THE ROLE OF TEMPORAL FINE STRUCTURE IN EVERYDAY HEARING

by

Agudemu Borjigin

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

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Hari M. Bharadwaj, Chair

Weldon School of Biomedical Engineering

Dr. Michael G. Heinz

Weldon School of Biomedical Engineering

Dr. Edward L. Bartlett

Weldon School of Biomedical Engineering

Dr. Daniel B. Polley

Mass. Eye & Ear Infirmary, Harvard University

Approved by:

Dr. Tamara L Kinzer-Ursem

ACKNOWLEDGMENTS

I thank my advisor, Dr. Hari M. Bharadwaj, for his continuous guidance and support throughout my Ph.D study over the past few years. I thank Dr. Michael G. Heinz, Dr. Edward L. Bartlett, and Dr. Daniel B. Polley for serving on my thesis committee. I also thank Josh Stohl for giving me the internship opportunity at MED-EL to work on the DNN project. Thank you Kostas Kokkinakis for the feedback and continued collaboration on the DNN project. My thanks also go to my friends and colleagues for their support and company on my Ph.D. journey. Special thanks to Andrew Sivaprakasam, Homeira Islam Kafi, Varsha M Athreya, and Ivy Schweinzger for working with me on the systematic review project. Finally, I wish to express my gratitude to my beloved parents for their endless love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis aims to investigate how one fundamental component of the inner-ear (cochlear) response to all sounds, the temporal fine structure (TFS), is used by the auditory system in everyday hearing. Although it is well known that neurons in the cochlea encode the TFS through exquisite phase locking, how this initial/peripheral temporal code contributes to everyday hearing and how its degradation contributes to perceptual deficits are foundational questions in auditory neuroscience and clinical audiology that remain unresolved despite extensive prior research. This is largely because the conventional approach to studying the role of TFS involves performing perceptual experiments with acoustic manipulations of stimuli (such as sub-band vocoding), rather than direct physiological or behavioral measurements of TFS coding, and hence is intrinsically limited. The present thesis addresses these gaps in three parts: 1) developing assays that can quantify TFS coding at the individual level 2) comparing individual differences in TFS coding to differences in speech-in-noise perception across a range of real-world listening conditions, and 3) developing deep neural network (DNN) models of speech separation/enhancement to complement the individual-difference approach. By comparing behavioral and electroencephalogram (EEG)-based measures, Part 1 of this work identified a robust test battery that measures TFS processing in individual humans. Using this battery, Part 2 subdivided a large sample of listeners (N=200) into groups with “good” and “poor” TFS sensitivity. A comparison of speech-in-noise scores under a range of listening conditions between the groups revealed that good TFS coding reduces the negative impact of reverberation on speech intelligibility, and leads to reduced reaction times suggesting lessened listening effort. These results raise the possibility that cochlear implant (CI) sound coding strategies could be improved by attempting to provide usable TFS information, and that these individualized TFS assays can also help predict listening outcomes in reverberant, real-world listening environments. Finally, the DNN models (Part 3) introduced significant improvements in speech quality and intelligibility, as evidenced by all acoustic evaluation metrics and test results from CI listeners (N=8). These models can be incorporated as “front-end” noise-reduction algorithms in hearing assistive devices, as well as complement other approaches by serving as a research tool to help generate and rapidly

sub-select the most viable hypotheses about the role of TFS coding in complex listening scenarios.

1. INTRODUCTION

Hearing loss is one of the most prevalent disabilities: over 5% of the world's population have disabling hearing loss and it is estimated that over 900 million people, or one in every ten people will have disabling hearing loss by 2050 (WHO, 2020). The communication disruption due to hearing loss is one of the main reasons for social isolation (Mick et al., 2014), which has evolved to be a growing epidemic that is linked to some other mental and physical challenges such as depression (Mener et al., 2013), heart disease (Rosen & Olin, 1965), dementia, and Alzheimer's disease (Lin et al., 2011). Although hearing assistive technologies such as hearing aids and cochlear implants (CI) could help restore near-normal audibility in quiet, people with hearing loss struggle to converse in noisy listening environments. The failure of current hearing assistive technologies to restore hearing functions in noise is partly due to our limited understanding of how sounds are coded in the human auditory system.

This thesis aims to investigate how one fundamental component of all sounds—temporal fine structure (TFS)—is used by a typical human auditory system for everyday hearing. Acoustic information in the auditory system is represented by cycle-by-cycle variations in phase, the TFS, and dynamic variations in amplitude, the envelope (ENV; Hilbert, 1906). Theses information is conveyed through the firing rate and/or timing of the neural spikes (i.e., rate-place vs temporal coding) of cochlear neurons. Neurons phase-lock to both TFS (Johnson, 1980), and ENV (Joris & Yin, 1992) robustly, with TFS phase-locking extending at least up to 1000 Hz (Verschooten et al., 2019). While the peripheral rate-place code has consistent counterparts throughout the auditory system, the upper limit of phase-locking progressively shifts to lower frequencies along the ascending pathway (Joris et al., 2004). How this metabolically expensive initial/peripheral temporal code (Hasenstaub et al., 2010; Laughlin et al., 1998) contributes to everyday hearing and how its degradation contributes to perceptual deficits are foundational questions in auditory neuroscience and clinical audiology. Yet, the significance of TFS coding is debated (Drullman, 1995; Oxenham, 2013; Oxenham & Simonson, 2009; Swaminathan & Heinz, 2012).

Current literature has established that sound localization and pitch perception could benefit from TFS cues in quiet (J. G. Bernstein & Oxenham, 2003; Houtsma & Smurzynski, 1990; Moore, 1973; Smith et al., 2002; Yin & Chan, 1990). However, whether TFS is important for spatial or pitch-based masking release in noise is further debated, especially when other redundant cues can also convey pitch or location, and when room reverberation can degrade temporal cues (Best et al., 2005; Ihlefeld & Shinn-Cunningham, 2011; Oxenham & Simonson, 2009). The challenge to pin down the functional significance of TFS in noise mainly arises from the limitations of the most commonly used approach to investigating TFS—vocoding, a technique to independently manipulate and study ENV and TFS cues (Ardoint & Lorenzi, 2010; Hopkins & Moore, 2009; Hopkins et al., 2008; Lorenzi et al., 2009; Smith et al., 2002). Although perceptually straightforward, the acoustic manipulations cannot eliminate subsequent confounding of ENV, TFS, and place cues without detailed knowledge of cochlear processing at the individual level (Oxenham, 2013; Swaminathan & Heinz, 2012). An alternative approach is to directly measure TFS sensitivity from individual listeners and compare it to individual differences in other perceptual measures. The individual-difference approach has been successfully used to address other fundamental questions (Bharadwaj et al., 2015; McDermott et al., 2010; Whiteford et al., 2020). Unfortunately, the lack of established measures of TFS sensitivity at the individual level limits this enterprise.

The goal of this thesis is to investigate the functional role of TFS in everyday hearing using an individual-difference approach. Two steps were taken to achieve this objective. The first step (chapter 2) was to identify candidate assays of TFS processing at the individual level, by employing a battery of both behavioral and electroencephalography (EEG)-based, classic TFS-sensitivity measures on a cohort of typical-hearing (TH) individuals. After establishing the array of TFS-sensitivity measurements at the individual level, the second step (chapter 3) was to extensively measure speech intelligibility in various types of listening situations on a large cohort of TH individuals, along with the TFS-sensitivity measures. The comparison between the individual differences in TFS sensitivity and speech-in-noise intelligibility allowed us to reveal the functional role of TFS in everyday hearing. In parallel to using the individual-difference approach to study TFS, this thesis explored the potential of deep neural network (DNN) models for mimicking or restoring speech-in-noise hearing in a human auditory system (chapter 4). Two state-of-the-art DNN models for speech segregation and enhancement were implemented and the performance of the models was evaluated by both acoustical evaluation metrics and CI listeners. These models can be incorporated as “front-end” noise-reduction algorithms in hearing assistive devices to help address the limitations of current signal processing strategies for restoring normal-level of speech hearing in noisy listening environments. These models also have the potential of complementing the individual-difference approach by serving as a research tool to help generate and rapidly sub-select the most viable hypotheses about the role of TFS coding in complex listening scenarios.

2. INDIVIDUALIZED ASSAYS OF TEMPORAL CODING IN THE ASCENDING HUMAN AUDITORY SYSTEM

This chapter has been published in eNeuro Journal (DOI).

2.1 Introduction

All acoustic information we receive is conveyed through the firing rate and/or timing of the neural spikes (i.e., rate-place vs temporal coding) of cochlear neurons. Temporal information in the basilar-membrane vibrations consists of cycle-by-cycle variations in phase, the temporal fine structure (TFS), and dynamic variations in amplitude, the envelope (ENV; Hilbert, 1906). Cochlear neurons phase-lock to both TFS (Johnson, 1980), and ENV (Joris & Yin, 1992) robustly, with TFS phase-locking extending at least up to 1000 Hz (Verschooten et al., 2019). While the peripheral rate-place code has consistent counterparts throughout the auditory system, the upper limit of phase-locking progressively shifts to lower frequencies along the ascending pathway (Joris et al., 2004). How this metabolically expensive initial/peripheral temporal code (Hasenstaub et al., 2010; Laughlin et al., 1998) contributes to everyday hearing and how its degradation contributes to perceptual deficits are foundational questions in auditory neuroscience and clinical audiology. Yet, the significance of TFS coding is debated (Drullman, 1995; Oxenham, 2013; Oxenham & Simonson, 2009; Swaminathan & Heinz, 2012).

Previous studies have explored whether sound localization and pitch perception benefit from TFS cues. While it is established that lateralization of low-frequency sounds depends on TFS (Smith et al., 2002; Yin & Chan, 1990), whether TFS is important for pitch perception is difficult to ascertain. Behavioral studies suggest that low-frequency periodic sounds elicit a stronger pitch than high-frequency sounds (J. G. Bernstein & Oxenham, 2003; Houtsma & Smurzynski, 1990; Moore, 1973), suggesting a possible role for TFS. However, these results permit alternate interpretations in terms of place coding and harmonic resolvability (Oxenham, 2012). Regardless of its role in quiet, whether TFS is important for masking release in noise is further debated, especially when other redundant cues can also convey pitch or location, and when room reverberation can degrade temporal cues (Best et al., 2005; Ihlefeld & Shinn-Cunningham, 2011; Oxenham & Simonson, 2009).

To investigate the role of TFS, studies have used sub-band vocoding to independently manipulate ENV and TFS cues (Ardoint & Lorenzi, 2010; Hopkins & Moore, 2009; Hopkins et al., 2008; Lorenzi et al., 2009; Smith et al., 2002). However, acoustic manipulations cannot eliminate subsequent confounding of ENV, TFS, and place cues without detailed knowledge of cochlear processing at the individual level (Oxenham, 2013; Swaminathan & Heinz, 2012). Thus, establishing the precise role of TFS through vocoding experiments is difficult, although the use of high-fidelity vocoders can help (Viswanathan, Bharadwaj, et al., 2021). An alternative approach is to directly measure TFS sensitivity from individual listeners and compare it to individual differences in other perceptual measures. The individual-differences approach has been successfully used to address other fundamental questions (Bharadwaj et al., 2015; McDermott et al., 2010; Whiteford et al., 2020). Unfortunately, the lack of established measures of TFS sensitivity at the individual level limits this enterprise.

Conventional behavioral TFS-sensitivity measurements have attempted to eliminate confounding cues such that primary task would rely on TFS processing (Hopkins & Moore, 2010; Moore & Sek, 2009; Sęk & Moore, 2012; Strelcyk & Dau, 2009). However, they did not assess the influence of extraneous factors on the measured scores. Unfortunately, nonsensory factors can contribute significantly to individual variability even when the tasks themselves rely on specific acoustic cues (G. R. Kidd et al., 2007). Objective electrophysiological measures of TFS sensitivity can circumnavigate this problem; however, such studies are scarce (Parthasarathy et al., 2020; Verschooten et al., 2015). Here, we employ a battery of both behavioral and electroencephalography (EEG)-based measures of TFS sensitivity on a cohort of typical-hearing (TH) individuals to identify candidate assays of TFS processing at the individual level. Our results suggest that extraneous variables dominate both behavioral and raw EEG measures. However, with adjustments, we observed robust behavior-EEG correlations in binaural assays, rendering them well suited for quantifying individual TFS processing.

2.2 Materials and Methods

The primary goal of the current study was to evaluate an array of both behavioral and electrophysiological measures as candidate assays of TFS sensitivity at the individual level. Based on the finding that nonsensory factors contribute significantly to behavioral TFS measures, a large-N supplementary behavioral experiment was conducted to assess whether nonsensory factors also influence ENV sensitivity when measured from naive participants.

2.2.1 Participants

One hundred and fifty-three listeners, aged 18–60 years, were recruited from the local community near Purdue University. All human subject measures were conducted following protocols approved by the Purdue University Internal Review Board and the Human Research Protection Program. Participants were recruited via posted flyers and bulletin-board advertisements and provided informed consent. All participants had pure-tone air-conduction thresholds of 25 dB Hearing Level (HL) or better at octave frequencies from 500 to 8000 Hz. Of the 153 subjects, 44 (20 males) participated in the main experiments designed to evaluate candidate assays of TFS processing. The remaining N = 109 participated in the supplementary experiment aimed at testing whether nonsensory factors also influence ENV sensitivity. Although the goal of the main experiment was to conduct all behavioral and electrophysiological TFS measures on each participant, some were not able to finish the full study battery because of limited availability. Among the 44 listeners who participated in the main study, 43 completed the frequency modulation (FM) detection task, and 36 completed the interaural time difference (ITD) detection task. The intersection, 33 subjects, completed both behavioral measurements. Among all participants (n = 44), 42 subjects completed EEG-ITD sensitivity measurements; 25 of those 42 subjects also completed EEG-frequency following response (FFR) measurements. Among the subjects who completed both behavioral measurements, all except one (n = 32) completed the EEG-ITD measurement; these subjects include all participants who completed the EEG-FFR measurements (n = 25). The subjects who completed both behavioral measurements (n = 32, age: mean = 26.8, SD = 11.2) were included for the main analyses including brain-behavior correlations. Although the age range was wide, only six out of 33 subjects were older than 35 years at the time of the testing, and age did not significantly correlate with any measure of this study.

2.2.2 Experimental Design and Statistical Analysis

Behavioral Measures of the TFS Coding

Each of the following behavioral measurements was conducted on a different day from the others to randomize the influence of factors that may be idiosyncratic to a specific test day/session. A single lab visit contained only one behavioral measurement to reduce the impact of cognitive fatigue from hour-long experiments.

FM DETECTION THRESHOLDS. To obtain monaural TFS sensitivity, FM thresholds were measured separately in each ear, using a weighted (3:1) one-down-one-up (Kaernbach, 1991), two-alternatives-forced-choice (2AFC) adaptive procedure. The stimulus in the target interval was a 500-ms-long 500-Hz tone with FM at a 2-Hz rate and variable depth. The reference interval was a 500-Hz pure tone. The interstimulus gap was 900 ms. The stimulus was ramped on and off with a rise/fall time of 5 ms to eliminate audible transitions. The stimulus level was 70 dB SPL. The subjects were instructed to press a button to indicate the interval containing the FM. Each measurement block was terminated after 11 reversals and the median of all the reversals from the adaptive procedure was extracted as the threshold. Four blocks of measurements were obtained in each ear from each subject. Except for an additional "demo" block to orient the participants before the formal testing, there was no further training. Sennheiser HDA 300 over-the-ear headphones were used for stimulus delivery. The slow FM rate of 2 Hz was chosen because it is thought that TFS cues are used to detect FM at rates below 10 Hz (Moore & Sek, 1996; Strelcyk & Dau, 2009). However, recent evidence suggests that this may not be the case (Whiteford et al., 2020). Nonetheless, given the large body of literature using and interpreting slow-FM detection as a measure of TFS sensitivity, we chose to include this in the battery of candidate measures.

ITD DETECTION THRESHOLDS. To obtain a binaural measure of TFS sensitivity, we measured ITD detection thresholds using a three-down-one-up, 2AFC adaptive procedure.

The stimulus consisted of two consecutive 400-ms-long, 500-Hz tone bursts with an ITD. The leading ear for the ITD was switched from the first burst to the second. The stimulus was ramped on and off with a rise/fall time of 20 ms to eliminate audible transitions and to reduce reliance on onset ITDs. The stimuli were presented at 70 dB SPL. Subjects were asked to report the direction of the jump (left-to-right or right-to-left) between the intervals through a button press. It was preferable to have subjects indicate the direction of change because absolute lateralization can be influenced by multiple factors (Moore & Sek, 2009). The threshold was defined as the geometric mean of the last nine reversals, and measured repeatedly across eight blocks, with a short break scheduled after the fourth block. Etymotic Research (ER-2) insert earphones were used for delivering the stimuli. A separate "demo" block was included before the experimental blocks to familiarize the subject with the task.

"NONSENSORY" SCORE. Because the main goal of the study is to evaluate candidate measures of TFS coding in naive subjects, i.e., individuals without extensive training/practice on the measured tasks, we anticipated that extraneous "nonsensory" variables may influence the measured thresholds. Accordingly, percent-incorrect scores on easy "catch" trials were calculated to quantify the subject's engagement. Errors made in these catch trials likely reflect nonsensory factors such as lapses in attention, variations in motivation, alertness, etc., rather than the strength of sensory coding. For the FM detection task, trials with frequency deviations (modulation depths) >15 Hz were deemed to be catch trials, and the percent-incorrect scores were calculated for just these trials for each subject as an estimate of lapse rate. Similarly, the criterion for designating a trial as a "catch" trial for the ITD detection task was that the ITD exceeded 80 μs. The number of catch trials available varied from subject to subject because of the adaptive nature of the task. On average, the FM and ITD detection tasks included 3–10 catch trials per block. To mitigate the influence of extraneous variables such as engagement and motivation on the measured thresholds, a simple linear model was constructed with this nonsensory score as the sole predictor, and the residuals from the model were treated as "clean" thresholds and used in all analyses thereafter.

SUPPLEMENTARY AMPLITUDE MODULATION (AM) DETECTION TASK. To further investigate the influence of nonsensory factors on behavioral measures in general, we conducted a supplementary experiment using a task that is unrelated to TFS processing, an AM detection task similar to the one used in Bharadwaj et al., 2015. A similar 2AFC procedure as in the FM and ITD detection threshold measurements was employed. The target was a 500-Hz, 75 dB SPL band of noise centered at 4 or 8 kHz, and amplitude modulated at 19 Hz. Two unmodulated tones, flanked at two equivalent rectangular bandwidths (ERBs; Glasberg and Moore, 1990; Moore, 1968) away from the center frequency, each at 75 dB SPL, were used to minimize off-frequency listening. The signal in the reference interval was statistically identical but unmodulated. Using a noise carrier helps eliminate spectral cues for the AM detection task (Viemeister, 1979). The threshold for the modulation depth detection was determined by an adaptive weighted one-up-one-down procedure (Kaernbach, 1991).

Electrophysiological Measures of the TFS Coding

While behavioral measures directly assess perceptual sensitivity to TFS, they may also reflect common nonsensory factors such as attention and motivation. To dissociate TFS coding from nonsensory factors, we designed two passive EEG measures of TFS coding and compared them to individual behavioral measures. For EEG measurements, participants watched a silent, captioned video of their choice while passively listening to the auditory stimuli. EEG recordings were obtained using a 32-channel EEG system (Biosemi Active Two), while the stimuli were presented via ER-2 insert earphones.

GENERAL EEG SETUP AND PREPROCESSING PROCEDURES. The Biosemi EEG system employs active common-mode noise rejection using a pair of ground electrodes in a “driven-right leg” configuration (Metting van Rijn et al., 1990). EEG recordings were re-referenced to the average voltage across the two ear lobes. For cortical response analyses (EEG-ITD; see Cortical correlates of TFS-based ITD processing), the raw data were bandpass filtered from 1 to 50 Hz, whereas for subcortical responses (EEG-FFR; see FFR), raw data were filtered from 400 to 1300 Hz. The 400- to 1300-Hz bandpass filter eliminates artifacts from eye blinks. For the 1- to 50-Hz cortical data, ocular artifacts were removed using the signal-space projection technique (Uusitalo & Ilmoniemi, 1997). After the eye-blink correction, epochs with large voltage excursions (above 150 μV for cortical recordings; above 50 μV for subcortical recordings) were excluded to reduce movement artifacts. For both cortical and subcortical recordings, analyses focused on recordings from vertex electrodes (i.e., Fz and Cz channels).

