc Am*, *68*(4), 1115–1122.

Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J Acoust Soc Am*, *130*(3), 1475–1487.

Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *J Acoust Soc Am*, *134*(1), 436–446.

Joris, P. X., & Yin, T. C. (1992). Responses to amplitude-modulated tones in the auditory nerve of the cat. *J Acoust Soc Am*, *91*(1), 215–232.

Joris, P., Schreiner, C., Rees, A., et al. (2004). Neural processing of amplitude-modulated sounds. *Physiol Rev*, *84*(2), 541–578.

Kates, J. M. (2011). Spectro-temporal envelope changes caused by temporal fine structure modification. *J Acoust Soc Am*, *129*(6), 3981–3990.

Katz, J. S., & Abel, T. J. (2019). Stereoelectroencephalography versus subdural electrodes for localization of the epileptogenic zone: What is the evidence? *Neurother*, *16*(1), 59–66.

Khanna, S. M., & Leonard, D. G. (1982). Basilar membrane tuning in the cat cochlea. *Science*, *215*, 305–306.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & 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*(4), 2395–2405.

Kim, S., Emory, C., & Choi, I. (2021). Neurofeedback training of auditory selective attention enhances speech-in-noise perception. *Front Hum Neurosci*, *15*, 337.

Kim, S., Schwalje, A. T., Liu, A. S., Gander, P. E., McMurray, B., Griffiths, T. D., & Choi, I. (2020). Pre-and post-target cortical processes predict speech-in-noise performance. *Neuroimage*, 117699.

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with r* (Vol. 65). Springer.

Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Comput Biol*, *10*(12), e1003985.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *J Acoust Soc Am*, *34*(11), 1689–1697.

Kumar, G., Amen, F., & Roy, D. (2007). Normal hearing tests: is a further appointment really necessary? *J R Soc Med*, *100*(2), 66–66.

Lachaux, J., Rodriguez, E., Martinerie, J., & Varela, F. (1999). Measuring phase synchrony in brain signals. *Hum Brain Mapp*, *8*(4), 194–208.

Launer, S., Zakis, J. A., & Moore, B. C. (2016). Hearing aid signal processing. *Hearing aids*, 93–130.

Le Van Quyen, M., Foucher, J., Lachaux, J.-P., Rodriguez, E., Lutz, A., Martinerie, J., & Varela, F. J. (2001). Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony. *J Neurosci Meth*, *111*(2), 83–98.

Lee, A. K., Wallace, M. T., Coffin, A. B., Popper, A. N., & Fay, R. R. (2019). *Multisensory processes: The auditory perspective* (Vol. 68). Springer.

Li, N., & Loizou, P. C. (2008). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *J Acoust Soc Am*, *124*(6), 3947–3958.

Lin, F. R., Niparko, J. K., & Ferrucci, L. (2011). Hearing loss prevalence in the United States. *Arch Intern Med*, *171*(20), 1851–1853.

Llinás, R. R., Leznik, E., & Urbano, F. J. (2002). Temporal binding via cortical coincidence detection of specific and nonspecific thalamocortical inputs: A voltage-dependent dye-imaging study in mouse brain slices. *Proc Natl Acad Sci USA*, *99*(1), 449–454.

Loizou, P. C. (2013). *Speech enhancement: Theory and practice, part iii: Evaluation* (2nd). CRC press.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc Natl Acad Sci U S A*, *103*(49), 18866–18869.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *Int J Audiol*, *42 Suppl 1*, S49–58.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010.

Magnusson, L. (2011). Comparison of the fine structure processing (fsp) strategy and the cis strategy used in the med-el cochlear implant system: Speech intelligibility and music sound quality. *Int J Audiol*, *50*(4), 279–287.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends Cogn Sci*, *10*(8), 363–369.

McCloy, D., Souza, P., Wright, R., Haywood, J., Gehani, N., & Rudolph, S. (2013). The PN/NC Corpus. Version 1.0. *Seattle: University of Washington. Retrieved from http://depts.washington.edu/phonlab/resources/pnnc*.