CORTICAL CORRELATES OF TFS-BASED ITD PROCESSING. Cortical EEG was recorded in response to 70 dB SPL 500-Hz tones that were amplitude-modulated (100% depth) at 40.8 Hz. 40.8 Hz can elicit a strong auditory steady-state response (ASSR) in EEG recordings (Picton et al., 2003); this response was used here as a measure of recording quality (Fig. 1C). The stimulus duration was of 1.5 s. As with the behavioral measurement, the leading ear for the ITD switched 1 s into the trial. The direction of the ITD switch was randomized across trials. To minimize monaural cues, the ITD switch coincided with a trough of the 40.8 Hz modulation (Fig. 2.1). This approach mirrors the method used in Papesh et al., 2017, where the stimulus switches between in-phase and out-of-phase states (phase shift of 180°). Our measurements involved ITD jumps of 20, 60, 180, or 540 μs in magnitude. The magnitude and direction of the ITD jump were randomized across trials. A total of 1200 trials were presented to the listener. The interstimulus interval was uniformly distributed between 500 and 600 ms. Besides amplitude and latency of the averaged evoked response across trials in each condition, we calculated the intertrial coherence (ITC), which quantifies the consistency in the phase of the evoked response components across trials. ITC of 0 indicates no phase locking (the response is dominated by background noise), and ITC of 1 indicates perfect phase-consistency across trials (no background noise added to the phase-locked response). Thus, the ITC is directly related to the signal-to-noise ratio of the evoked response (Bharadwaj & Shinn-Cunningham, 2014). The frequency band for ITC analysis was restricted to ∼1–20 Hz, because it is known that cortical transient-evoked responses primarily consist of low-frequency components, and because we sought to separate these responses from the 40.8-Hz ASSR response.

FFR. Subcortical FFRs were measured in response to tones in a forward-masking stimulus configuration (Verschooten & Joris, 2014). The stimuli consisted of three consecutive segments: a 500-Hz probe tone that was 100 ms long and at 75 dB SPL, a “forward-masker” tone of the same frequency and duration but at 85 dB SPL, and the same probe tone. A 50-

ms silent gap was included between the first probe tone and forward-masker, but only a 1-ms gap was included between the forward-masker and the second probe tone. Each stimulus segment was ramped on and off over 5 ms to reduce audible transitions. The polarity of the stimulus was alternated across a total of 8000 trials. The 500-Hz component of differential response obtained across the two stimulus polarities reflects response components that are phase-locked to the TFS, whereas the summed 500-Hz response represents the response to the ENV. However, the TFS component can contain both preneural (e.g., cochlear microphonic; CM) as well as neural responses. Verschooten and Joris, 2014 argued that the nonlinear residual obtained by subtracting the TFS response to the second probe tone from the TFS response to the first probe tone will isolate the neural component and suppress the approximately linear CM. This is because the forward masking of response to the second probe tone only masks the neural component, whereas the CM is intact. Owing to the inner-hair-cell rectification, the summed response across the two polarities also contains a component at twice the stimulus frequency (1000 Hz) that reflects physiological currents phase-locked to the TFS in the stimulus. Although TFS-related, whether this double-frequency response is purely neural as has been previously interpreted (Parthasarathy et al., 2020), or whether it includes preneural contributions is unknown. Thus, we considered two candidate subcortical correlates of TFS processing: (1) the 500-Hz component derived from the differential response across the two polarities of stimulus presentation, and (2) the 1000-Hz component derived from the summed response across two polarities of stimulus presentation.

Statistical Analysis

Pearson correlations were calculated to illustrate simple associations between pairs of measurements. Statistical inference about behavior-physiology correlations was made using a multiple linear stepwise regression analysis by adding new potential predictors one by one to model the dependent variable. All reported significant associations met a false discovery rate criterion of 5% to control for multiple comparisons (Benjamini & Hochberg, 1995). Statistical analyses were performed using R (R Core Team).

CODE ACCESSIBILITY. Stimulus generation and data analyses were done using custom scripts. They stimulus can be accessed at: stimulus-github; The analysis scripts for EEG data can be accessed at: EEGAnalysis-github; The analysis scripts for behavioral data can be accessed at: BehaviorDataAnalysis-github.

2.3 Results

2.3.1 Non-sensory factors contribute to large individual differences in behavioral measures of TFS coding

Similar to previous reports of large individual differences in the AM and ENV-based ITD detection thresholds across TH listeners (Bharadwaj et al., 2015), both the FM and TFS-based ITD detection thresholds varied widely across our TH listeners. FM detection thresholds across 43 TH listeners ranged from 7 to 22 dB relative to 1 Hz [i.e., a frequency deviation (Fdev) of 2–13 Hz from 500 Hz]. ITD detection thresholds varied from 21 to 39 dB relative to 1 μs (i.e., 11–89 μs) across 37 TH listeners. These FM and ITD detection thresholds are shown along with the results from similar studies, in Figures 2.8 and 2.1, respectively, and were largely comparable.

Across listeners, neither FM (averaged across two ears) nor ITD thresholds (each averaged across repetitions) correlated with the audiograms (across-ear average of thresholds at 500 Hz; across-ear average of the mean thresholds at high frequencies: 4 and 8 kHz); however, the two measures were significantly correlated with each other in a simple linear regression analysis (r = 0.44, p = 0.01, n = 33). While the correlations may arise from individual differences in TFS coding, they can also reflect nonsensory factors such as attention, motivation, etc. To disambiguate these competing explanations, we assigned each listener a nonsensory score. When those scores were factored out from each measurement, the correlation between the monaural FM and binaural ITD thresholds dropped such that the association no longer met conventional statistical significance criteria (R = 0.31, p = 0.08, n = 33), suggesting that nonsensory factors play a large role in raw scores. Furthermore, when just the blocks with the largest (i.e., worst) FM and ITD thresholds for each subject were compared, considerably stronger correlations were observed (r = 0.6, p = 9e-4, n = 33), underscoring the involvement of nonsensory factors in behavioral measurements. Figure 2.2 shows the correlations between the measured and predicted thresholds solely based on the lapse rates (i.e., the nonsensory score). The involvement of nonsensory factors is evident, especially for the poorer performers.

To confirm the involvement of nonsensory factors in raw behavioral scores, a similar comparison of thresholds and lapse rates was conducted for the supplementary AM detection task. The predicted thresholds based on the nonsensory score significantly correlated with the measured AM thresholds (R = 0.52, p = 1e-8, n = 109; Fig. 2.3). This result indicates the significant weight of nonsensory factors, not only for FM and ITD detection measurements but behavioral measures in general.

2.3.2 Raw electrophysiological TFS measures are strongly influenced by extraneous sources of variance

Two passive electrophysiological measurements were conducted to objectively evaluate individual TFS coding. Because passive electrophysiological measures are likely to be influenced by distinct extraneous factors (e.g., head size) compared with behavioral measures (e.g., motivation/engagement), these measurements provide a complementary window into individual TFS coding.

CANDIDATE CORTICAL CORRELATES OF TFS PROCESSING. Cortical responses evoked by the polarity shift of the ITD are quantified through the phase-locking strength shown in the phase-locking spectrograms (Fig. 2.1C). Clear responses to the onset, offset, and ITD jump are apparent in the low-frequency portion of the phase-locking spectrogram. The sustained ASSR is also clear around 40.8 Hz. The average response from 42 TH listeners shows monotonically increasing phase-locking strength of the ITD-evoked response across the ITD magnitudes (Fig. 2.1D), confirming that the response is indeed sensitive to TFS processing and the size of the ITD jump. Perhaps more important for the search of candidate TFS processing assays, large individual differences are apparent in the phase-locking strength across subjects (Fig. 2.4). Most subjects did not show a salient response for the 20-μs condition, and only about half showed robust responses for the 60-μs condition. Focusing therefore on the 180- and 540-μs conditions, the 180-μs condition is still part of the increasing slope of the response-versus-ITD-jump-size trend, but the response amplitude may have saturated for the 540 μs. Accordingly, we used each individual's response for the 180-μs condition for comparison to behavior. Note that the ITD being referred to here is the size of the jump; for instance, for the 20-μs condition, the stimulus started with an ITD of 10 μs with one ear leading and jumped to the other side about halfway through the stimulus to end with a 10-μs ITD with the other ear leading.

Unfortunately, one striking aspect of the result in Figure 2.4 is that even at 540 μs, the individual differences that were present in the lower ITD conditions persist. The ITD jump is obviously perceptible at 540 μs, and the EEG response appears to be near saturation level for most individuals; this suggests that a significant portion of the individual differences in the magnitude of the cortical response arises from factors extraneous to TFS-based processing. Extraneous factors that may contribute include anatomic factors such as head size, and the geometry/orientation of the neural sources relative to the scalp sensors (Bharadwaj et al., 2019). Thus, although the cortical response to ITD jumps is indeed elicited and parametrically modulated by TFS-based processing, raw response amplitude metrics may be unsuitable for use as an individualized assay of TFS coding.

2.3.3 Candidate subcortical correlates of TFS processing

Figure 2.5 shows an example FFR recording from a single individual in response to the stimulus sequence with a probe tone, a forward masker, and a second probe tone. The top row (green traces) shows the differential response across two stimulus polarities. This response to the probe tone (labeled "d1" in Fig. 2.5) tracks the 500-Hz TFS in the stimulus, but contains both preneural (e.g., CM) and neural components. Because forward masking is thought to arise from synaptic processing (Verschooten & Joris, 2014), the forward masker would be expected to only suppress the neural (i.e., postsynaptic) component of the response to the second probe tone, leaving the preneural component intact (labeled "d2" in Fig. 2.5). Thus, subtracting d2 from d1 should leave a purely neural response phase-locked to the TFS.

The bottom row in Figure 2.5, blue traces, shows the summed response across two polarities. Because of inner hair-cell rectification, this response contains a 1000-Hz component arising from the stimulus TFS (also see Materials and Methods). This 1000-Hz component in response to the probe (labeled "s1" in Fig. 2.5) has previously been interpreted as a neural response (Parthasarathy et al., 2020). If that were indeed the case, the forward-masker would considerably suppress the 1000-Hz component in response to the second probe (labeled "s2" in Fig. 2.5).

Figure 2.6 shows the average d1 (Fig. 2.6A), d1–d2 (Fig. 2.6B), s1 (Fig. 2.6C), and s1-s2 (Fig. 2.6D) response obtained across subjects, quantified in the frequency domain. It is evident from the reduced size of the (d1–d2) response compared with the d1 response, and the reduced size of the (s1–s2) response compared with the s1 response that, forward masker only has a partially suppressing effect. This provides evidence that both candidate TFS measures, the 500-Hz component from the difference across stimulus polarities, and the 1000-Hz component from the sum across stimulus polarities, have significant preneural contributions. This is in contrast to the previous interpretation that the component at double the tone frequency is purely neural (Parthasarathy et al., 2020).

These results indicate that a forward-masking paradigm will need to be employed to extract the purely neural "residual" response. Unfortunately, unlike transtympanic recordings that are difficult to perform (Verschooten et al., 2015), this residual is small and not readily measurable from all individual subjects. Thus, while subcortical envelope-following responses (EFRs) provide a robustly measurable correlate of envelope processing (Bharadwaj et al., 2015), tracking the TFS via FFRs are not promising, and not readily measured across all individuals despite our cohort being comprised of TH listeners.

2.3.4 "Adjusted" behavioral and cortical measures are strongly correlated, likely reflecting TFS coding

Based on individual differences in the cortical amplitude measure persisting for the large-ITD-jump (540 μs) condition, we concluded that the amplitude measure of cortical phase-locking was dominated by extraneous variance, likely from anatomic factors. Thus, we focused our attention on the latency of the ITD-jump response, because the latency is expected to be unaffected by the scaling effects of individual anatomy. In particular, we extracted the latency of the cortical response to the 180-μs jump condition to avoid floor and ceiling effects. The latency was the mean of N1 and P2 latency (the latency is the time difference between the red dashed line and either N1 or P2 peak in Fig. 2.1B). The use of the latency metric was also motivated by the previous successful use of this EEG-latency measure to predict individual behavioral measures of spatial release from masking (Papesh et al., 2017). In addition to this latency metric, the slope of the cortical-response amplitude with increasing ITD-jump (i.e., the increase from the 60-μs condition to the 180-μs condition, divided by the 540-μs condition, in the ITC plot of Fig. 2.4) was extracted as a normalized measure of TFS processing that would mitigate the overall scaling influence of anatomic factors. This normalization was also motivated by the previous successful use of a similarly normalized electrophysiological measure in the context of modulation processing (Bharadwaj et al., 2015).

Both of these "adjusted" cortical measures exhibited significant correlations with behaviorally measured ITD thresholds. Specifically, individual differences in latency of the cortical ITD-jump response (for 180 μs) correlated with individual differences in the ITD detection thresholds ($R = 0.35$, $p = 0.048$, $n = 32$). The correlation improved when the behavioral scores were also adjusted to factor out the nonsensory score ($R = 0.45$, $p = 0.01$, $n = 32$). The slope metric from the cortical EEG response also correlated with ITD thresholds both with and without adjustments to the behavioral scores ($R = 0.43$, $p = 0.021$, $n = 32$, original ITD scores; $R = 0.42$, $p = 0.028$, with nonsensory score factored out). There were no significant brain-behavior correlations with "unadjusted" or "raw" metrics, such as the ITC amplitude of the ITD-evoked response, even after normalization by the ITC amplitude of the onset response.

With the subcortical measures, because results indicated a significant preneural contribution for both candidate TFS measures, and the residual neural component extracted from the forward-masking paradigm was not robustly measurable for many participants, we did not explore FFR-behavior associations in detail. A simple correlational analysis between the residual (d1–d2) 500-Hz response and ITD thresholds suggested that the correlations were not statistically distinguishable from zero (data not shown).

A multiple linear regression model was used to predict ITD detection thresholds using both the nonsensory score, the EEG latency, as well as the EEG normalized slope metric (both from cortical ITD-jump response). The model could predict the behavioral ITD threshold well (Fig. 2.7) with the predictors together accounting for more than half of the variance observed in the behavioral thresholds (Table 2.1). We interpreted this result as suggesting that both "adjusted" behavioral scores, and electrophysiological latency or slope metrics in response to TFS-based binaural processing are promising candidate assays of TFS processing that may be suitable for use at the individual level.

Table 2.1. Model prediction of the behavioral ITD detection thresholds, with factors including the nonsensory score, EEG latency, and EEG slope. The variations accounted for by the nonsensory score are more than three times as by either one of the two EEG metrics. Together, more than half the variance can be explained.

Predictor	Variance Explained
Non-sensory score	37.48%
EEG latency	10.03%
EEG slope	9.28%
Explained	56.79%
Unexplained	43.21%

2.4 Discussion

In the present study, we sought to identify viable assays that can index the fidelity of TFS processing at the individual subject level. To obtain insight into whether individual differences in various candidate measures reflected TFS-based processing or extraneous factors, we compared individual differences in behavioral scores across FM and ITD detection tasks to differences in cortical and subcortical EEG-based measures. Results revealed the strong influence of extraneous factors on both behavioral scores and amplitude-based EEG metrics.

With behavioral measures, nonsensory factors quantified using the lapse rate in catch trials, could account for a third of the variance across individuals. Although previous work has explored a range of behavioral TFS measures (Hopkins & Moore, 2010; Moore & Sek, 2009; Sęk & Moore, 2012), the results from the present study underscore the importance of adjusting raw behavioral scores to reduce the impact of nonsensory factors. Indeed, although raw FM and ITD measures correlated significantly with each other, similar to the correlation between monaural AM detection and binaural envelope-ITD thresholds (Bharadwaj et al., 2015), this was driven in part by nonsensory factors. Because phase-locking to the TFS is essential for low-frequency ITD processing (Yin & Chan, 1990), it is plausible that ITD detection thresholds can provide an index of TFS sensitivity. On the other hand, whether FM detection relies on TFS coding has been controversial because of the possible role of recovered ENV cues that result from cochlear filtering of FM stimuli; indeed, FM stimuli lead to perceptible out-of-phase ENV fluctuations at cochlear places tuned to frequencies just above and below the FM carrier (Whiteford et al., 2017; Whiteford & Oxenham, 2015). Whiteford et al., 2020 extensively tested the role of place coding in FM detection and found that place coding by itself can account for the observed variations in FM sensitivity across all carrier frequencies and modulation rates. This finding is in contrast to the widely accepted view of the utilization of time coding in the detection of slow-rate FM (Moore & Sek, 1996; Parthasarathy et al., 2020; Strelcyk & Dau, 2009). Together with our finding that nonsensory factors influence raw behavioral scores, this uncertainty about the link between TFS coding and FM detection calls into question the previous use of FM detection scores as a correlate of TFS processing. In contrast, unambiguous theoretical links can be made between ITD detection and TFS coding, suggesting that once ITD thresholds are adjusted to reduce the influence of nonsensory scores, they may serve as a useful metric of TFS processing. This was corroborated by our finding that passive EEG measures, when combined with nonsensory scores, can account for more than half of the variance in ITD thresholds. Here, we used lapse rates in the catch trials to obtain a correlate of nonsensory factors. Alternately, a surrogate behavioral task that does not rely on TFS coding (e.g., interaural level difference sensitivity) may also be used to adjust ITD thresholds with similar benefits.

Another key finding from the present study is that although passive EEG measurements can potentially reflect TFS-based processing objectively, they too are susceptible to the influence of extraneous factors. Indeed, consistent with the interpretation that individual anatomic factors can have a scaling influence on response amplitudes, we found that cortical responses phase-locked to ITD changes showed large individual differences even for a large

ITD jump (540 μs) where the response amplitude was near saturation for most individuals. Therefore, we argued that the evoked-response latency and/or percent growth/slope metrics may be better assays of TFS processing. Accordingly, latency and slope metrics showed significant correlations with behavioral ITD detection thresholds. For candidate subcortical FFR-based measures of TFS processing, our results showed that preneural physiological currents (CM, inner hair-cell currents) contribute significantly to the measure, thus complicating their applicability. Indeed, brainstem response measures from individuals with compromised inner hair-cell synaptic transmission show that preneural transduction currents can contribute to the measured response (Santarelli et al., 2009). Moreover, when employing a forward-masking-based design to isolate the neural component of the FFR, the resulting signal is relatively weak even in our TH cohort. This result from non-invasive ear-canal recordings is in contrast to neurophonic measurements from the auditory nerve (Snyder & Schreiner, 1985) or round window (Henry, 1995) from animals, or FFR measurements from humans using transtympanic electrodes where the forward-masking design has been used successfully (Verschooten et al., 2018). Although FFRs have previously been used as a putative correlate of TFS-based processing (Parthasarathy et al., 2020), our results suggest that additional experiments are needed to clarify the interpretation of those results.

Our finding that the subcortical FFR may be a poor correlate of neural TFS processing is in contrast to previous results suggesting that subcortical EFRs are correlated with behavioral measures of ENV processing. For example, Bharadwaj et al., 2015 showed that the AM detection thresholds and ENV-based ITD thresholds correlated strongly with normalized EFR-based metrics. This is likely both because EFR measurements more readily exclude preneural contributions (which primarily track the TFS), and because Bharadwaj et al., 2015 obtained asymptotic behavioral scores from a large number of trials (1200–1500 trials) from trained subjects. Indeed, with naive subjects in this study, an AM detection task similar to the one used in Bharadwaj et al., 2015 also showed a strong influence of nonsensory factors.

In summary, the present study examined various candidate assays for quantifying TFS processing at the individual subject level. These included behavioral FM and ITD detection thresholds, and EEG-based cortical and subcortical physiological measures. Among these, our experiments suggest that the latency of cortical responses to ITD jumps, normalized cortical response amplitude (i.e., percent growth/slope), and "adjusted" ITD thresholds may all be useful. Indeed, when a multiple linear regression model was constructed to predict behavioral ITD thresholds, the combination of the nonsensory score (lapse rate in catch trials), EEG latency, and slope measures could account for >50% of the variance across individuals. Our results are consistent with the findings by Papesh et al., 2017, who also found a correlation between ITD-evoked EEG latency and spatial-hearing outcomes such as spatial release from masking. Given that multiple candidate measures were explored to identify the most promising assays, future experiments should be conducted to independently confirm the efficacy of the assays endorsed by our results. The most promising assays rely on binaural TFS-based processing. Indeed, similarly to our results, steady-state cortical responses that track continuous interaural phase modulations have also been found to correlate with behavioral binaural sensitivity, further corroborating the potential utility of cortical binaural measures as electrophysiological assays of TFS processing (Koerner et al., 2020; Undurraga et al., 2016).

Reliable measures of TFS processing are critical for future investigations into the role of TFS in everyday hearing using intact speech-in-noise stimuli without vocoding manipulations. While sub-band vocoding can allow for independent manipulation of acoustic TFS and envelope cues, subsequent cochlear processing can confound these factors once again (Gilbert & Lorenzi, 2006; Swaminathan & Heinz, 2012). Furthermore, when both rate-place/ENV cues and TFS cues are redundant, vocoding experiments cannot provide insight into how they are perceptually weighted. The candidate TFS measures identified in the present study can help address these gaps.

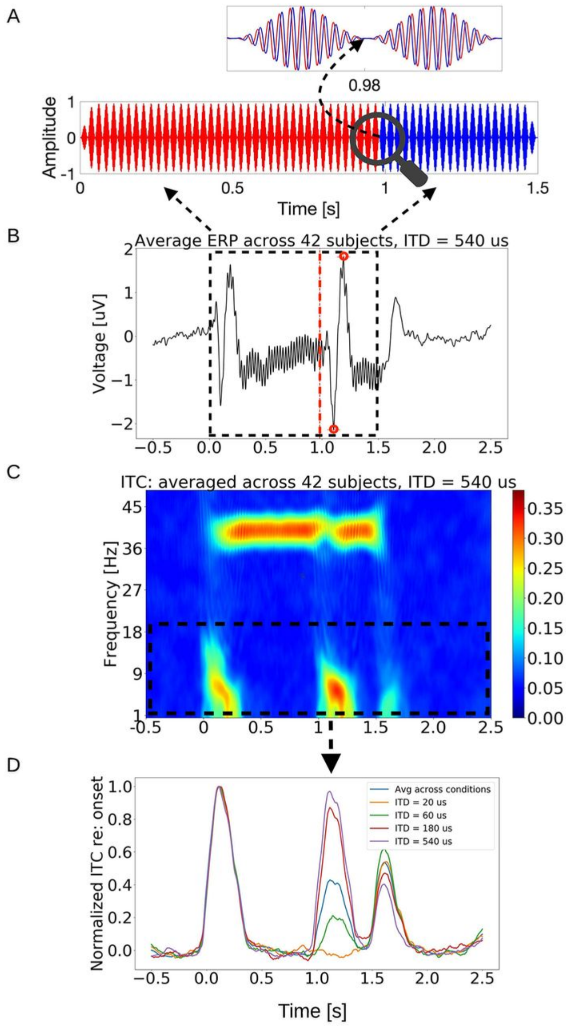


Figure 2.1. Stimulus paradigm and response from the EEG-TFS sensitivity measurement. A, The stimulus is a 1.5-s-long, 500-Hz pure tone that is amplitude modulated at 40.8 Hz. The red color represents the sound in the right ear, whereas the blue stands for the sound in the left ear. In the figure, the stimulus in the right ear leads in time till 0.98 s (indicated by the red segment of zoomed-out view of the stimulus), after which the ITD shifts in polarity, i.e., the stimulus in the left ear takes the lead. The ITD jump occurs when the stimulus amplitude is zero to minimize the involvement of monaural cues (pointed out by the dashed arrow). B, Averaged evoked response potential (ERP) from all trials across 42 subjects in "ITD = 540 μs" condition from Cz electrode. The red dashed line indicates where the ITD switched polarity, which resulted in N1 and P2 responses (denoted by red dots). C, ITC spectrogram of the EEG response, averaged across 42 subjects, with the colormap indicating the ITC. Robust ASSRs can be seen around the AM frequency of 40.8 Hz. There are also salient responses time locked to the stimulus onset, offset, and importantly, to the ITD jump. D, The average time course of the ITC for frequencies below 20 Hz is shown for each ITD jump condition. The response evoked by the shift in the ITD polarity increases monotonically with the size of the ITD jump, confirming that the response is parametrically modulated by TFS-based processing.

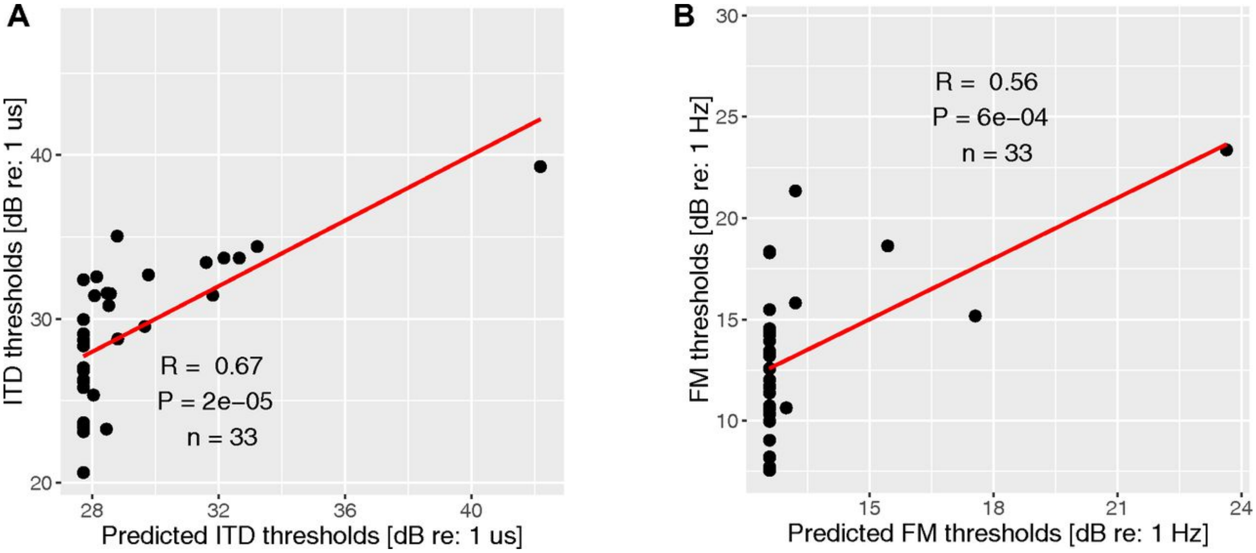


Figure 2.2. Measured vs. predicted thresholds based on lapse rate. [A] Measured vs predicted ITD detection thresholds; [B] Measured vs predicted FM detection thresholds. The significant contribution of non-sensory factors is apparent, especially for the poorer performers.

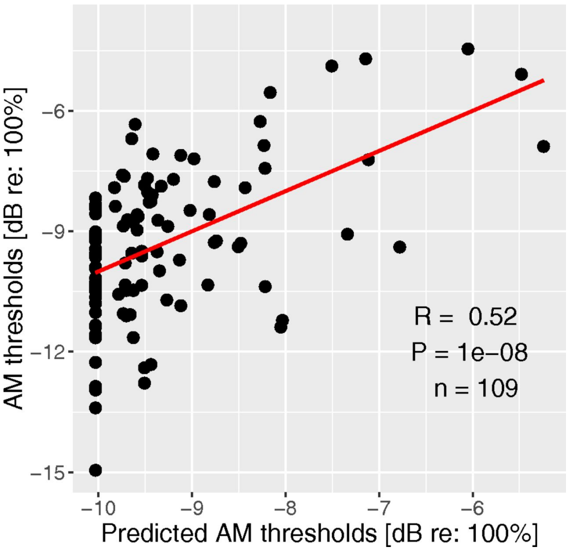


Figure 2.3. Measured vs predicted AM thresholds based on lapse rate. The thresholds are the average detection thresholds of AM tones at 4 kHz and 8 kHz. The significant contribution of non-sensory factors is apparent.

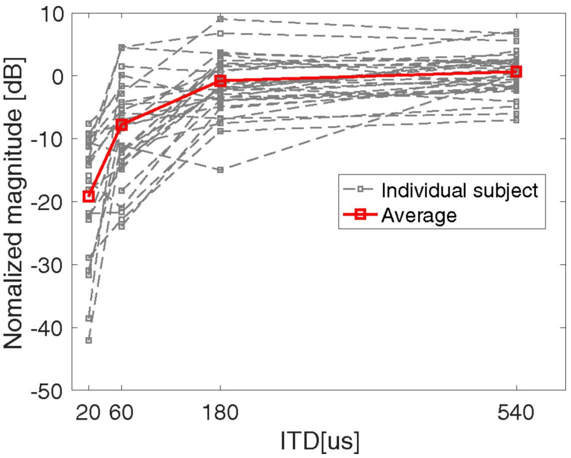


Figure 2.4. Individual EEG ITC (averaged under 20 Hz) values as a function of the jump size of the ITD. The ITC increases with the ITD for almost all subjects. Robust responses above noise floor are detected for most subjects for the "ITD = 180 μs" condition. Interestingly, individual differences present at 180 μs persist even at 540 μs despite the ITD jump being obviously perceptible and the response amplitude appearing to saturate.

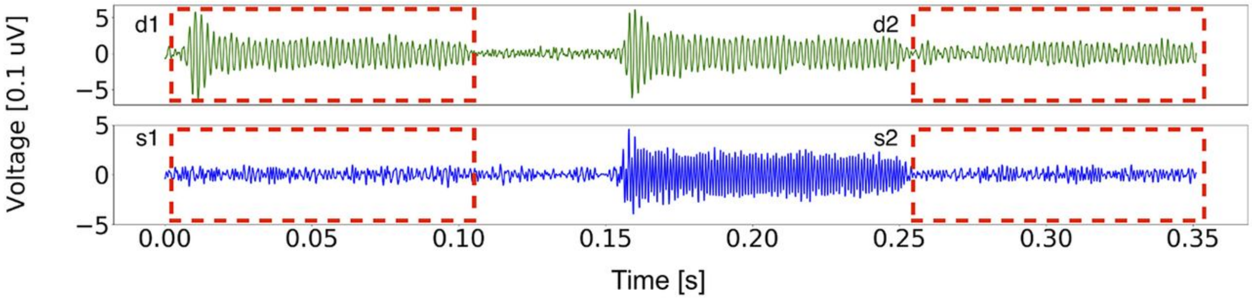


Figure 2.5. FFR to the probe-forward-masker-probe stimulus sequence for an individual subject. The top row (green trace) represents the differential response across two stimulus polarities, whereas the bottom row (blue trace) represents the summed response across two stimulus polarities. The first boxed segments in both rows (red, dashed box, labeled d1 or s1) reflect the raw response to the probe tone, which is likely a mixture of neural and preneural responses (e.g., CM), whereas the second boxed segments in both rows (red, dashed box, labeled d2 or s2) is the adapted response after forward masking. For d2 and s2, the preneural (e.g., CM) component is expected to be intact, whereas the neural response is attenuated by forward masking (because of the very short 1-ms gap). The forward masker only partially suppresses the responses, suggesting a strong preneural contribution to d1 and s1. The weaker residuals obtained by subtraction, i.e., (d1 – d2) and (s1 – s2) are likely purely neural.

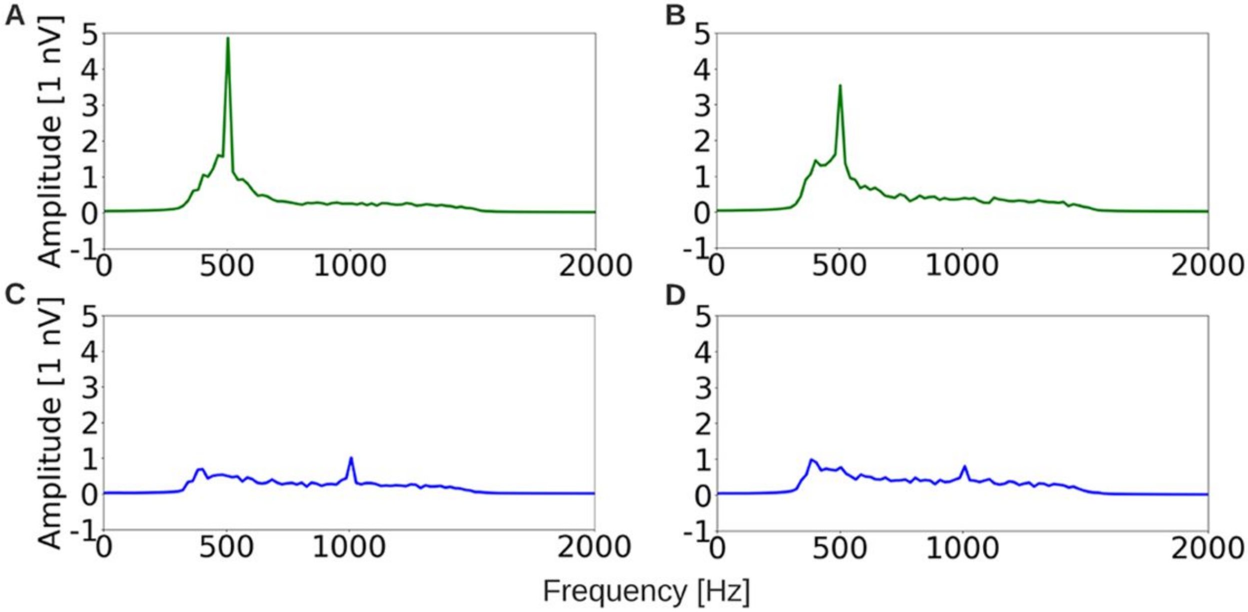


Figure 2.6. Frequency-domain representations of the d1 (A), d1–d2 (B), s1 (C), and s1–s2 (D) segments from Figure 2.5, but averaged across subjects. Forward masking partially attenuates both the 500-Hz component of d1 response, and the 1000-Hz component of the s1 response, suggesting that both responses reflect a mix of preneural and neural sources.

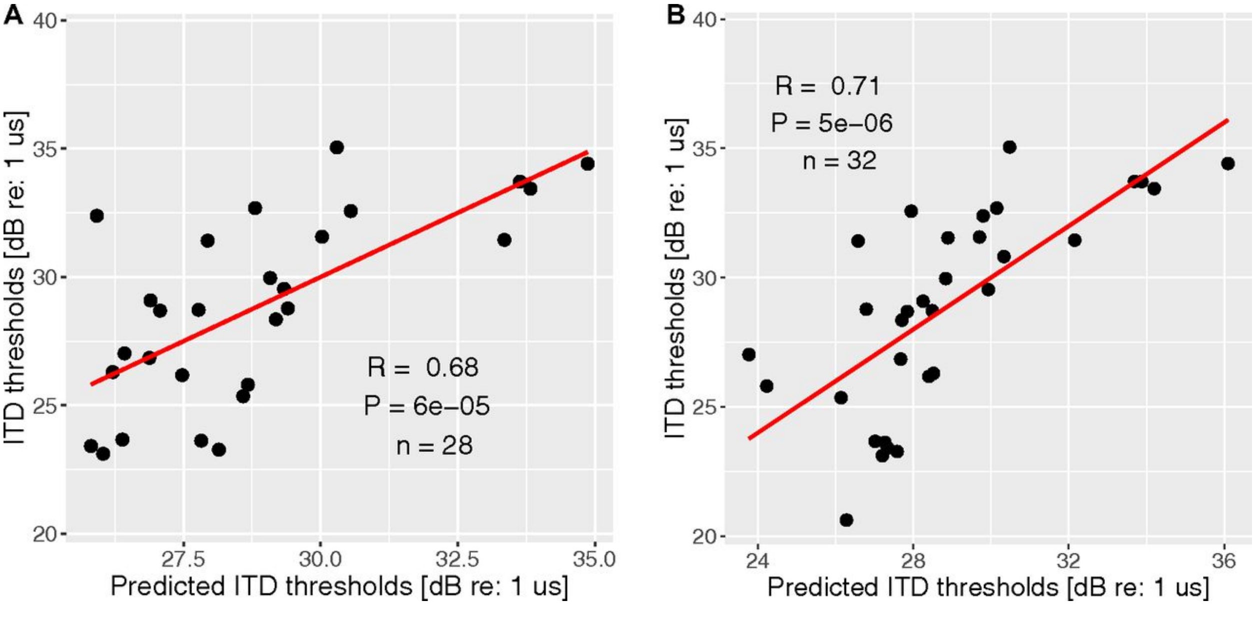


Figure 2.7. Model prediction of the ITD detection thresholds, based on the combination of lapse rate and slope (60- to 180-μs condition; A), or the combination of lapse rate and EEG latency (B). Please refer to Table 2.1 for the variance explained by each factor.

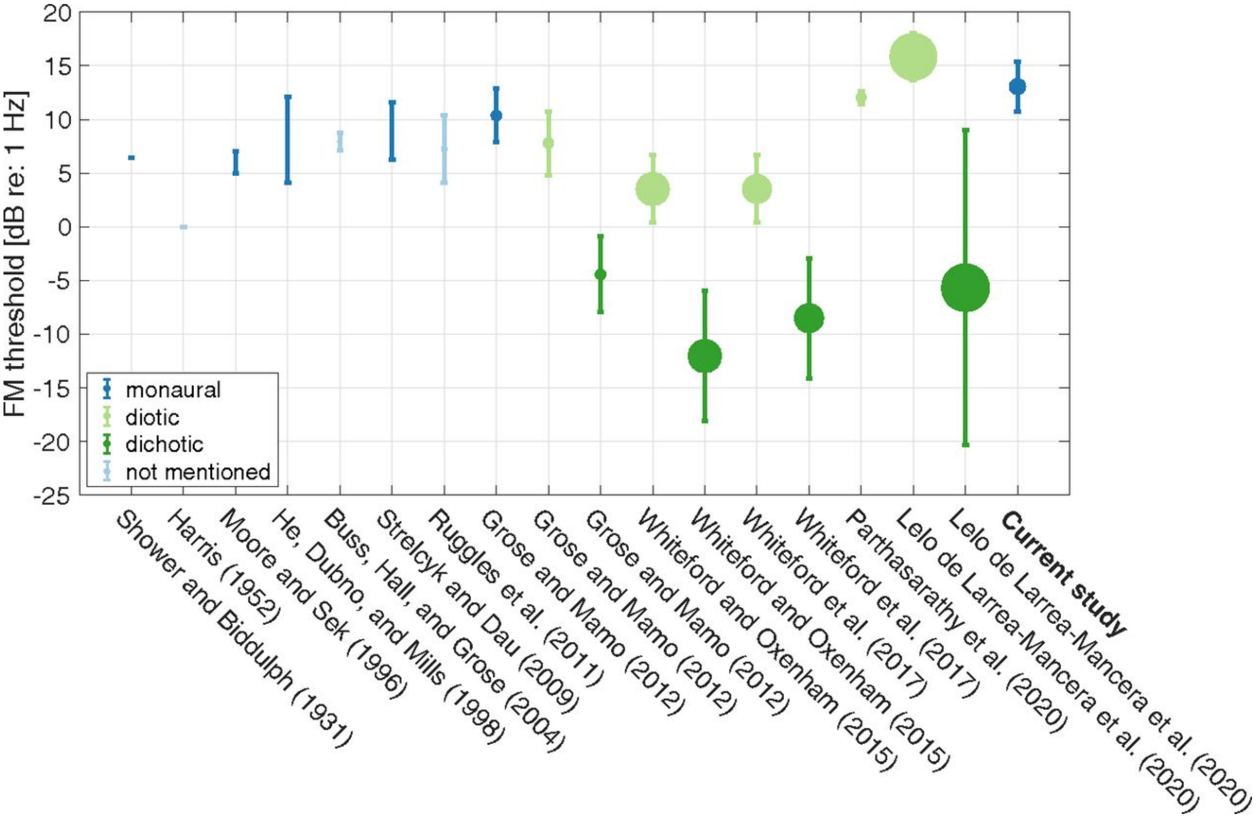


Figure 2.8. A sample of published reports of FM detection thresholds for comparison (Buss et al., 2004; Grose & Mamo, 2012; Harris, 1952; He et al., 1998; Lelo de Larrea-Mancera et al., 2020; Moore & Sek, 1996; Parthasarathy et al., 2020; Ruggles et al., 2011; Shower & Biddulph, 1931; Strelcyk & Dau, 2009; Whiteford et al., 2017; Whiteford & Oxenham, 2015). Error bar is 1 SD. The size of the dot represents the number of subjects (Whiteford and Oxenham, 2015 has the most subjects; N = 100). Stimulus parameters such as stimulus level, carrier frequency, and modulation frequency in the cited studies are similar to those used in the current study, with slight differences (Ruggles et al., 2011; Strelcyk and Dau, 2009 used carrier at 750 Hz). Some threshold values are approximate from figures [e.g., mean and SD had to be estimated based on median and range in the box whisker plots in Whiteford et al., 2017; Whiteford and Oxenham, 2015]. The mean and SD from the young and middle-aged group from Grose and Mamo, 2012 were combined to generate a single data point. Some authors expressed the threshold in terms of ΔF/Fc, where ΔF is frequency deviation, and Fc is the carrier frequency. Moore and Sek, 1996 used ΔF that was in two directions, i.e., peak-peak. Subjects from some studies were highly experienced in psychoacoustic tasks hence the thresholds were very low/good. Whiteford et al., 2017; Whiteford and Oxenham, 2015 obtained thresholds that fall in the lower end of the results of the current study from a very large number of subjects. This may be because their subjects were younger TH listeners and the stimuli were presented diotically and dichotically instead of monaurally.