Meddis, R., Delahaye, R., O'Mard, L., Sumner, C., Fantini, D. A., Winter, I., & Pressnitzer, D. (2002). A model of signal processing in the cochlear nucleus: Comodulation masking release. *Acta Acust united Ac*, *88*(3), 387–398.

Meddis, R., Lopez-Poveda, E. A., Fay, R. R., & Popper, A. N. (2010). *Computational models of the auditory system*. Springer.

Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *J Acoust Soc Am*, *102*(3), 1811–1820.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hear Res*, *266*(1-2), 36–51.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *J Acoust Soc Am*, *27*(2), 338–352.

Mok, B. A., Viswanathan, V., Borjigin, A., Singh, R., Kafi, H. I., & Bharadwaj, H. M. (2021). Web-based psychoacoustics: Hearing screening, infrastructure, and validation. *bioRxiv*, DOI: 10.1101/2021.05.10.443520. https://doi.org/10.1101/2021.05.10.443520

Moore, B. C. (2007). *Cochlear hearing loss: Physiological, psychological and technical issues* (2nd). John Wiley & Sons.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C., Glasberg, B. R., Flanagan, H. J., & Adams, J. (2006). Frequency discrimination of complex tones; assessing the role of component resolvability and temporal fine structure. *J Acoust Soc Am*, *119*(1), 480–490.

Moore, B. C., & Rosen, S. M. (1979). Tune recognition with reduced pitch and interval information. *Q J Exp Psychol*, *31*(2), 229–240.

Morillon, B., Lehongre, K., Frackowiak, R. S., Ducorps, A., Kleinschmidt, A., Poeppel, D., & Giraud, A.-L. (2010). Neurophysiological origin of human brain asymmetry for speech and language. *Proc Natl Acad Sci USA*, *107*(43), 18688–18693.

Musiek, F. E., & Chermak, G. D. (2013). *Handbook of central auditory processing disorder, volume i: Auditory neuroscience and diagnosis* (Vol. 1). Plural Publishing.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum Brain Mapp*, *15*(1), 1–25.

Obleser, J., & Weisz, N. (2011). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cereb Cortex*, *22*(11), 2466–2477.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *J Neural Eng*, *14*(5), 056001.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cereb Cortex*, *25*(7), 1697–1706.

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *J Neurosci*, *35*(18), 7256–7263.

Oxenham, A. J., Bernstein, J. G., & Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception. *Proc Natl Acad Sci USA*, *101*(5), 1421–1425.

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

Oxenham, A. J., & Simonson, A. M. (2009). Masking release for low-and high-pass-filtered speech in the presence of noise and single-talker interference. *J Acoust Soc Am*, *125*(1), 457–468.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, *2*(7).

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *J Acoust Soc Am*, *121*(4), 2312–2326.

Phatak, S. A., & Grant, K. W. (2012). Phoneme recognition in modulated maskers by normal-hearing and aided hearing-impaired listeners. *J Acoust Soc Am*, *132*(3), 1646–1654.

Picton, T. W. (2010). *Human auditory evoked potentials*. Plural Publishing.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? a late locus of selective attention to natural speech. *Eur J Neurosci*, *35*(9), 1497–1503.

Pressnitzer, D., Meddis, R., Delahaye, R., & Winter, I. M. (2001). Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus. *J Neurosci*, *21*(16), 6377–6386.

Pritchard, W. S. (1992). The brain in fractal time: 1/f-like power spectrum scaling of the human electroencephalogram. *Int J Neurosci*, *66*(1-2), 119–129.

Qin, M., & Oxenham, A. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J Acoust Soc Am*, *114*(1), 446–454.

Rabiner, L. (1993). Fundamentals of speech recognition. *Fundamentals of speech recognition*.