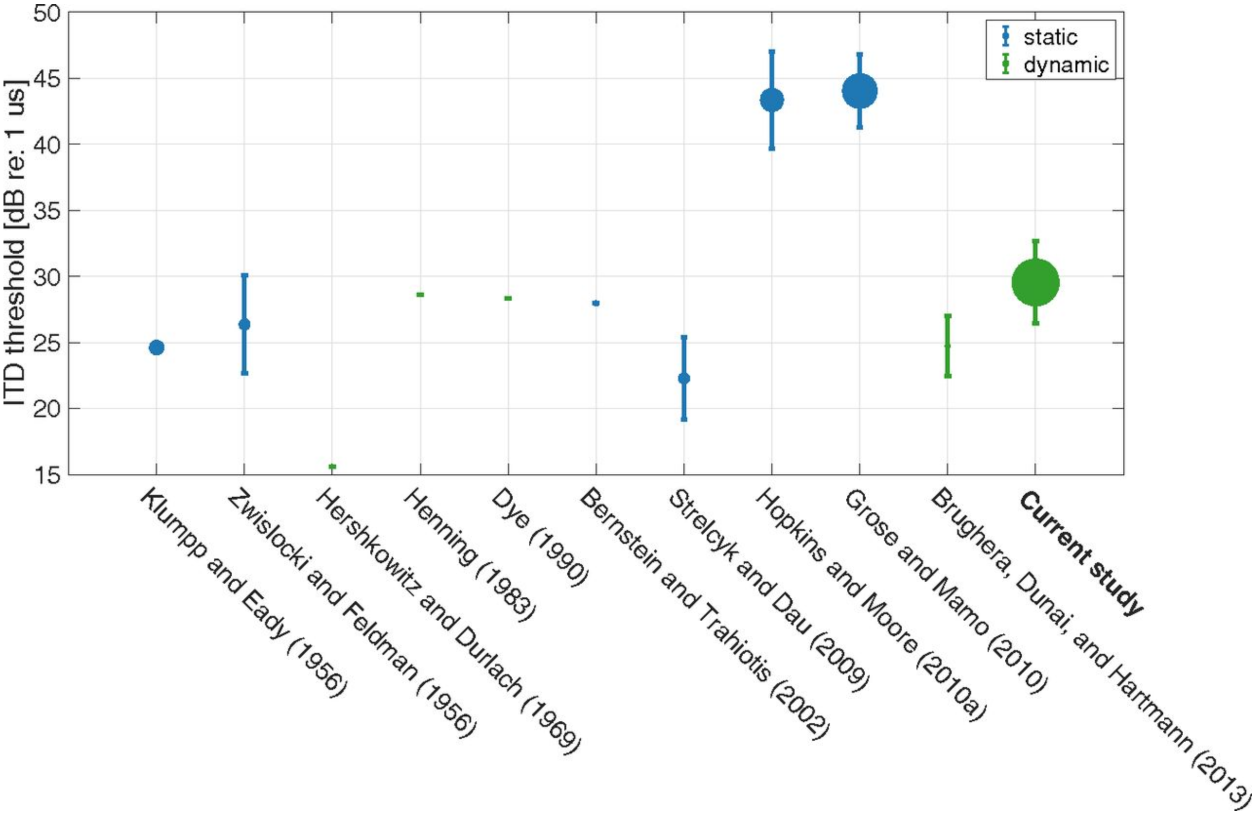


Figure 2.9. A sample of published reports of ITD detection thresholds for comparison (L. R. Bernstein & Trahiotis, 2002; Brughera et al., 2013; Dye, 1990; Grose & Mamo, 2010; Henning, 1983; Hershkowitz & Durlach, 1969; Hopkins & Moore, 2010; Klumpp & Eady, 1956; Strelcyk & Dau, 2009; Zwicker, 1956). Error bar is 1 SD. The size of the dot represents the number of subjects (the current study has the most subjects; N = 36). Stimulus parameters such as level and carrier frequency in the cited studies are similar to those used in the current study, with slight differences [Strelcyk and Dau, 2009 used carrier at 750 Hz]. Note that some threshold values were extracted approximately from figures rather than direct numerical reports. Some of the studies used stimuli with the leading ear switching from one side to the other (labeled "dynamic," marked in green color), whereas others presented an ITD only in the target intervals, with the reference being the midline (labeled "static," marked in blue color). Note that the values from Hershkowitz and Durlach, 1969 and Brughera et al., 2013 were halved since the authors used ITD/2 in each interval. The mean and SD from young and middle-aged cohort from Grose and Mamo, 2010 were combined to generate a single data point. Subjects from some studies were highly experienced in psychoacoustic tasks.

3. BETTER TEMPORAL FINE STRUCTURE PROCESSING REDUCES LISTENING EFFORT AND IMPACTS FROM REVERBERATIONS IN HUMANS

3.1 Introduction

Everyday human communication relies heavily on the auditory system's capacity to process complex sounds, such as speech, in the presence of interfering background noise. However, regardless of the complexity, all sounds that we hear can be decomposed into only two elements: the rapid variations in phase—temporal fine structure (TFS) and slower variations in amplitude superimposed on TFS—temporal envelope (Hilbert, 1906). The auditory system can robustly phase-lock to both TFS (Johnson, 1980) and envelope (Joris & Yin, 1992), with TFS phase-locking extending at least up to 1000 Hz (Verschooten et al., 2019). This fast phase-locking activity is unmatched by other sensory systems. For instance, only a selective population of mechanoreceptive neurons (e.g., rapidly adapting afferent—RA) in the somatosensory system can phase-lock to mechanical vibrations, and the phase-locking frequency limit is only around 50 Hz (Ahissar & Vaadia, 1990; Romo & Salinas, 1999; Shadlen & Newsome, 1994; Singer & Gray, 1995; Talbot et al., 1968). In addition, Ochoa and Torebjörk, 1983 showed that the ultimate perception of flutter sensation does not even depend on this phase-locked temporal information. Rather, it is the RA firing rate that delivers the sensory information. Similarly, despite exceptionally fast phase-locked activity in the auditory system (i.e., TFS coding), the upper-frequency limit progressively decreases along ascending auditory pathway, and the phase-locked temporal code gradually transitions into the rate code (Joris et al., 2004). How this metabolically consuming initial/peripheral phase-locked temporal code (Hasenstaub et al., 2010; Laughlin et al., 1998) contributes to the perception, and how its degradation is attributable to perceptual deficits is a fundamental question not only in the auditory-system research and clinical audiology but also among overall neuroscience community studying other sensory systems. Yet, the significance of phase-locking (in particular, TFS coding in the auditory system) is debated (Drullman, 1995; Oxenham, 2013; Oxenham & Simonson, 2009; Swaminathan & Heinz, 2012).

Studies suggest that TFS plays a role in sound localization (Smith et al., 2002; Yin & Chan, 1990) and pitch perception (J. G. Bernstein & Oxenham, 2003; Houtsma & Smurzynski, 1990; Moore, 1973) in quiet settings. Both spatial (F. J. Gallun et al., 2015; F. J. Gallun et al., 2013; Ihlefeld & Shinn-Cunningham, 2008; Jakien Kasey M. & Gallun Frederick J., 2018; Jakien Kasey M. et al., 2017; Srinivasan et al., 2016) and pitch cues (Bird & Darwin, 1997; Brokx & Nooteboom, 1982; Summers Van & Leek Marjorie R., 1998) could ease hearing in the presence of noise maskers, providing a release from masking (or masking release) of about 5 dB. Despite the implication of the relationship between TFS and masking release based on either pitch or spatial cues, the perceptual role of TFS in speech-in-noise hearing is still further debated. It is because the other component of sounds—the temporal envelope—could provide a similar amount of masking release through either pitch or spatial cues (Best et al., 2005; Oxenham & Simonson, 2009). Furthermore, low-frequency TFS-based spatial cues may be more susceptible to corruption by reverberation than high-frequency envelope-based spatial cues (Ihlefeld & Shinn-Cunningham, 2011). Indeed, reverberation, as one of the two primary sources of distortions of signals during transmission (the other being amplitude fluctuations), is lowest in the range of 2-8 kHz but increases substantially in the low-frequency region (Richards & Wiley, 1980), where phasing-locking of TFS is the most prominent. Therefore, in the presence of many other distortions and redundant cues, whether the phase-locked temporal coding of TFS would introduce additional masking-release benefits beyond those delivered by rate-code-based envelope information is still unclear.

Studies have investigated the role of TFS by acoustically isolating TFS and envelope from each other and independently manipulating the content of TFS, through a signal-processing technique called vocoding (Ardoint & Lorenzi, 2010; Hopkins & Moore, 2009; Hopkins et al., 2008; Lorenzi et al., 2009; Smith et al., 2002). However, establishing the precise role of TFS through vocoding is confounded by partial conversion of TFS to envelope due to cochlear processing (Oxenham, 2013; Swaminathan & Heinz, 2012). Recovery of the mixture of TFS and envelope renders the vocoding strategy ineffective for studying the role of TFS through isolation, which is why the perceptual roles of TFS and envelope coding in everyday hearing are still an ongoing debate. An alternative approach is to directly measure individual differences in TFS sensitivity and compare them to individual differences in speech-in-noise sensitivity. Many studies, such as Bharadwaj et al., 2015; McDermott et al., 2010; Whiteford et al., 2020, have successfully used this approach to address other fundamental questions. Unfortunately, this approach has not been used to investigate the role of TFS because there have not been established measures of TFS sensitivity at the individual level until our very recent study (Borjigin et al., 2021). In that study, we established a set of TFS-sensitivity measures at the individual level, through a comprehensive assessment of both behavioral and electrophysiological measurements. Here, we adopted the behavioral elements of those measures and modified them accordingly for an online data-collection platform that we custom-designed to circumnavigate the COVID-19-related restrictions on in-person measurements (Mok et al., 2020). In addition to TFS-sensitivity measurements, we also assessed the participants' speech-in-noise sensitivity under various types of noise interference. The association between the individual differences in TFS and speech-in-noise sensitivity suggests that better TFS processing does not benefit the listeners with more masking release, but reduces the impacts from reverberation and leads to less reaction time during the task suggesting lessened cognitive load or listening effort associated with speech hearing in adverse listening environments.

3.2 Materials and Methods

The primary goal of this study was to uncover the perceptual role of temporal fine structure (TFS) processing in everyday hearing, particularly speech-in-noise listening, by evaluating both TFS and speech-in-noise sensitivity at the individual level from a large cohort of typical-hearing, naïve participants. To circumnavigate COVID-19 related restrictions on in-person measurements, all data for this study battery, including TFS-sensitivity measurements, a reference measurement—interaural level difference (ILD), and speech-in-noise measurements, were collected on the remote-testing platform that was developed previously in our lab (Mok et al., 2020)—snaplabonline.

3.2.1 Participants

Two hundred participants, aged 19–54 years (mean = 30, std = 8, 102 males), were recruited from an online subject-recruitment platform—Prolific.co. 87% of the participants self-reported their first language as English. All participants self-reported having no hearing loss, neurological disorders, or persistent tinnitus, and they passed a headphone check and speech-in-noise hearing screening. The participants consented to participate in the study following Institutional Review Board (IRB) protocols at Purdue University and were compensated for their time. All participants completed the full study battery.

3.2.2 Experimental Design and Statistical Analyses

Screening Measurements

All measurements in this study including the screening measurements are listed in Table 3.1. Unlike in a research sound booth, with web-based testing, subjects used their personal computers and completed the tasks in the environments of their choice. One of the challenges of an online study for auditory research is the limited control over the participant's acoustic setup, such as the participant's headphones and listening environments. As an investigation of lower-level aspects of auditory processing (i.e., TFS processing), our study is especially sensitive to those testing parameters. So, we implemented a couple of measurements to ensure that the participants were using two channels, which is crucial since TFS, the interaural level difference (ILD), as well as speech-in-noise hearing tasks with spatial information require two-ear listening. Furthermore, given that hearing loss is a highly prevailing health condition (1 in 8 people aged 12 and older have hearing loss in both ears) (Lin et al., 2011), we also conducted a speech-in-noise test as a hearing screening to ensure that the participants were of typical hearing. For all screening measurements, as well as the measurements from the formal test battery that followed, the participants were instructed to adjust the loudness of the stimulus to a comfortable level before the experiments.

HEADPHONE-CHECK MEASUREMENTS. Two measurements were taken to ensure the appropriate use of headphones. 1) Listeners were asked to identify the softest of three low-frequency tones. The target was 6 dB softer than the two control (foil) tones. One of the control tones had a phase difference of pi between the two ears. This difference would induce an intensity reduction via acoustic cancellation if the stimulus was played by two speakers in the free-field. Therefore, without headphones, the listener would hear two soft tones instead of only one: one being the target and the other being the control tone with the pi-phase difference. Therefore, the task becomes much more difficult without wearing headphones hence more errors in the response. This measure is based on the methodology used by Woods et al., 2017. However, false positives could happen if the participants' computer only had one speaker (channel), where acoustic cancellation would not be possible through the free field. This method could not detect if participants' headphones have one or two channels either. So, the second test was carried out to ensure the use of the two-channel headphones. Subjects were asked to discriminate the direction (rising/falling/flat) of a low-frequency chirp (150-400 Hz) that was embedded in a background noise, configured in an N_0S_π setting. N here represents noise; S represents target signal; N_0S_π means the noise was identical in two ears, while the target signal's polarity was flipped in one ear, which introduces the so-called binaural masking level difference (BMLD) (Licklider, 1948). We chose the signal-to-noise ratio (SNR) such that the chirp would be difficult to detect with only one channel without the BMLD benefits through two channels. Therefore, the participants would make a lot more errors in their response if they were not wearing 2-channel or stereo headphones.

Table 3.1. All measurements conducted in this study. ITD: interaural time difference. ILD: interaural level difference. FM: frequency modulation.

Screening	**TFS Sensitivity**	**Speech Intelligibility**	**reference**
Headphone check	ITD detection	Speech in 4-talker masker	ILD detection
Speech-in-noise hearing screening	Binaural FM detection	Speech in steady-noise masker	

TFS Sensitivity Measurements

We previously established that a combination of binaural behavioral and electrophysiological (EEG) measurements of interaural time difference (ITD) sensitivity could more reliably reflect individual differences in TFS processing fidelity than monaural behavioral measurement of frequency-modulation (FM) detection and diotic (i.e., identical in two ears, no binaural cues) EEG measurement of frequency-following-response (Borjigin et al., 2021). Therefore, in this study, we adopted the behavioral component—the ITD detection—and added a binaural version of the FM detection. The EEG components were not included due to the restrictions from remote-testing. Importantly, our previous study also pointed out that those binaural metrics were effective only if the "non-sensory" factors that are irrelevant to TFS processing, such as attention and motivation, were controlled for (Borjigin et al., 2021). The "non-sensory" factors were quantified by using percent errors in catch trials that were embedded in the same measurements. In this study, we took a step further and implemented a stand-alone measurement—interaural level difference (ILD) detection. Similar to the two binaural temporal metrics used for quantifying individual differences in TFS sensitivity, ILD is also binaural but dependent on level instead of temporal coding of TFS processing. On one hand, having ILD detection as a surrogate measure helped control for individual variability in high-level capability to perform a behavioral task. On the other hand, it also further augmented the extrication of TFS processing from extraneous factors, by stripping away any individual variability that may have existed in the low-level brainstem circuitry for non-TFS binaural processing. This ensured that the individual differences we observed mostly reflected TFS processing, rather than the overall sensitivity of binaural hearing. These measurements (ITD, binaural FM, and ILD) are detailed below.

INTERAURAL TIME DIFFERENCE (ITD) DETECTION. The stimulus consisted of two consecutive 400-ms-long, 500-Hz tone bursts. Each tone burst was delivered to both ears, but with a certain time delay in one ear (i.e., ITD). This led to the perception of the sound coming from a lateralized position, from the same side as the ear receiving the stimulus without a time delay (i.e., leading ear). The leading ear for the ITD was switched from the first to the second tone burst, which led to a perception of the tone burst "jumping" from one side to the other. ITDs in octave steps from 2 to 128 us (8 repetitions for each step) were presented in random order. The tone bursts were ramped on and off with a rise and fall time of 20 ms to exclude audible transitions and to reduce reliance on onset ITDs. The gap between the two tone bursts was 200 ms. Some studies have shown that stimulus level could influence ITD detection performance (Hopkins & Moore, 2010; Strelcyk & Dau, 2009). Zwislocki and Feldman, 1956 measured the effect of sensation level on lateralization and found that the performance was worse for very low and very high sensation levels, whereas the performance was stable between 30 and 90 dB. As mentioned above, the participants were instructed to adjust the loudness level to a comfortable level to ensure that their performance was stable. We measured the subject's sensitivity to the "jumping" perception, instead of static lateralized location, because absolute lateralization could be complicated by the following two factors. First, sounds with a large ITD could sometimes be heard either from the left or from the right. Second, individuals with asymmetric hearing may perceive the sound with a static ITD coming from the better ear, regardless of the magnitude and direction of ITD (Hopkins & Moore, 2010). The measurement was conducted in a two-alternatives forced-choice setting, where subjects were asked to report the direction of the "jump" (left-to-right or right-to-left) between the intervals using a button press. A separate "demo" block was included before the experimental blocks to familiarize the subject with the task. The detection thresholds were determined by a Bayesian approach (Kuss et al., 2005), using the toolbox developed by Schütt et al., 2016. All other measurements in this study, including TFS-sensitivity, ILD sensitivity, and speech-in-noise tests (Table 3.1), adopted the same method for estimating the detection thresholds.

BINAURAL FREQUENCY MODULATION (FM) DETECTION. Borjigin et al., 2021 suggested that monaural FM measurement mostly reflected "non-sensory" factors instead of TFS sensitivity. Since the binaural ITD measurement was proven to be better for capturing TFS sensitivity and there is an unambiguous theoretical link between binaural temporal hearing and TFS processing (Yin & Chan, 1990), monaural FM detection measurement was replaced by binaural FM detection measurement as an additional metric for TFS-sensitivity assessment. The stimulus consisted of a target and control. They were turned on and off with a rise and fall time of 5 ms to exclude audible transitions, with an inter-stimulus gap of 900 ms. The control was a 500-ms, 500-Hz pure tone. The target was the same, except that it had a 2-Hz frequency modulation. The direction of the modulation was set to be opposite in two ears to introduce binaural cues. A low FM rate (under 10 Hz) was chosen for TFS-sensitivity assessment because FM detection at fast modulation rates primarily reflects an FM-to-AM (envelope) conversion mechanism (Moore & Sek, 1996). FMs of magnitudes in octave steps from 0.1 to 3.2 Hz (8 repetitions for each step) were presented in random order. In a two-alternatives forced-choice setting, the subject clicked on the option that indicated the target interval with frequency modulations. No training was provided except for the "demo" block to orient the participants before the formal testing.

INTERAURAL LEVEL DIFFERENCE (ILD) DETECTION. The ILD detection thresholds were measured as a baseline reference to quantify the involvement of extraneous "non-sensory" factors, such as lapses in attention and variations in motivation. The stimulus was consecutive two pure-tone bursts, with their frequency set at 4 kHz to make sure it was beyond the frequency limit of TFS processing (Verschooten et al., 2019). Similar to the measurement of ITD detection, one tone was lateralized on one side, and the other on the opposite side. The perception of lateralization was induced by a difference in the stimulus levels (i.e. ILD) in two ears. Within each trial, the participant perceived the tone burst "jumping" from one side to the other. ILDs in octave steps from 0.1 to 3.2 dB (8 repetitions for each step) were presented in random order. The subjects clicked on the option indicating the direction of the jump in a 2AFC setting. For both ITD and ILD, the spatial cues were delivered via the participant's headphones, which was acceptable as previous studies suggest that the perception of sound localization through the ear or headphones was nearly as good as that in free-field experiments (Wightman & Kistler, 1989).

Speech-in-noise Sensitivity Measurements

Each trial was a mixture of a target word and a masker. The masker was either a babble mixture of 4 speakers or steady-state noise. The target word was prompted with instructions in the same voice, saying: "Please select the word…". The masker began only slightly before the onset of the target word so that the listener had a chance to orient themselves about the target based on the unmasked prompt. A single word instead of a full sentence was chosen as the target to minimize the involvement of confounding factors such as individual differences in auditory working memory. Studies have shown that children as young as 18 months tended to achieve the same speech reception thresholds (SRT) as adults when a single word was used for a speech-in-noise intelligibility task (Litovsky, 2005; Murphy et al., 2011; Yuen Kevin C. P. & Yuan Meng, 2014). However, the detection of a full sentence takes years to develop: "fast-learning" children could achieve an adult level at around 8 years of age (Cameron & Dillon, 2007) but in other cases, not until 12 years of age (Vaillancourt et al., 2008; Werner, 2017).

Participants were tested with 10 different target-masker configurations, as shown in Table 3.2. Half of them contained 4 speech-babble-masker and 1 non-speech-masker conditions. 4 speech-babble-masker conditions include conditions with pitch cues, spatial cues, both pitch and spatial cues, and no cues at all (i.e., control condition). The non-speech masker condition had a steady noise as the masker without any cues. The other half had the same 5 conditions but with an addition of room reverberation. Reverberation was added to investigate the role of TFS processing in reverberation because the presence of reverberation generally degrades hearing performance (Ihlefeld & Shinn-Cunningham, 2011). The presentation order of these 10 different test conditions was randomized across trials to reduce the formation of expectations, which could improve the subject's performance (Russo & Pichora-Fuller, 2008). More details regarding the configurations of pitch and spatial cues, reverberation, and the creation of steady-noise masker are as follows.

Table 3.2. List of conditions for speech-in-noise tasks.

	Babble Noise				**Steady Noise**
Reverberant	Control: no cues	Pitch	Space	Pitch+Space	No Cues
Non-reverberant	Control: no cues	Pitch	Space	Pitch+Space	No Cues

PITCH CUES. To make sure that the pitch difference between the target and masker remains constant across trials, audio recordings were first processed to remove the natural pitch fluctuations or intonation, and then the flattened pitch contour of the target and masker was moved to a fixed frequency, as shown in Figure 3.1. More specifically, the pitch of the male target was fixed at 95 Hz, and the pitch of the female target was at 245 Hz; among the 4-talker babble background, the male masker's pitch was set at 85, 90, 100, and 105 Hz, respectively, and the female masker's pitch was set at 235, 240, 250, and 255 Hz, respectively. Note that the target and masker of the same sex were at similar pitch values but with a small difference to ensure that the participant could distinguish the target from the masker. The pitch was flattened for all other configurations (i.e., control, space, pitch+space, and non-speech noise masker) as well to maintain consistency throughout the study.

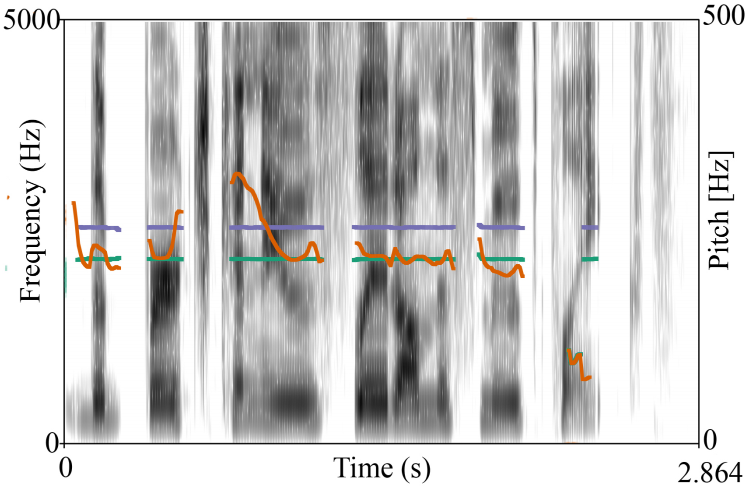


Figure 3.1. The spectrogram of a sentence: "The birch canoe slid on the smooth planks." The orange curve is the pitch contour with natural fluctuations; flattened pitch contour is shown in green; the flattened pitch contour that was moved to 255 Hz is shown in purple.

Spatial Cues. Spatial cues help listeners better localize and hence more easily identify the target of interest in poor listening conditions (Fay, 2005). To simulate the perception of spatial separation, the polarity of the target signal in one ear was flipped while the masker signal was kept the same in both ears. This configuration is usually denoted as N_0S_π. The conditions without any spatial cues are referred to as N_0S_0. A lower speech reception threshold (better performance) is typically observed in the N_0S_π condition. The difference in speech reception thresholds between the N_0S_π and N_0S_0 conditions is commonly referred to as binaural masking level difference (BMLD, i.e., spatial masking release) (Levitt & Rabiner, 1967). Studies have shown that a difference in the spatial location, as well as the pitch of the target and masker, would result in the benefit of a masking release of around 5 dB (Best et al., 2005; Oxenham & Simonson, 2009). Note that masking release is quantified as the threshold difference between the conditions with cues (i.e., hence easier) and the control condition without any cues.

STEADY-NOISE MASKER AND REVERBERATION. A steady noise masker was used as a reference for the energetic masking, in contrast to the informational masking from a speech-babble masker. To minimize the intrinsic modulations, which is a source of informational masking that is inherent in noise (Culling & Stone, 2017; Hartmann & Pumplin, 1988; Kohlrausch et al., 1997), a previous method from (Stone & Moore, 2014) was used to create the steady-noise masker. All original audios were recorded or generated in an idealistic anechoic chamber with no reverberation. To simulate reverberate conditions, the signals were convoluted with an impulse response that was recorded in a bar.

Statistical Analyses

Due to the challenge to recruit anonymous online subjects back for multiple follow-up sessions, a 3-hour, multiple-session, and in-person test battery was condensed into a single 1-hour online session, leading to fewer trials in each test. It is particularly true for speech-in-noise measurements since it contained a total of 50 conditions (5 signal-to-noise ratios, i.e., difficulty levels, for 10 different target-masker configurations). Consequently, only 4 trials were scheduled for each condition due to time constraints. They were not sufficient to generate a psychometric curve for threshold estimates. Data from multiple participants had to be converged and a statistical method—Jackknife resampling—was used to create "more" trials for each participant. More specifically, within the same group, a subject's data were removed and replaced with the data that were merged from the rest of the subjects. This process was repeated for all participants in the same group. Group-level mean and variance were estimated by M and V/n, respectively, where M and V are the mean and variance of the Jackknife thresholds, respectively, and n is the number of samples within the group (Efron & Stein, 1981). Note that the measurements that were used to group the participants (i.e., TFS or ILD measurements, see section 3.2.2) had a sufficient number of trials for threshold estimates since all three measurements contained only 6 conditions as opposed to 50 conditions in the speech-in-noise measurements. The difficulty in keeping the subjects over multiple sessions might be mostly due to the anonymous nature of the online participants. In future studies, the recruitment of identified participants with contact information and/or setting a reward for completion would help maintain online participants over multiple sessions, hence more trials in the measurements.

As mentioned previously in section 3.2.2, among 10 speech-in-noise conditions, 5 of them were simulations of speech-in-noise mixtures in anechoic environments, and the other 5 conditions were generated with room reverberation. To emphasize the effects of reverberation, as shown in 3.5a, data from all 5 speech-in-noise configurations were merged into one for both reverberant and non-reverberant conditions. The variance of the merged data was estimated by using inverse-variance weighting, where the variance from each condition was weighted in inverse proportions to its variance (Hedges, 1982; Sinha et al., 2011).

3.3 Results

3.3.1 Performance trend in online data is in excellent agreement with lab-based data.

Web-based online experiments that have been used in many areas of human behavioral research have not yet been adopted in psychoacoustic research. It is mainly due to the challenges in fulfilling stringent requirements for high audio quality, proper headphone setup (single vs dual channels), quiet listening environments, as well as a good estimate of the participant's hearing status. Despite the challenges, to cope with the suspension on in-person measurements posed by COVID-19, we custom-developed a web-based psychoacoustic platform and demonstrated that the remote setup was viable to produce a close match between online and in-person data across several measurements (Mok et al., 2020). One of the tremendous advantages of online study is to allow simultaneous hence fast data collection from a large number of participants. We rapidly collected all data presented here from 200 typical-hearing participants using this online data-collection platform.

In our recent previous study (Borjigin et al., 2021), we established a battery of temporal fine structure (TFS) sensitivity measurements on the individual level, via thorough examination of several behavioral and electrophysiological measures. Based on the conclusions, a couple of binaural-hearing (i.e., two-ear) sensitivity measurements were chosen in this study as the metrics for assessing individual differences in TFS processing, including binaural frequency modulation (FM) detection and interaural time difference (ITD) detection measurements. It is established that binaural temporal processing depends on TFS (Smith et al., 2002; Yin & Chan, 1990). In the ITD detection measurement, two ears receive copies of the stimulus TFS with a small timing delay (i.e., ITD) in microseconds because of the difference in traveling distance from sound to two ears. Similarly, binaural FM introduces frequency fluctuations in opposite directions in two ears, leading to a timing difference in TFS on two sides. Figures 3.6 and 3.7 are across-study comparisons in binaural FM detection and ITD detection thresholds, respectively. Binaural-hearing online data are comparable to our previous in-person results (Borjigin et al., 2021) as well as the results from other studies. For both binaural FM and ITD detection measurements, the standard deviation is small considering the number of participants involved in this study (n = 200).

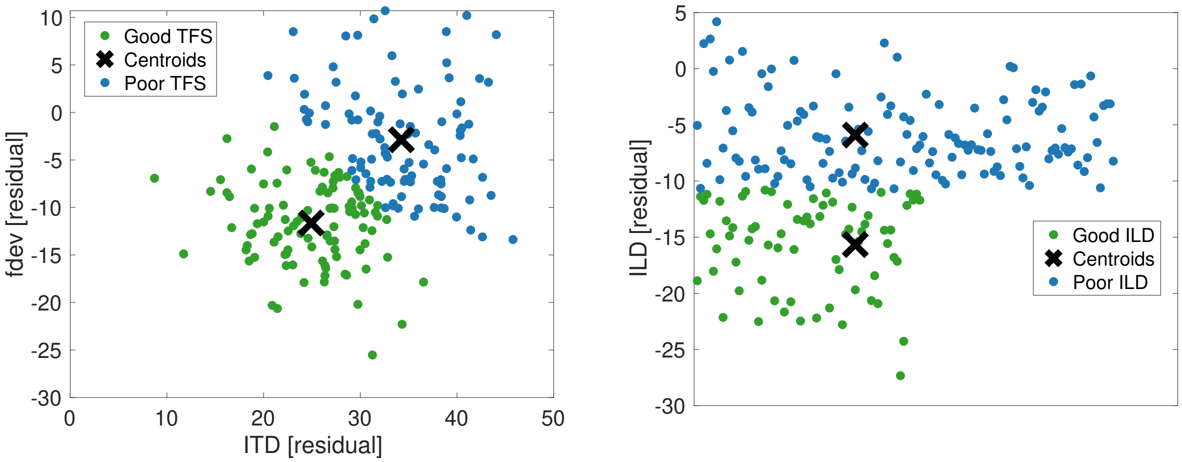
3.3.2 Binaural temporal sensitivity measures could capture individual differences in TFS processing fidelity.

The scatter plot (Figure 3.2a) is a direct illustration of the individual differences that we observed in the binaural temporal sensitivity measurements—binaural frequency modulation (FM) and interaural time difference (ITD) detection measurements. In our previous study, we found that the metrics that had been commonly used in literature to assess TFS sensitivity (including the ITD and binaural FM detection) were prone to the impact of extraneous variables, such as attention and motivation, hindering them to be ineffective (Borjigin et al., 2021). “Non-sensory” factors that were irrelevant to TFS processing had to be controlled for in order for the measurements to reflect individual differences that were truly driven by TFS processing. The “non-sensory” factors were quantified for each subject using percent-wrong scores in catch trials. Here, in this study, we took a step further and implemented a stand-alone surrogate measure that does not depend on TFS processing—interaural level difference (ILD) detection measurement—to control for individual variability from “non-sensory” factors.

Having ILD detection as a surrogate measure not only helped control for "non-sensory" factors but also further augmented the extrication of TFS processing from extraneous factors, by disentangling TFS sensitivity from binaural-hearing processing. The two TFS metrics used in this study both involved binaural temporal processing, by introducing a small timing difference in the stimulus TFS across two ears. Similar to the timing difference between two ears (i.e., ITD), the copies of sound reaching two ears also differ in sound intensity (i.e., ILD) due to the attenuation by the head in between two ears. ILD is a level-dependent, non-temporal binaural cue that does not rely on TFS processing but still requires the coordination of two ears. Regressing out the ILD scores from TFS-related binaural metrics not only reduces the influence from individual variability in high-level capability to perform a behavioral task but also further strips away any individual variability that may exist in low-level brainstem circuitry for binaural processing. This ensured that the individual differences we observed mostly reflected TFS processing, rather than binaural-hearing sensitivity.

The thresholds from TFS-sensitivity measures—binaural FM and ITD detection—are plotted against each other in Figure 3.2a. Note that ILD detection thresholds were controlled for both measurements so that the influence from "non-sensory" factors as well as general binaural-processing sensitivity was minimized. There is a statistically significant correlation between the adjusted binaural FM and ITD detection thresholds (R = 0.3, P = 1.83e-5). Using a clustering algorithm, subjects were divided into two groups based on their TFS-sensitivity thresholds—good vs poor-TFS group. Figures 3.3a and 3.3b demonstrate a clear separation of two groups' psychometric curves in the binaural FM and ITD detection measurements, respectively. However, there's no such separation in the ILD detection measurements between the two groups (Figure 3.3c). This contrast assures that the two groups in Figure 3.2a were formed based on individuals' TFS sensitivity, rather than other extraneous factors.

In addition to TFS sensitivity, individuals were also grouped based on their general ability to perform a behavioral task as well as TFS-irrelevant binaural processing sensitivity, using their ILD detection thresholds as shown in Figure 3.2b. Note that, similar to Figure 3.2a, TFS sensitivity was controlled for the ILD measurements by regressing out the combination of ITD and binaural FM detection thresholds from ILD thresholds. A group-level separation



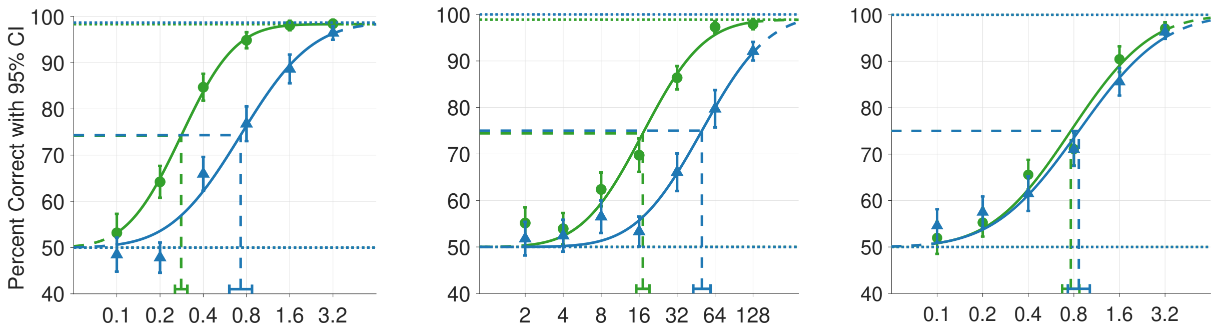
(a) grouping based on TFS sensitivity (b) grouping based on ILD sensitivity

Figure 3.2. (a). Cluster assignment based on ITD and FM thresholds. Residual means that ILD thresholds were regressed out from both ITD and FM thresholds. fdev is the frequency deviation or fluctuation in the FM stimulus. (b). Cluster assignment based on ILD sensitivity, with TFS sensitivity regressed out.

in the psychometric curves that is present in Figure 3.3f, but not in Figure 3.3e or 3.3d suggests that individuals were grouped based on "non-TFS" variability in Figure 3.2b. Regrouping the participants using this non-TFS metric also helped validate the observed group-level effect of better TFS processing on speech-in-noise measurement outcomes that will be detailed in section 3.3.4.