Rabinowitz, W., Eddington, D., Delhorne, L., & Cuneo, P. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *J Acoust Soc Am*, *92*(4), 1869–1881.

Rallapalli, V. H., & Heinz, M. G. (2016). Neural spike-train analyses of the speech-based envelope power spectrum model: Application to predicting individual differences with sensorineural hearing loss. *Trends Hear*, *20*, 2331216516667319.

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*(4), 2670–2679.

Ribary, U. (2005). Dynamics of thalamo-cortical network oscillations and human perception. *Prog Brain Res*, *150*, 127–142.

Rimmele, J. M., Golumbic, E. Z., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, *68*, 144–154.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci*, *336*(1278), 367–373.

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *J Acoust Soc Am*, *133*(4), 2431–2443.

Roß, B., Borgmann, C., Draganova, R., Roberts, L. E., & Pantev, C. (2000). A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones. *J Acoust Soc Am*, *108*(2), 679–691.

Rothauser, E. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans Audio Electroacoust*, *17*, 225–246.

Schooneveldt, G. P., & Moore, B. C. (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*(6), 1944–1956.

Senkowski, D., Talsma, D., Herrmann, C. S., & Woldorff, M. G. (2005). Multisensory processing and oscillatory gamma responses: Effects of spatial selective attention. *Exp Brain Res*, *166*(3-4), 411–426.

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*(5), 2631–2644.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Shannon, R. V., Zeng, F.-G., & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *J Acoust Soc Am*, *104*(4), 2467–2476.

Sheft, S., Ardoint, M., & Lorenzi, C. (2008). Speech identification based on temporal fine structure cues. *J Acoust Soc Am*, *124*(1), 562–575.

Shera, C. A., Guinan, J. J., & Oxenham, A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc Natl Acad Sci U S A*, *99*(5), 3318–3323.

Shinn-Cunningham, B., Ruggles, D. R., & Bharadwaj, H. (2013). How early aging and environment interact in everyday listening: From brainstem to behavior through modeling. *Basic aspects of hearing* (pp. 501–510). Springer.

Shinn-Cunningham, B., Varghese, L., Wang, L., & Bharadwaj, H. (2017). Individual differences in temporal perception and their implications for everyday listening. *The frequency-following response* (pp. 159–192). Springer.

Shinn-Cunningham, B. (2008). Object-based auditory and visual attention. *Trends Cogn Sci*, *12*(5), 182–186.

Shinn-Cunningham, B., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplif*, *12*(4), 283–299.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci*, *18*(1), 555–586.

Slaney, M. et al. (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, *35*, 8.

Slepian, D. (1978). Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: The discrete case. *Bell Syst Tech J*, *57*(5), 1371–1430.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J Acoust Soc Am*, *67*(1), 318–326.

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*(4), 2562–2576.

Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdiscip Rev Cogn Sci*, *11*(1), e1517.

Stinstra, J., & Peters, M. (1998). The volume conductor may act as a temporal filter on the ecg and eeg. *Med Biol Eng Comput*, *36*(6), 711–716.

Stone, M. A., Füllgrabe, C., & Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *J Acoust Soc Am*, *132*(1), 317–326.

Stone, M. A., & Moore, B. C. (2014). On the near non-existence of "pure" energetic masking release for speech. *J Acoust Soc Am*, *135*(4), 1967–1977.

Swaminathan, J., & Heinz, M. G. (2011). Predicted effects of sensorineural hearing loss on across-fiber envelope coding in the auditory nerve. *J Acoust Soc Am*, *129*(6), 4001–4013.

Swaminathan, J., & Heinz, M. G. (2012). Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. *J Neurosci*, *32*(5), 1747–1756.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans Audio Speech Lang Process*, *19*(7), 2125–2136.

Tallon-Baudry, C., & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn Sci*, *3*(4), 151–162.

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

Thomson, D. (1982). Spectrum estimation and harmonic analysis. *Proc IEEE*, *70*(9), 1055–1096.

Uusitalo, M. A., & Ilmoniemi, R. J. (1997). Signal-space projection method for separating meg or eeg into components. *Med Biol Eng Comput*, *35*(2), 135–140.

Van Eyndhoven, S., Francart, T., & Bertrand, A. (2017). Eeg-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Trans Biomed Eng*, *64*(5), 1045–1056.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *J Assoc Res Otolaryngol*, 1–11.

Verhulst, S., Bharadwaj, H. M., Mehraei, G., Shera, C. A., & Shinn-Cunningham, B. G. (2015). Functional modeling of the human auditory brainstem response to broadband stimulation. *J Acoust Soc Am*, *138*(3), 1637–1659.

Verschooten, E., Robles, L., & Joris, P. X. (2015). Assessment of the limits of neural phase-locking using mass potentials. *J Neurosci*, *35*(5), 2255–2268.

Viswanathan, V., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2019). Electroencephalographic signatures of the neural representation of speech during selective attention. *eNeuro*, *6*(5). https://doi.org/10.1523/ENEURO.0057-19.2019

Viswanathan, V., Bharadwaj, H. M., Shinn-Cunningham, B. G., & Heinz, M. G. (2021). Modulation masking and fine structure shape neural envelope coding to predict speech intelligibility across diverse listening conditions. *bioRxiv*, DOI: 10.1101/2021.03.26.437273. https://doi.org/10.1101/2021.03.26.437273

Viswanathan, V., Shinn-Cunningham, B. G., & Heinz, M. G. (2021). Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble. *bioRxiv*, DOI: 10.1101/2021.05.11.443678. https://doi.org/10.1101/2021.05.11.443678

Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans Audio, Speech, Language Process*, *26*(10), 1702–1726.

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiol Rev*, *90*(3), 1195–1268.

Wang, X., Lu, T., Bendor, D., & Bartlett, E. (2008). Neural coding of temporal information in auditory thalamus and cortex. *Neurosci*, *154*(1), 294–303.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, *58*(301), 236–244.

Wayne, R. V., Hamilton, C., Jones Huyck, J., & Johnsrude, I. S. (2016). Working memory training and speech in noise comprehension in older adults. *Front Aging Neurosci*, *8*, 49.

White, J. A., Banks, M. I., Pearce, R. A., & Kopell, N. J. (2000). Networks of interneurons with fast and slow γ-aminobutyric acid type a (gabaa) kinetics provide substrate for mixed gamma-theta rhythm. *Proc Natl Acad Sci USA*, *97*(14), 8128–8133.

Whitton, J. P., Hancock, K. E., Shannon, J. M., & Polley, D. B. (2017). Audiomotor perceptual training enhances speech intelligibility in background noise. *Curr Biol*, *27*(21), 3237–3247.

Wilson, B. S., & Dorman, M. F. (2008). Cochlear implants: A remarkable past and a brilliant future. *Hear Res*, *242*(1-2), 3–21.

Wilson, T. P. (1982). Relational networks: An extension of sociometric concepts. *Soc Networks*, *4*(2), 105–116.

Winn, M. B., Chatterjee, M., & Idsardia, W. J. (2013). Roles of voice onset time and f0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering. *J Speech Lang Hear Res*, *56*, 1097–1107.

Winter, I. M., & Palmer, A. R. (1990). Responses of single units in the anteroventral cochlear nucleus of the guinea pig. *Hear Res*, *44*(2-3), 161–178.

Wong, D. D., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., & De Cheveigne, A. (2018). A comparison of regularization methods in forward and backward models for auditory attention decoding. *Front Neurosci*, *12*, 531.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Atten Percept Psychophys*, *79*(7), 2064–2072.

Wöstmann, M., Herrmann, B., Maess, B., & Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proc Natl Acad Sci USA*, *113*(14), 3873–3878.

Zaar, J., & Dau, T. (2015). Sources of variability in consonant perception of normal-hearing listeners. *J Acoust Soc Am*, *138*(3), 1253–1267.