3.3.3 Better TFS sensitivity does not introduce additional masking-release benefit.

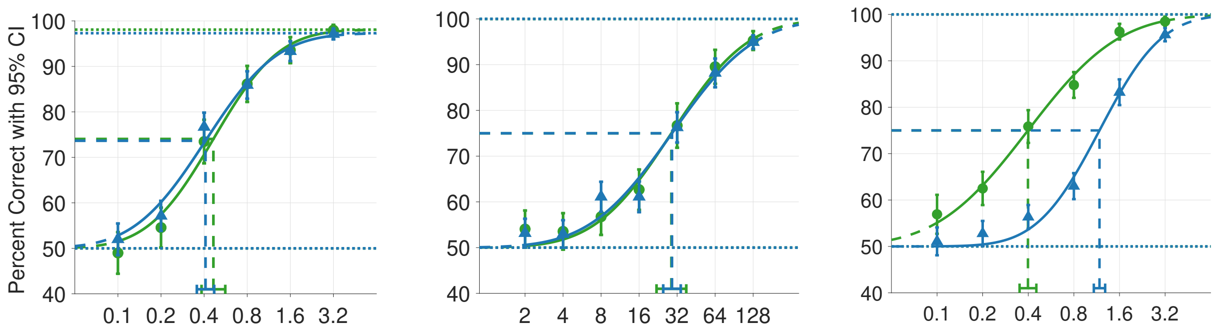
To understand the functional role of TFS in everyday hearing, in addition to measuring TFS sensitivity, we also measured participants' speech intelligibility under various types of noise interference, which was designed to simulate typical social listening settings. During the measurement, participants were asked to identify the target speech while ignoring the interfering background noise. Figure 3.4a shows two TFS-sensitivity groups' speech intelligibility performance in 4 different types of background noise: pitch, space, pitch+space, and steady noise. Except for the "steady-noise" being a non-speech noise interference, all



(a) fdev [dB re: 1Hz]

(b) ITD [us]

(c) ILD [dB]



(d) fdev [dB re: 1Hz]

(e) ITD [us]

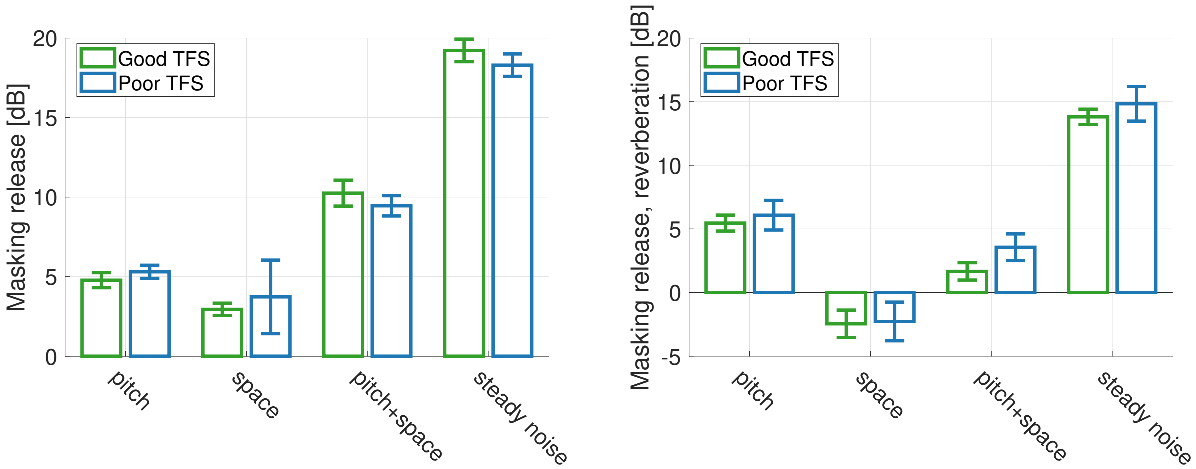
(f) ILD [dB]

Figure 3.3. (a)-(c): psychometric curves for good vs poor TFS-sensitivity groups. Left: ITD; mid: binaural FM; right: ILD. (d)-(f): psychometric curves for good vs poor ILD-sensitivity groups. Note that the percent correct at each experimental parameter is the average across individuals in the same group, whereas the standard error bar was calculated after Jackknife resampling (see Methods section 3.2.2).

other backgrounds are a 4-talker speech babble. In the "pitch" condition, the target and background have a pitch difference: e.g., if the target is a female speaker, the background babble would be produced by 4 different male talkers. With this pitch cue/difference, a listener typically could more easily identify the target, as compared to a condition where both target and background have the same pitch (i.e., reference condition. Note that the target and background are also spatially co-located). This benefit is often referred to as a release from masking or simply masking release, which is shown on the vertical axis in Figure 3.4a. In the "space" condition, instead of a pitch difference, the target and background are spatially separated from each other, which also introduces masking release as opposed to the condition where the target and background are co-located at the same spatial location (i.e., reference condition). In Figure 3.4a, the masking release based on either pitch or spatial cues is about 5 dB, which is consistent with the current literature (Bird & Darwin, 1997; Brokx & Nooteboom, 1982; F. J. Gallun et al., 2015; F. J. Gallun et al., 2013; Ihlefeld & Shinn-Cunningham, 2008; Jakien Kasey M. & Gallun Frederick J., 2018; Jakien Kasey M. et al., 2017; Srinivasan et al., 2016; Summers Van & Leek Marjorie R., 1998). Furthermore, the masking-release benefit seems to be additive (i.e., ∼10 dB) when both pitch and spatial cues were presented to the listeners, as shown in the "pitch + space" condition. The masking release is almost 20 dB when the background noise was switched to non-speech steady noise from babble speech (i.e., reference condition), as shown in the "steady noise" condition, suggesting that the masking power mostly comes from a background that is similar to the target, which is often referred to as "informational masking" (Arbogast et al., 2002; G. Kidd & Colburn, 2017; Viswanathan, Shinn-Cunningham, et al., 2021).

More importantly, between good and poor TFS-sensitivity groups, there is no significant group-level difference in terms of masking release across all conditions. This is consistent with several other studies suggesting that better TFS processing does not necessarily benefit a listener with more masking release (Best et al., 2005; Freyman et al., 2012; Oxenham & Simonson, 2009). Figure 3.4b demonstrates the masking release in the same four conditions as in Figure 3.4a, except for an addition of reverberation to all conditions. Note that the reference condition (i.e., babble-speech condition, with no pitch or spatial cues) was also simulated with reverberation. In general, reverberation reduced masking-release benefits across conditions, except for the "pitch" condition. Indeed, studies have shown that reverberation has less impact on monaural cues such as amplitude modulation and harmonics (Culling et al., 1994; Ruggles et al., 2011; B. Shinn-Cunningham et al., 2017). In contrast, the "space" condition was affected the most, which resulted in a negative masking release. It aligns with the previous findings that spatial cues are more subjective to corruption and less reliable in reverberation (Ihlefeld & Shinn-Cunningham, 2011; Palomäki et al., 2004; Ruggles et al., 2011). Again, similar to the non-reverberant conditions, there is no group-level difference in terms of masking release, due to better TFS sensitivity.

Masking-release measure was shown to be dependent on the signal to noise ratio (SNR) of the test conditions (J. G. W. Bernstein & Brungart, 2011; J. G. W. Bernstein & Grant, 2009; Freyman et al., 2008; Freyman et al., 2012; Oxenham & Simonson, 2009). Little masking release is observed if SNR is too low (i.e., noisy) or too high (i.e., quiet), where listening cues such as pitch and spatial cues are often not used towards masking-release advantage when the task is either too difficult or too easy. In this study, we calculated masking release at threshold levels, which fell in between the ceiling and floor SNRs where masking release was reported to be minimum/saturated with the least amount of individual variations. Therefore, the factor of SNRs should not be the reason why we did not see a group difference in masking release across conditions.



(a) masking release, without reverberation

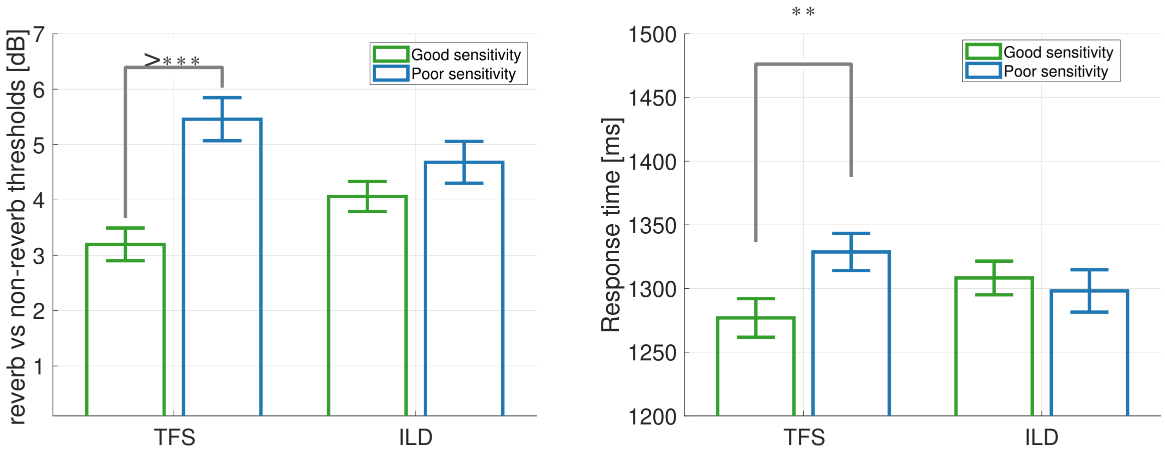
(b) masking release, with reverberation

Figure 3.4. Masking release across conditions. The height of the bars represents the mean, error bars represent 1 standard deviation. Masking release was calculated by subtracting the speech reception threshold in each condition from that of the control condition. Note that the control condition in (a) does not have reverberation, whereas the control condition in (b) contains reverberation. A positive masking release means that the speech reception threshold was lower/better than that of the control condition.

3.3.4 Better TFS processing reduces the impact from reverberation and lessens cognitive load associated with speech-in-noise hearing.

Although the pitch-based masking release shown in Figure 3.4b was not affected by the addition of reverberation, as compared to Figure 3.4a, the absolute speech reception thresholds in both pitch and reference conditions were dramatically increased (i.e., worsened) by reverberation, so were those from other conditions. Note that the masking release is the difference between the speech reception threshold in a certain condition and that of the reference condition. Therefore, in addition to masking release, we also calculated the threshold increase from non-reverberant to reverberant conditions, as indicated by the height of the bars in Figure 3.5a. When the participants were grouped based on their TFS sensitivity (Figure 3.5a, left), two groups differed significantly in terms of the increase/worsening in speech reception threshold due to reverberation: the poor-TFS sensitivity group suffered much more in reverberant settings than their counterpart ($p<0.0001$). When the participants were grouped based on their non-TFS sensitivity (i.e., ILD sensitivity), there was not anymore a significant group difference (Figure 3.5a, right), which validates that the group difference was driven by the individual differences in TFS sensitivity, instead of other "non-sensory" factors. This result shows that better TFS sensitivity could help reduce the impact of reverberation in noisy listening settings, which is often the case for most everyday listening situations. Indeed, previous studies suggest that TFS-based spatial cues are more susceptible to corruption due to reverberation (Ihlefeld & Shinn-Cunningham, 2011). Having better TFS sensitivity might have provided the listeners with a protective shield against reverberation, leading to a significantly less increase/worsening in speech reception threshold than their poor-TFS counterpart.

The Speech reception threshold is one of the most common metrics that have been used to estimate an individual's speech-hearing sensitivity in noise. However, a behavioral threshold does not give a full picture. Two listeners with the same speech reception threshold could have experienced the task with different levels of difficulty. To investigate the involvement of cognitive efforts beyond threshold estimates, we calculated each individual's reaction time, by subtracting the time stamp of the button press from that of the stimulus offset. The reaction time is indicated by the height of the bars in Figure 3.5b, which agrees with the result from the reaction-time measurement in Apoux et al., 2001. When the participants were grouped based on their TFS sensitivity (Figure 3.5b, left), the group with better TFS sensitivity had a significantly shorter reaction time than their poor-TFS counterpart (p=0.007), suggesting less cognitive effort required to perform the task. Similar to the previous analysis, when the listeners were re-grouped based on their non-TFS sensitivity (i.e., ILD sensitivity) as shown in Figure 3.5b (right), the two groups did not differ significantly from each other. This contrast helps confirm that it is the TFS sensitivity that was the main driving factor for the group difference in reaction time, indicating a possible benefit of reduced cognitive load due to better TFS processing.



(a) threshold increase due to reverberation

(b) reaction time

Figure 3.5. (a) group differences in the increase of speech reception threshold due to reverberation and (b) in reaction time. Data were merged across all 4 conditions from Figure 3.4.

3.4 Discussion

In this study, we investigated the functional significance of temporal fine structure (TFS) in everyday hearing. The relative contribution of TFS and envelope to speech hearing in noise is still debated mainly due to the limitation of the vocoding technique, which has been one of the most commonly used methods to study TFS and envelope in the literature (Ardoint &

Lorenzi, 2010; Hopkins & Moore, 2009; Hopkins et al., 2008; Lorenzi et al., 2009; Smith et al., 2002). Vocoding allows researchers to acoustically isolate TFS and envelope from the intact sound stimulus and study them independently. Despite being conceptually straightforward, vocoded stimulus fails to maintain the TFS content as some portion of the TFS would be converted into the envelope after the stimulus is processed by the cochlea (Oxenham, 2013; Swaminathan & Heinz, 2012). Therefore, vocoding doesn't isolate TFS and envelope from each other as expected. In addition, participants may just use TFS cues differently when they are presented with this novel stimulus that never exists in everyday listening environments. In this study, we avoided this limitation by using intact stimulus through an individual-difference approach. We modified and adopted the individual measurements of TFS processing from our recent study (Borjigin et al., 2021). We also conducted a comprehensive speech-in-noise intelligibility measurement on every individual. The direct comparison between the individual differences in TFS processing and speech intelligibility under various types of noise interference allowed us to reveal the perceptual role of TFS coding in everyday hearing: the results from 200 online participants suggest that better TFS processing does not necessarily introduce more masking-release benefits for speech-in-noise hearing, but better TFS sensitivity provides listeners with the advantage of having a more resilient hearing in reverberation and helps decrease reaction time hence indicating reduced listening efforts when conversing in noisy settings.

It is somewhat surprising not to see any additional benefit of spatial masking release from the "good-TFS" group in contrast to the "poor-TFS" counterpart (Figure 3.4a) as there is a well established theoretical link between phase-locked TFS processing and binaural temporal processing aspect of spatial hearing (Smith et al., 2002; Yin & Chan, 1990). Low-frequency acoustic TFS is coded temporarily in the auditory nerve through phase locking, which refers to the phenomenon where neurons only fire to a certain phase within a periodic signal. The time intervals between successive neural firings are extracted by the brain as a temporal code of the stimulus. Temporal code is used by binaural brainstem circuits for sound localization. Therefore, we hypothesized that individuals with better TFS sensitivity would benefit more from the spatial cues in speech-in-noise tasks. The null result could be due to the choice of interaural level difference (ILD) detection task as a reference measurement for non-TFS factors. Similar to the binaural temporal measurements of TFS processing (see Methods section 3.2.2), ILD also activates binaural circuits, but the ones that do not depend on the temporal code from TFS processing (Tollin et al., 2008). Assuming that ILD-activated binaural processing sensitivity shares some commonality with binaural temporal processing fidelity, regressing out the ILD scores from the binaural TFS measurements might have just stripped away individual differences that may have existed in binaural temporal processing. It could explain the lack of significant group differences in spatial masking release. Note that there was no significant group-level difference when participants were grouped based on their ILD sensitivity either (with TFS sensitivity measures regressed out from the ILD thresholds, not shown). Another possibility lies in the fact that the ability to reliably report binaural perception is challenging and typically requires extensive training (Stecker & Gallun, 2012). In this study, no training was provided to the participants, except for the initial demonstration trials for orientation. Indeed, the variation across individuals within the poor-TFS group in spatial hearing condition was unusually broad compared to other conditions (Figure 3.4a). Passive physiological measurements (e.g., EEG) of binaural temporal sensitivity, where active engagement is not required, could help better reflect an individual's spatial-temporal hearing in noisy settings. Papesh et al., 2017 demonstrated a significant brain-behavior correlation between auditory evoked potential measures of interaural phase difference (IPD: it introduces ITD) and spatial release from masking in speech-in-noise listening task. The null result might also be due to the small individual differences in the ITD thresholds considering all of our participants were of typical hearing. These small differences might not have translated into spatial release from masking. Regardless, our result is not a piece of evidence against the agreement regarding the role of TFS processing in the binaural temporal processing aspect of spatial hearing (Brughera et al., 2013; Churchill et al., 2014; Macpherson & Middlebrooks, 2002; Smith et al., 2002; Wightman & Kistler, 1992; Yin & Chan, 1990). Rather, the result emphasizes the fact that TFS processing is a necessary but not sufficient condition for spatial hearing benefit in noise. In other words, good TFS processing is key for successful binaural temporal processing. But a listener also needs a good binaural processing circuitry and/or needs to be sufficiently trained to better utilize the temporal code to benefit from spatial cues in speech-hearing-in-noise tasks.

Similarly, there was no significant group difference in pitch-based masking release. While it is established that TFS processing is crucial for the spatial localization of low-frequency sounds, whether TFS is needed for pitch perception has still been debated since over 150 years ago (Ohm, 1843; Seebeck, 1841). Behavioral studies in humans indicate that low-frequency periodic sounds result in a stronger pitch percept than high-frequency sounds (J. G. Bernstein & Oxenham, 2003; Houtsma & Smurzynski, 1990; Moore, 1973). The pitch discrimination ability deteriorates at high frequencies above 4-5 kHz (Attneave & Olson, 1971; Oxenham et al., 2011), which coincides with the phase-locking limit in the auditory nerve (Palmer & Russell, 1986; Rose et al., 1967). It all suggests a possible role for TFS in pitch perception. Furthermore, recent studies suggest that the deficits in TFS coding among individuals with hearing loss as well as the lack of TFS information in cochlear implant sound coding strategies might explain the difficulties in speech hearing in fluctuating noise, where pitch perception is believed to play an important role (Lorenzi et al., 2006; Qin & Oxenham, 2003; Stickney et al., 2007). However, these results are subject to alternate interpretations, such as place coding (Oxenham, 2012, 2013). Place coding, or tonotopic coding, refers to a frequency-to-place mapping. Within the inner ear, for example, different place along the basilar membrane of the cochlea responds to different frequencies. This mapping is maintained from low to high frequencies, not only in the cochlea but also throughout the rest of the auditory pathways up to the auditory cortex. This is in contrast to temporal code from phase-locking as phase locking to higher frequency component is not preserved in higher stations of the auditory pathways: e.g., at the level of auditory cortex, the limit of phase locking reduces to 100-200 Hz (Wallace et al., 2000). Indeed, in some studies, pitch perception remained possible with pure tones at very high frequencies (Henning, 1966; Moore, 1973; Moore & Ernst, 2012), where phase-locking information of TFS is unlikely useful. It is possible that pitch perception is based on phase-locked TFS information at low frequencies, and the timing information becomes so weak at high frequencies that the available place coding information takes the lead in pitch perception. Some studies even suggest that place coding information may be even important at low frequencies and timing information alone may not be sufficient to produce robust pitch perception (Dreyer & Delgutte, 2006; Kohlrausch et al., 1997; Oxenham et al., 2004). Our results are in agreement with the place-coding interpretation, showing that there was no additional masking-release benefit from pitch cues due to better TFS processing. Similar to the “space” and “pitch” conditions, there was no significant group difference when two cues were combined, as shown in the “pitch+space” condition (Figure 3.4a). The result, however, suggests an additive nature of the benefits from pitch and space cues, leading to a masking release of around 10 dB. Indeed, it was shown that the differences in pitch can help listeners more easily make sense of competing sound sources and better segregate when spatial cues are available (Darwin, 2005). Having speakers with distinguishable voices coming from different spatial locations is perhaps also a more ecologically relevant listening scenarios we encounter in everyday listening.

When the background was switched from a speech-babble masker to a non-speech masker, shown as a “steady noise” condition in Figure 3.4a, we observed a masking release of almost 20 dB, suggesting a significant impact from informational masking on daily speech-in-noise communication. This is consistent with the result from Arbogast et al., 2002, where the informational masking was systematically studied. Informational masking refers to the listening challenges brought by the similarity or confusion between the target and background when there is no significant spectrotemporal overlap between them. The masking due to spectrotemporal overlap is called energetic masking. In the “steady noise” condition (Figure 3.4a), a steady noise masker was created to make sure that it is purely energetic masking with the continuous spectrum and minimum intrinsic fluctuations or modulations to eliminate informational masking (based on Stone and Moore, 2014, see Methods). In the control condition, on the other hand, the 4-talker speech masker introduces informational masking because of the speech-on-speech masking configuration, in addition to the energetic masking. While the energetic masking was not controlled for in the reference condition, the improvement of almost 20 dB in the speech reception thresholds from the control condition to the “steady noise” condition implies a significant role of informational masking in everyday listening. While researchers generally believe that energetic masking occurs at the auditory periphery (Culling & Stone, 2017), informational masking is believed to involve central factors such as object formation, auditory selective attention, perceptual scene segregation, auditory working memory, and linguistic processing (G. Kidd & Colburn, 2017). Studies have suggested that TFS-based cues, such as pitch (Smith et al., 2002), play an important role in supporting the formation of the aforementioned high-level functions that are critical for the release from informational masking (Darwin, 1997; Oxenham & Simonson, 2009; B. G. Shinn-Cunningham, 2008; Viswanathan, Shinn-Cunningham, et al., 2021). Furthermore, TFS cues were shown to influence the coding of the target speech envelope in the brain, which predicts the intelligibility performance in various distortions, including speech-on-speech masking, where informational masking is dominant (Viswanathan, Bharadwaj, et al., 2021). Despite these indirect implications of the relationship between TFS coding and the release from informational masking, our results show that there was no significant difference in the release from informational masking when the subjects were grouped based on their TFS sensitivity (see Figure 3.4a). However, the group with better TFS sensitivity had a significantly shorter response time than their poor counterpart, as shown in Figure 3.5b. The response time was measured as a quantification of the subject's overall central engagement, with shorter response time or faster reaction during the task indicating superior central functions. Therefore, our results support previous findings regarding the role of TFS coding in central processing, which is crucial for the release from informational masking. The lack of group difference in the informational masking release is probably because the group difference in central processing quantified by the reaction time did not translate into the masking-release metric. It highlights the importance of investigating the metrics that are beyond the final threshold measurements, which may reveal many aspects that were hidden below the surface.

We also investigated the relationship between TFS processing and speech-in-noise hearing in the presence of reverberation. The same four conditions listed in Figure 3.4a were examined with the addition of reverberation, including "pitch", "space", "pitch+space", and "steady noise" conditions, along with the reference condition. As shown in Figure 3.5a, reverberation notably increased (i.e., worsened) the speech reception thresholds, indicated by the difference in thresholds from conditions with and without reverberation. More importantly, the group with poor TFS sensitivity had significantly more threshold increase than their counterpart with better TFS sensitivity ($p < 0.0001$), indicating a possible role of TFS processing in "de-reverberation". Indeed, TFS-based spatial cues were shown to be more susceptible to the corruptions from reverberation (Ihlefeld & Shinn-Cunningham, 2011) and reverberation also tends to cripple spatial selective attention (Ruggles et al., 2011), where TFS processing was believed to be involved as mentioned earlier. Our result combined with these reports suggests that a better TFS processing might serve as a protective shield against the negative impact of reverberation on speech-in-noise hearing. Besides the aforementioned role of TFS processing in central processing and hence reduced listening efforts, it is another potential benefit of good TFS processing that has been revealed in this study. One thing to note is that the threshold increase due to reverberation was calculated from the data merged across all test conditions with and without reverberation (see Methods). In contrast, the masking release metric within individual reverberant conditions did not differ between groups, as shown in Figure 3.4b. Interestingly, the spatial release from masking turned out to be negative in reverberation (Figure 3.4b), indicating an impairment from the spatial cues instead of benefits. The impairment could be due to the use of interaural phase cues (see Methods), which is not as ecologically realistic as precisely implemented ITD cues. The mix of these artificial spatial cues and ecologically relevant acoustic distortions such as reverberation might have caused confusion during the task. It also further reflects the unreliable nature of spatial cues in reverberation as mentioned above (Ihlefeld & Shinn-Cunningham, 2011; Palomäki et al., 2004; Ruggles et al., 2011). In contrast to the distortion from reverberation on spatial cues, their effects on other cues such as amplitude modulation or harmonic structure were shown to be less profound (Culling et al., 1994; Ruggles et al., 2011; B. Shinn-Cunningham et al., 2017). In line with this, our result shows that the pitch-based masking release was not affected at all by the reverberation, as shown in Figure 3.4b. The additive nature of the pitch and spatial cues observed in the non-reverberant "pitch+space" condition (Figure 3.4a) is also present in reverberant settings (Figure 3.4b), where the impairment from spatial cues was neutralized to some extent by the pitch cues. There was still a significant amount of release from informational masking despite the small reduction, as compared to the non-reverberant condition. This is similar to the findings from Deroche et al., 2017, where reverberation was shown to reduce the release from informational masking. Nevertheless, no significant group difference due to TFS processing was observed in the masking release measure across these conditions in reverberation, which could be due to the same factors that were discussed for the results from the non-reverberant conditions.

To our knowledge, this study constitutes the most comprehensive investigation of the perceptual role of TFS by covering all major aspects of speech-in-noise hearing where TFS processing has been believed to play a role, including pitch perception, spatial hearing, informational masking, and listening in reverberation. Despite the extensive prior literature on TFS, the contribution of TFS to speech perception in noise is poorly understood. It is mainly due to the limitation of the vocoding approach that has been traditionally used to study TFS. We avoided this limitation by taking the individual-difference approach, systematically measuring TFS sensitivity and speech-in-noise intelligibility from a very large subject population (n=200). Our results not only confirmed many findings from previous literature but also, most importantly, demonstrated that TFS is indeed crucial for speech-in-noise hearing, by reducing the impacts from reverberation and reducing reaction time indicating lessened listening efforts. This is an encouraging message for CI companies, for introducing TFS information into the sound coding strategies to ameliorate the listening difficulties in noise. These findings as well as our success with the individual difference approach also show the promise of adopting TFS and speech-in-noise measurements in audiology clinics. First, these results indicate that it is worth measuring TFS and speech-in-noise sensitivity since they could reflect suprathreshold hearing deficits in more complex yet everyday listening environments. Establishing assessment tools for TFS and speech-in-noise sensitivity might help diagnose certain clinical populations, such as the “hidden hearing loss” patients who have hearing difficulties in noise, but with a “normal” clinical diagnosis based on the current diagnostic tools that are only designed to assess audibility. Second, our large-scale individual-level measurement battery for TFS and speech-in-noise sensitivity provides a blueprint for implementing those tests in clinics. The targeted sensitivity of a test battery to individual differences that are free of influences from extraneous factors is essential for the test battery developed in research to be adopted in clinical applications. It is the sensitivity to individual differences of a test battery that help the clinicians make better-informed treatment decisions for individual patients.

Lastly, our study is also one of the first few online psychoacoustics studies and our online data-collection platform could serve as a template for future virtual testing in both research and clinics. All data for this study battery, including TFS-sensitivity measurements, ILD reference measurements, and speech-in-noise measurements, were collected on the remote-testing platform to cope with COVID-related restrictions on in-person measurements. Online data are comparable to those from in-person measurements from our lab as well as from many other studies in the literature (see Figure 2.8 and Figure 2.9). This shows the promise of using virtual measurements in future studies. Our remote testing not only produced data of matching quality but also demonstrated tremendous advantages over traditional in-person testing. Conventional psychoacoustic experiments are time-consuming and hence usually involve a small number of subjects that might have been recycled many times. A larger number of participants with more diverse ethnic and cultural backgrounds, hearing history, musical training, and daily experience with sounds is key to exploiting individual differences for a better understanding of the fundamental aspects of auditory processing. We were able to administer a large battery of TFS sensitivity assessments and comprehensive speech-in-noise measurements under 10 different target-masker configurations on a 200-subject population. Because of the synchronous and highly-automated nature of our online study, we finished the entire data collection within just a few days. In contrast, data collection for such a large study battery from 200 participants could have taken months if not years through traditional in-person measurements. The online study further saves the time and cost that comes with a laboratory visit, especially for scheduling "rare" subjects that are not local. For instance, it is not uncommon for research labs studying CI to fly in the CI users from other states for week-long experiments. The costs of travel reimbursement and compensation for time could be very high and this type of visit can also be extremely challenging to schedule because not everybody could make time for a week-long visit. Online platforms could greatly ameliorate this issue by allowing researchers to access more diverse and representative subject pools that may otherwise not be able to easily visit research facilities. Last but not the least, online platforms could help continue human-subject research under special circumstances such as COVID-19 when in-person experiments are restricted. The online format has also been more and more widely adopted by medical facilities, for the same aforementioned benefits: providing medical services for those in areas with a lack of medical resources as well as continuing to provide medical care to patients amidst a global pandemic. With the success of online study, we foresee wider adoption of virtual testing in both research and clinic in the near future.

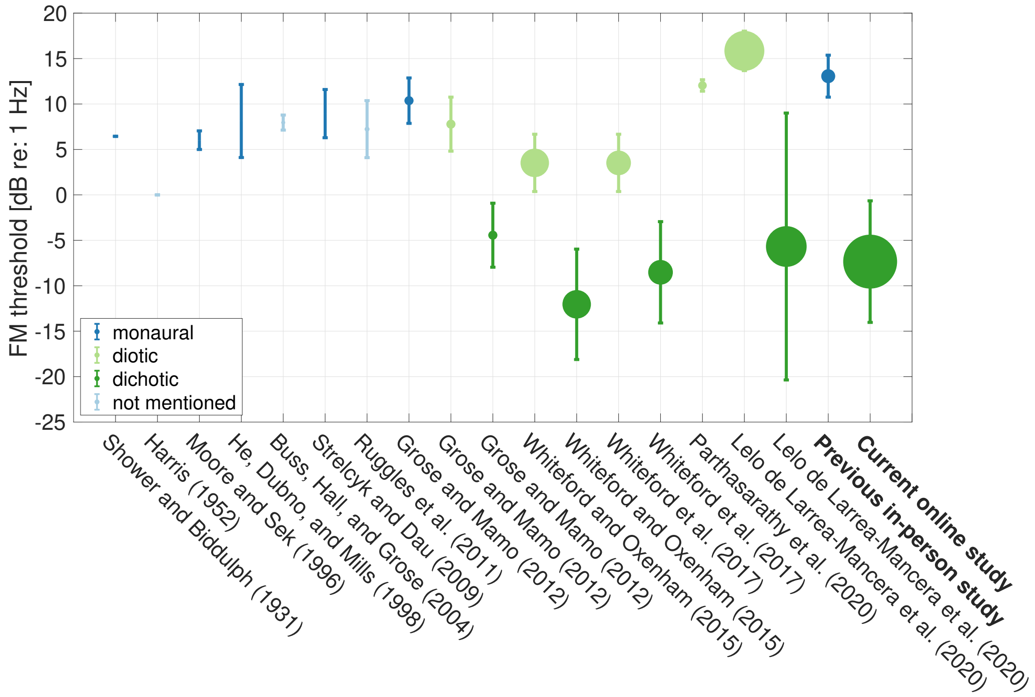


Figure 3.6. A sample of published reports of FM detection thresholds for comparison (Borjigin et al., 2021; Buss et al., 2004; Grose & Mamo, 2012; Harris, 1952; He et al., 1998; Lelo de Larrea-Mancera et al., 2020; Moore & Sek, 1996; Parthasarathy et al., 2020; Ruggles et al., 2011; Shower & Biddulph, 1931; Strelcyk & Dau, 2009; Whiteford et al., 2017; Whiteford & Oxenham, 2015). Borjigin et al., 2021 is our previous in-person study. Error bar is 1 standard deviation (std). The size of the dot represents the number of subjects (current study has the most subjects; N=200). Stimulus parameters such as stimulus level, carrier frequency, and modulation frequency in the cited studies are similar to those used in the current study, with slight differences (e.g., Ruggles et al., 2011; Strelcyk and Dau, 2009 used carrier at 750 Hz). Some threshold values are approximate from figures (e.g., mean and std had to be estimated based on median and range in the box whisker plots from Whiteford and Oxenham, 2015 and Whiteford et al., 2017). The mean and std from the young and middle-aged group from Grose and Mamo, 2012 were combined to generate a single data point. Some authors expressed the threshold in terms of $\Delta F/F_c$, where ΔF is frequency deviation, and F_c is the carrier frequency. Moore and Sek, 1996 used ΔF that was in two directions, i.e., peak-peak. Subjects from some studies were highly experienced in psychoacoustic tasks hence the thresholds were very low/good. Whiteford et al., 2017; Whiteford and Oxenham, 2015 obtained thresholds that fall in the lower end of the results from a very large number of subjects. This may be because their subjects were younger typical-hearing listeners and the stimuli were presented diotically and dichotically instead of monaurally.

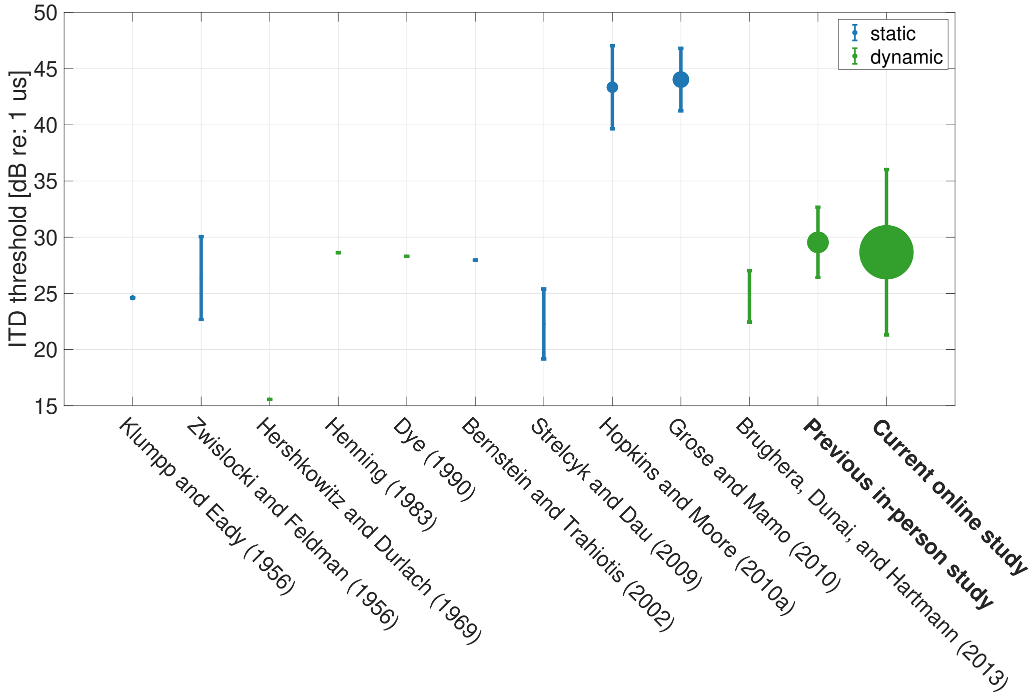


Figure 3.7. A sample of published reports of ITD detection thresholds for comparison (L. R. Bernstein & Trahiotis, 2002; Borjigin et al., 2021; Brughera et al., 2013; Dye, 1990; Grose & Mamo, 2010; Henning, 1983; Hershkowitz & Durlach, 1969; Hopkins & Moore, 2010; Klumpp & Eady, 1956; Strelcyk & Dau, 2009; Zwicker, 1956). Borjigin et al., 2021 is our previous in-person study. Error bar is 1 standard deviation (std). The size of the dot represents the number of subjects (Current study has the most subjects; N=200). Stimulus parameters such as level and carrier frequency in the cited studies are similar to those used in the current study, with slight differences (e.g., Strelcyk and Dau, 2009 used carrier at 750 Hz). Note that some threshold values were extracted approximately from figures rather than direct numerical reports. Some of the studies used stimuli with the leading ear switching from one side to the other (labeled "dynamic", marked in green color), whereas others presented an ITD only in the target intervals, with the reference being the midline (labeled "static", marked in blue color). Note that the values from Brughera et al., 2013; Hershkowitz and Durlach, 1969 were halved since the authors used $ITD/2$ in each interval. The mean and std from young and middle-aged cohort from Grose and Mamo, 2010 were combined to generate a single data point. Subjects from some studies were highly experienced in psychoacoustic tasks, hence very low/good thresholds.

4. DEEP NEURAL NETWORK ALGORITHMS FOR NOISE REDUCTION AND THEIR APPLICATION TO COCHLEAR IMPLANTS

4.1 Introduction

Cochlear implant (CI) listeners struggle to understand speech in noisy environments, despite often achieving satisfactory speech intelligibility in quiet settings. This is especially true when the background noise is modulated, non-stationary noise interference (Cullington & Zeng, 2008; Fu et al., 1998). Some front-end processing algorithms have been shown to improve speech intelligibility in fluctuating noise. For example, studies demonstrated that two-microphone directionality could introduce a masking release of up to 10 dB (Hersbach et al., 2012; Wouters & Vanden Berghe, 2001). However, these approaches require the listener to always face the target, which helps create spatial separation between the target and background. In this work, we implemented and evaluated two single-microphone noise reduction algorithms that do not rely on spatial separation and do not require the use of two microphones. Traditionally, single-channel noise reduction algorithms that have been commonly used for hearing assistive devices, are driven by signal processing strategies based on signal statistics. Classic examples include spectral subtraction (Boll, 1979) and wiener filtering (Scalart & Filho, 1996). These models could improve speech intelligibility to a certain extent in statistically predictable backgrounds, such as stationary noise (Dawson et al., 2011; Loizou et al., 2005; Mauger et al., 2012). For speech in a more complicated, non-stationary, multi-talker babble background, machine-learning techniques, such as deep neural networks (DNNs) or Gaussian mixture models (GMMs), were shown to be successful in improving speech intelligibility for listeners with typical hearing (G. Kim et al., 2009), for listeners with hearing loss (Bramsløw et al., 2018; Chen et al., 2016; Healy et al., 2019; Healy et al., 2015; Healy et al., 2013; Monaghan et al., 2017), and for CI listeners (Goehring et al., 2017; Hu & Loizou, 2010; Lai et al., 2018). Recent improvements over these models were introduced by DNN-based regression models, where the training target was a continuous or soft mask, instead of a binary mask (Bentsen et al., 2018; Madhu et al., 2013). However, all models mentioned above function based on *a priori* knowledge of the target and/or background by using the same target speaker (Chen et al., 2016; Lai et al., 2018), background interference (Goehring et al., 2017), or both (Bentsen et al., 2018; Bramsløw et al., 2018; Goehring et al., 2017; Healy et al., 2019; Healy et al., 2015; Healy et al., 2013; Hu & Loizou, 2010; G. Kim et al., 2009; Lai et al., 2018) for the training and testing process.

The generalization of a speech enhancement model is critical to ensure efficacy in CI devices. It is, however, impractical to present all speech-in-noise mixtures that a user would normally encounter in real-world listening situations to the models during the training process. Indeed, in most studies, the model performance evaluated by objective intelligibility metrics was reduced significantly when unseen testing data were presented to the DNN models (Chen & Wang, 2017; Goehring et al., 2017; May & Dau, 2014). Recent studies have demonstrated the promise of using recurrent neural network (RNN) models for better generalization by including recurrent connections, feedback, and gate elements (Chen & Wang, 2017; Graves et al., 2013; Kolbæk et al., 2017; Weninger et al., 2015). Classic architecture with these elements is the long short-term memory (LSTM) RNN structure that accumulates information from the past and hence enables the network to form a temporary memory (Hochmair et al., 2015; LeCun et al., 2015), which is essential for properly managing and learning speech context. RNN-LSTM based models have been shown to improve the speech-in-noise perception for listeners with hearing loss (Bramsløw et al., 2018; Healy et al., 2019; Keshavarzi et al., 2019; Keshavarzi et al., 2018) and CI users (Goehring et al., 2019).

Despite the wide adoption of RNN models in modern audio processing systems in many domains, including speech recognition and synthesis as well as speech enhancement and segregation, the sequential nature of the RNN architecture often renders the computation inefficient. The impact of this limitation becomes especially prominent with long speech sentences. Practically, this bottleneck can be avoided by using a mechanism known in the literature as the transformer. The transformer is a fully attention-based mechanism that can effectively replace the recurrence structure in a common RNN model (Vaswani et al., 2017). This architecture allows the model to attend to the entire sequence all at once and establish connections between distinct elements, which ultimately leads to more efficient learning of long-term dependency. The transformer has gained considerable popularity and competitive performance in speech recognition (Karita et al., 2019), speech synthesis (Li et al., 2019), speech enhancement (J. Kim et al., 2020), and audio source separation (Subakan et al., 2021). The SepFormer model developed by Subakan et al., 2021 is currently the top-performing model in speech separation applications, according to Papers with Code website.

In this work, we implemented the SepFormer model for speech processing in noise. Although being state of the art, this model might not be suitable for real-time applications such as CI due to its complicated architecture. Therefore, while using SepFormer as a reference for the flagship benchmark model, we also implemented a low-complexity RNN model of speech enhancement, to account for the constraints on processing time and computational power in real-time applications. The effectiveness of these two proposed DNN models was verified through commonly used objective intelligibility metrics. We then investigated the clinical effectiveness of this approach for CI recipients under noisy conditions by thorough behavioral testing using challenging noise types and signal to noise ratio (SNR) levels. The goal of this work is to serve as a proof of concept that aims to facilitate the adoption of DNN technology in CI devices for better speech hearing in more complex and dynamic background noise.