Zeng, F.-G., Popper, A. N., & Fay, R. R. (2011). *Auditory prostheses: New horizons* (Vol. 39). Springer Science & Business Media.

Zhu, L., Bharadwaj, H., Xia, J., & Shinn-Cunningham, B. (2013). A comparison of spectral magnitude and phase-locking value analyses of the frequency-following response to complex tones. *J Acoust Soc Am*, *134*(1), 384–395.

Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. *Acoustical factors affecting hearing aid performance*, *2*, 255–275.

VITA

Vibha Viswanathan

715 Clinic Drive, West Lafayette, IN 47907, viswanav@purdue.edu

Education

Aug 2021	Ph.D.	Purdue University, West Lafayette	Biomedical Engineering
Dec 2007	M.S.	University of Michigan, Ann Arbor	Electrical Engineering: Systems
May 2006	B.E.	Anna University, India	Electronics and Communication Engineering

Professional Experience

2017 – 2021 NIH F31 Predoctoral Fellow, Auditory Neurophysiology & Modeling Lab, Purdue University, West Lafayette, IN
Mentors: Michael Heinz, Barbara Shinn-Cunningham
Investigated the neurophysiological mechanisms of speech intelligibility under masking and distortion using a combination of electroencephalography (EEG), computational modeling, and behavioral experiments.

2016 – 2017 Lynn Fellow, CONNplexity Lab, Purdue University, West Lafayette, IN
Mentor: Joaquín Goñi
Developed fMRI-based graph statistical methods to detect temporal changes in the functional connectivity between brain regions.

2015 – 2016 Research Fellow, Auditory Neuroscience Lab, Boston University, Boston, MA
Mentor: Barbara Shinn-Cunningham
Investigated the neural mechanisms of speech-on-speech selective attention using EEG and graph theoretic approaches.

2011 – 2015	MATLAB Math Quality Engineer, MathWorks, Natick, MA *Wrote test suites for and helped design core MATLAB algorithms for linear algebra, signal processing, random number distributions, differential equation solvers, computational geometry, and graph/network algorithms.*
2008 – 2011	Engineer, Engineering Development Group, MathWorks, Natick, MA *Developed MATLAB tools for image/audio processing. Provided application support for Signal/Image Processing, Mathematics, and Statistics toolboxes.*

Membership

2020 –	Member, Acoustical Society of America
2019 –	Member, Society for Neuroscience
2015 –	Member, Association for Research in Otolaryngology

Awards and Honors

2019	Acoustical Society of America Conference Travel Award and Invited Talk
2018 – 2021	Ruth L. Kirschstein National Research Service Award (NRSA) Individual Predoctoral Fellowship (*F31*), National Institutes of Health (NIH)
2016 – 2017	Lynn Fellowship, Purdue University
2016	Travel Scholarship, ACNN Workshop on Big Neuroscience Data, Tools, Protocols & Services
2015	Travel Award, Center for Computational Neuroscience and Neural Technology, Boston University
2006	"First class with distinction" in Bachelor of Engineering, Anna University

Teaching

Fall 2020	Guest Instructor, Biomedical Engineering, Purdue University *BME 511: Biomedical Signal Processing*

Developed course content and delivered four lectures on Linear Algebraic Methods for Signal Processing

Fall 2017 Graduate Teaching Assistant, Biomedical Engineering, Purdue University
BME 305: Bioinstrumentation Circuit and Measurement Principles

Mentoring Experience

2020 Luis Fernando Aguilera de Alba (Undergraduate Researcher, Purdue University SURF Program)

Publications

Journal Articles

[J1] **Viswanathan, V.**, Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2019). "Electroencephalographic signatures of the neural representation of speech during selective attention," eNeuro, 6(5).

Preprints

[P1] **Viswanathan, V.**, Shinn-Cunningham, B. G., & Heinz, M. G. (2021). "Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble," bioRxiv 2021.05.11.443678.

[P2] Mok, B. A., **Viswanathan, V.**, Borjigin, A., Singh, R., Kafi, H. I., & Bharadwaj, H. M. (2021). "Web-based Psychoacoustics: Hearing Screening, Infrastructure, and Validation," bioRxiv 2021.05.10.443520.

[P3] **Viswanathan, V.**, Bharadwaj, H. M., Shinn-Cunningham, B. G., & Heinz, M. G. (2021). "Modulation masking and fine structure shape neural envelope coding to predict speech intelligibility across diverse listening conditions," bioRxiv 2021.03.26.437273.

Conference Proceedings

[C1] Gopi, E. S., **Viswanathan, V.**, Sankaralingham, P., & Ramakumar, S. (2005). "A new approach to create high level features from low level features of audio clips," Proceedings of IEEE International Conference on Communications, Circuits and Systems.

Forthcoming Manuscripts

[F1] **Viswanathan, V.**, Shinn-Cunningham, B. G., & Heinz, M. G. (In Preparation). "Speech categorization reveals the role of early-stage temporal-coherence processing in auditory scene analysis."

Invited Talks

[T1] **Viswanathan, V.**, Bharadwaj, H. M., Shinn-Cunningham, B. G., & Heinz, M. G. "Evaluating Human Neural Envelope Coding as the Basis of Speech Intelligibility in Noise," 177th Meeting of the Acoustical Society of America, Louisville, KY, 13–17 May, 2019.

Conference Abstracts/Presentations

[A1] **Viswanathan, V.**, Shinn-Cunningham, B. G., & Heinz, M. G. (2021). "Effects of Masker Modulation Spectra and Fine Structure on Consonant Confusions," Virtual Mid-Winter Meeting of the Association for Research in Otolaryngology, 20–24 Feb.

[A2] Mok, B. A., **Viswanathan, V.**, Borjigin, A., Singh, R., & Bharadwaj, H. M. (2020). "Anonymous Multipart Web-based Psychoacoustics: Infrastructure, Hearing Screening, and Comparison with Lab-based Studies," The Journal of the Acoustical Society of America, 148(4), 2713-2714.

[A3] **Viswanathan, V.**, Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2020). "Attentional Modulation of the Neural Representation of Speech: Spectral Profile and

Individual Differences," Mid-Winter Meeting of the Association for Research in Otolaryngology, San Jose, CA, 25–29 Jan.

[A4] **Viswanathan, V.**, Bharadwaj, H. M., Shinn-Cunningham, B. G., & Heinz, M. G. (2019). "Evaluating Human Neural Envelope Coding as the Basis of Speech Intelligibility in Noise," The Journal of the Acoustical Society of America, 145(3), 1717-1717.

[A5] **Viswanathan, V.**, Bharadwaj, H. M., Shinn-Cunningham, B. G., & Heinz, M. G. (2019). "Neurophysiological Validation of Envelope-based Models of Speech Intelligibility," Mid-Winter Meeting of the Association for Research in Otolaryngology, Baltimore, MD, 8–13 Feb.

[A6] **Viswanathan, V.**, Dzemidzic, M., Kareken, D., & Goñi, J. (2017). "Dynamic multivariate kurtosis as a measure of temporal variations in task-rest functional connectivity," NetSci, Indianapolis, IN, 19–23 June.

[A7] **Viswanathan, V.**, Dzemidzic, M., Kareken, D., & Goñi, J. (2017). "Dynamic multivariate kurtosis as a measure of temporal variations in task-rest functional connectivity, with applications to family history of alcoholism," Society for Neuroscience, Greater Indiana Chapter, IN, Indiana, 31 March.

[A8] **Viswanathan, V.**, Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2016). "Neural signatures of speech-on-speech selective attention," Mid-Winter Meeting of the Association for Research in Otolaryngology, San Diego, CA, 20–24 February.