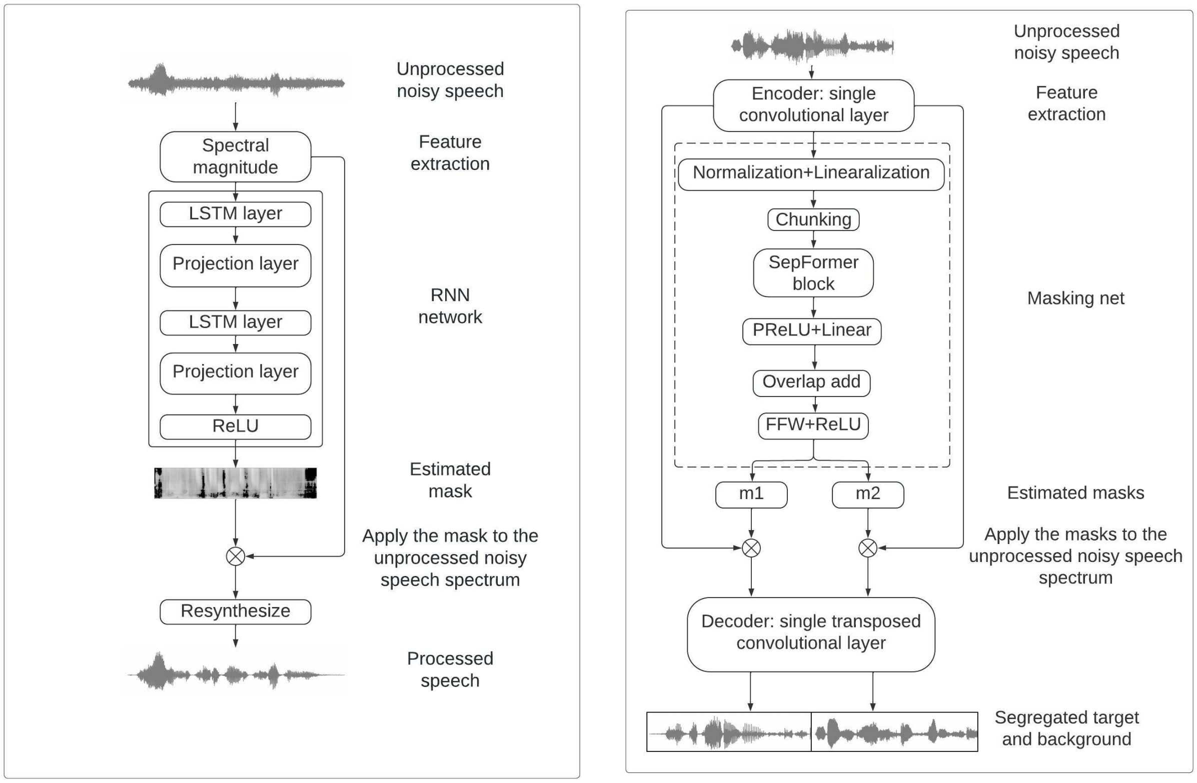
4.2 Network Architecture

4.2.1 RNN

The schematic of the single-channel, RNN-based speech enhancement algorithm is illustrated in Figure 4.1a. A clean target speech and either babble-speech or non-speech noise were mixed to create the unprocessed noisy speech. The features were spectral magnitude from short-time Fourier transformation (STFT). The features were extracted using Hamming-windowed frames with a window size of 32 and a hop size of 16. The "add-one" log was applied to the spectral magnitude to reduce the influence of very small values. The predicted mask (i.e., the model outcome) was a continuous "soft" mask instead of an ideal binary mask, where the values of the mask were either one or zero. It was shown that "soft" masks introduced better speech quality and intelligibility (Madhu et al., 2013). In addition, no threshold had to be determined for the "soft" mask as for the ideal binary mask. The predicted mask from the RNN model was multiplied with the original unprocessed noisy mixture to generate the “de-noised” spectrum of the mixture with an enhancement on the target speech. This “de-noised” spectrum was compared with the spectrum of the clean target speech to compute the mean square error (MSE) loss for optimizing the training outcome. The final processed speech was recovered by resynthesizing (i.e., taking the inverse STFT) the “de-noised” spectrum. The RNN network consisted of two long short-term memory (LSTM) layers, with each followed by a projection layer. A PyTorch-powered speech toolkit—SpeechBrain—was used to implement, train, and test the RNN model. Adam optimizer was used for minimizing the MSE loss during the training process (Kingma & Ba, 2017), with the learning rate set at 0.0001. The model performance was evaluated and monitored with a validation dataset at the end of each full learning cycle with all training samples (i.e., epoch). The training was terminated after 100 epochs to avoid overfitting, where the model performance with the validation dataset stabilized with no further significant improvements.

4.2.2 SepFormer

While the simple, light-weight RNN model evaluated here serves as a proof of concept for DNN algorithms that are suitable for small devices such as CI, we also implemented the current state-of-the-art model for speech separation applications—SepFormer, to explore the limits of current DNN technology in the speech enhancement and source separation domains. The model architecture is depicted in Figure 4.1b. A single-layer convolutional network was used as an encoder to learn the STFT-like representation of the input noisy signal. Similarly, at the end of the process, a transposed convolution layer with the same stride and kernel size as in the encoder was used to turn the STFT-like representations back into separate sources within the mixture. The extracted STFT-like features of the noisy mixture go into the masking network, which estimates the masks for the foreground (i.e., target speech) and background. These masks are also “soft”, continuous masks as in the RNN model. In the masking net, the features are first normalized and processed by a linear layer. They are then chopped into chunks along the time axis with an overlap factor of 50%, which are next fed into the core of the masking net—SepFormer block. This block consists of two transformer structures that can learn both short and long-term dependencies. More details can be found in Subakan et al., 2021. The output of the SepFormer block is then processed by a PReLU and linear layer. The overlap-add scheme, described in Luo and Mesgarani, 2019, was used to sum up the chunks. This summed representation is passed into two feed-forward layers and a ReLU activation to finally generate the masks for both the foreground and background sources. The training procedure and infrastructure are the same as in the RNN model.



(a) RNN

(b) SepFormer

Figure 4.1. Schematic diagrams of the DNN architecture and signal processing frameworks used in this study: (a) RNN and (b) SepFormer

4.3 Materials

4.3.1 Training and testing datasets

The speech materials for training (including validation) the models were taken from LibriSpeech, an open-source corpus of about 1000 hours of read English sentences (sub-datasets: 100, 360, 500-hour dataset). The data were extracted from read audiobooks, with each sentence being carefully segmented and aligned. The sentences were recorded by a total of 2484 speakers. In this work, we chose the 100-hour sub-dataset for model training. Non-speech materials were from WHAM!, an open-source dataset of environmental recordings from various urban locations throughout the San Francisco Bay Area in late 2018, such as coffee shops, bars, and parks. Each recording was carefully processed to remove any intelligible speech. The speech-in-noise dataset for training was generated by mixing a speech sentence with either speech babble or non-speech environmental noise. A total of four conditions were created: a target speech mixed with non-speech noise, a target speech mixed with 1, 2, and 4-talker speech babbles. Each condition was created at 1-10 dB signal-to-noise ratios (SNR) in 1-dB steps. The loudness of the target was kept constant throughout the dataset. It was made sure that each babble-talker utterance contained distinct speakers and content. There were 5590 and 410 utterances for each condition in the training and validation dataset, respectively, yielding a total of ~30 hours of training data.

The models were tested and evaluated at 1, 5, and 10 dB SNR, in 2-talker speech babble and non-speech backgrounds. To test the generalization of the models, both speech and non-speech materials were extracted from different sources than those used for constructing the training dataset. The speech materials were from a subset of the IEEE "Harvard" corpus (Rothauser, 1969), which contains 720 sentences, recorded by 33 speakers (15 female speakers). Three speakers were chosen for testing: one male speaker for the target speech and two female speakers for the 2-talker mixtures. Within all sentences, 340 of them were used as target sentences, while the rest were used to generate 2-talker mixtures. Each 2-talker mixture consists of two different sentences that are spoken by different speakers. The non-speech background noise was a steady-state noise, which is much denser in both time and frequency than the environmental sounds from the training dataset. We used this test dataset for testing with both objective evaluation metrics and listeners with CI.

4.3.2 Objective intelligibility models

The DNN models were first evaluated quantitatively using three popular objective evaluation metrics: source to distortion ratio (SDR) (Vincent et al., 2006), short-time objective intelligibility (STOI) (Taal et al., 2010, 2011), and perceptual evaluation of speech quality (PESQ) (Hu & Loizou, 2008). These accurate and reliable objective evaluation methods provided useful information regarding the overall expected benefit before conducting behavioral listening tests with CI users. All three evaluation metrics compare a clean reference speech and the processed speech in an attempt to measure the overall benefit elicited due to processing and score the overall quality. The SDR metric decomposes the estimated source into four components representing respectively the true source, spatial distortions, interference, and artifacts. The final SDR score is computed by calculating the ratio of the source energy to the sum of all other projection energies (i.e., spatial distortions, interference, and artifacts) as described in Vincent et al., 2006. The STOI metric was initially designed to predict the intelligibility of speech processed by enhancement algorithms. Recently, Falk et al., 2015 demonstrated that for CI users, the STOI outperformed all other measures for predicting the intelligibility of enhanced speech in noise. The STOI first applies time-frequency analysis to both clean reference and processed speech. An intermediate intelligibility measure is obtained by estimating the linear correlation coefficient between clean and processed time-frequency units. The final STOI score is the average of all intermediate intelligibility estimates from all time-frequency units. The PESQ score ranges between –0.5 and 4.5 and employs a sensory model to compare the reference signal with the processed signal by relying on a perceptual model of the human auditory system. The PESQ is computed as a linear combination of average disturbance value and average asymmetric disturbance value. The parameters for the linear combination can be further modified towards predicting different aspects of speech quality. More details can be found in Hu and Loizou, 2008 and Kokkinakis and Loizou, 2011. In general, the PESQ has been shown to reliably predict the quality of

Table 4.1. Demographic information for the subjects who participated in this study. [1] Fine structure processing with *sequential* stimulation in the four apical channels. [2] Fine structure processing with *parallel* stimulation in the four apical channels.

Subject ID	**Gender**	**Age (yrs)**	**Duration of CI use (yrs)**	**Active electrodes**	**Clinical strategy**
ME149	F	57	6	8	FS4[1]
ME153	M	68	4	12	FS4
ME185	F	65	9	12	FS4-p[2]
ME196	M	66	2	11	FS4-p
ME202	M	56	3	12	FS4
ME203	F	63	4	11	FS4-p
ME204	M	63	2	12	FS4-p
SSD100	F	35	10	12	FS4-p

processed speech and hence in the present context, the PESQ is assumed to be capable of detecting and quantifying the overall effects of DNN processing on the signal quality.

4.4 Methods

4.4.1 Subjects

A total of eight post-lingually deafened adults fitted with MED-EL CIs (MED-EL GmbH, Innsbruck, Austria) took part in the study. They were all between the ages of 35 and 68 (4 male and 4 female). Their mean age at testing was 59.1 years (SD = 10.4 years). The average duration of CI use in years was 5 years (SD = 3 years). Demographic information is provided in Table 4.1, including each subject's default clinical sound coding strategy. The research study was approved by the Western Institutional Review Board. All subjects gave informed consent prior to testing. Only research subjects whose participation in the study would cause them to incur financial hardship received financial compensation for their participation.

4.4.2 Test setup

All participants were tested using their everyday program, including all front-end processing features that they normally used. Before the testing session, their audio processor was loaded with the recipient's daily program. The participants used the direct audio input (DAI) cable, which attenuates the microphone inputs by approximately 30 dB while passing the direct input signal through DAI cable without attenuation. At the beginning of the testing session, the audiologist provided instructions regarding the study procedures, and then connected the recipient's processor to the audio port of a Windows-based Microsoft Surface Pro touchscreen tablet through DAI cable. The proprietary psychophysical software suite PsyWorks v.6.1 (MED-EL GmbH, Innsbruck, Austria) was used to present the speech stimuli from the table to the audio processor. The calibration was performed using a built-in feature within the PsyWorks software and was designed specifically for use with each recipient's audio processor.

4.4.3 Procedure

The testing was carried out in a self-administered manner. The subjects used PsyWorks to present the speech-perception materials to their audio processor. Subjects were assigned a unique presentation order using a Latin square design and were blinded to the processing condition that was presented. The subjects vocalized responses through a microphone located in front of them. The responses were scored in real-time using an automatic speech-to-text recognizer module which captured all words that were correctly or incorrectly identified. Words containing additions, substitutions, or omissions were scored as incorrect. The percent correct scores for each condition were calculated by dividing the number of words correctly identified by the total number of words in a sentence list. After each list, the percent correct was displayed and stored electronically. All participants were native English speakers, and none had speech difficulties that prevented the investigator from understanding their responses. The total testing time for all conditions including multiple breaks was approximately 2.5 hours.

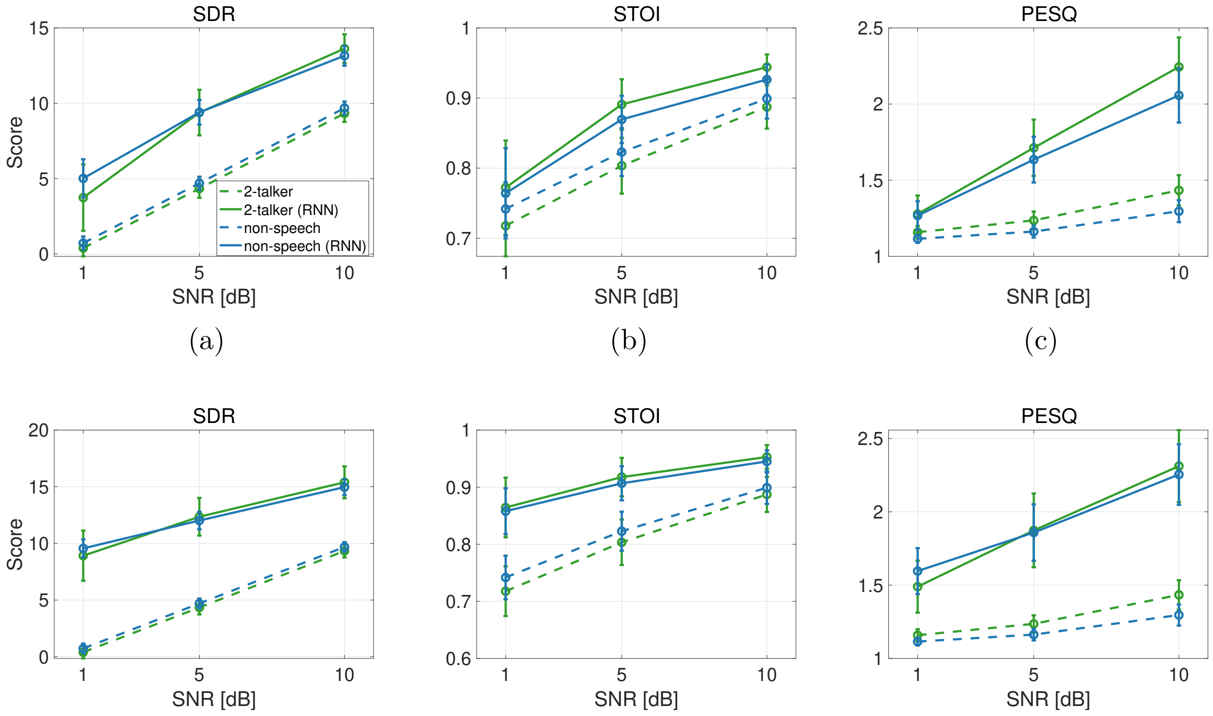
4.5 Results

4.5.1 Objective assessment

The models introduced significant improvements across all objective evaluation metric scores. The three objective evaluation scores of the RNN model that were tested with 340 samples are shown in Figures 4.2a-4.2c. The RNN processing improved the source-to-distortion ratio (SDR, Figure 4.2a) scores over the unprocessed condition (dashed lines) across all signal-to-noise ratios (SNRs) as well as across both masker types (green: 2-talker, blue: non-speech noise). The improvements in SDR introduced by the RNN model (i.e., the elevation from the dashed lines to solid lines) are all statistically significant (p values less than 0.0001). The improvements in speech quality and intelligibility introduced by RNN can also be evidenced by the other two metrics: perceptual evaluation of speech quality (PESQ, Figure 4.2b) and short-time objective intelligibility (STOI, Figure 4.2c). All improvements are also statistically significant, with a p value that is less than 0.0001. Although statistically significant, the improvement in speech intelligibility metric (i.e., STOI) was not as prominent as in the two speech quality metrics (i.e., SDR and PESQ). It is probably because the SNR tested was high overall (starting from 1 dB SNR) and speech intelligibility was not a significant issue (Tang et al., 2017). The objective evaluation scores for the unprocessed noisy mixtures (dashed lines) remained the same for the SepFormer model since the test materials did not change, as shown in Figure 4.2d-4.2f. However, the scores for the processed audio signals by the SepFormer model (solid lines) had even more separation from the dashed lines, indicating better performance by SepFormer than RNN. The superior performance of SepFormer over RNN is especially evident in the 1-dB-SNR, or the worst, conditions across all three evaluation metrics.

4.5.2 Behavioral testing with CI listeners

The models also introduced significant improvements in speech intelligibility scores in all CI listeners tested. The speech intelligibility scores observed for the three processing conditions: unprocessed, processed by RNN, and processed by the SepFormer model are



(a) (b) (c)

(d) (e) (f)

Figure 4.2. (a)-(c): Objective evaluation scores for signals processed with the RNN architecture; (d)-(f): objective evaluation scores for signals processed with the SepFormer architecture.

plotted in percent correct in Figure 4.3. As with the objective evaluation metrics, CI listeners were tested with both 2-talker and non-speech masker conditions. Figure 4.3a depicts the data collected in the 5 dB SNR condition, while Figure 4.3b shows the data obtained in the 10 dB SNR condition.

The percent correct scores were normally distributed according to the Shapiro-Wilk test for normality and hence we conducted the following statistical tests. First, a three-way analysis of variance ANOVA (with repeated measures) using the processing condition, SNR, and type of masker as within-subject factors indicated that there were no statistically significant three-way or two-way interactions. Hence, two separate sets of two-way ANOVA (with repeated measures) were conducted separately at each SNR level in order to examine the main effects of the processing condition and masker type separately at each SNR level.

In the 5 dB SNR condition, a repeated measures two-way ANOVA indicated a statistically significant effect of the masker type ($F[1,7] = 9.85, p = 0.016$), and a significant effect of the processing condition ($F[2,14] = 32.95, p < 0.001$). For SNR = 10 dB, a repeated measures two-way ANOVA indicated only a statistically significant effect of the processing condition ($F[2,14] = 26.99, p < 0.001$), confirming that speech intelligibility did not differ significantly between different types of maskers. Post-hoc comparisons (with Bonferroni corrections) between scores revealed that the RNN and SepFormer conditions were not significantly different from one another ($p = 0.183$). The RNN scores were also not significantly different from the unprocessed condition ($p = 0.0815$). However, speech intelligibility scores due to processing with the SepFormer model were significantly different from the unprocessed condition ($p = 0.00213$).

In the 5 dB SNR unprocessed condition, the median score was 27% in non-speech noise and 15% in speech noise. After processing with the RNN model, the median performance increased to 49% in non-speech noise and 46% in speech noise, respectively. Similar to what was observed in the objective evaluation metrics analysis, the SepFormer model demonstrated even better performance, raising the median score to 58% in non-speech and 60% in speech noise. In the 10 dB SNR condition plotted in Figure 4.3b, there was an improvement in the median scores in the unprocessed condition as compared to the 5 dB conditions: 39% in non-speech noise and 41% in speech noise. The RNN model further improved the scores to 63% and 64% in non-speech and speech noise, respectively. The SepFormer model outperformed the RNN, achieving 70% and 73% in non-speech and speech noise, respectively. Theses improvements are not only present in the medians, but also on the individual level. For example, in 5 dB, non-speech noise condition, all subjects but SSD100 and ME196 had improvements in their percent score due to RNN processing; all subjects had higher scores due to SepFormer processing. Considering the large individual variability that is typical of CI patients, the consistent improvement across almost all subjects is a very encouraging message for the application of DNN-based models for noise reduction in CIs.

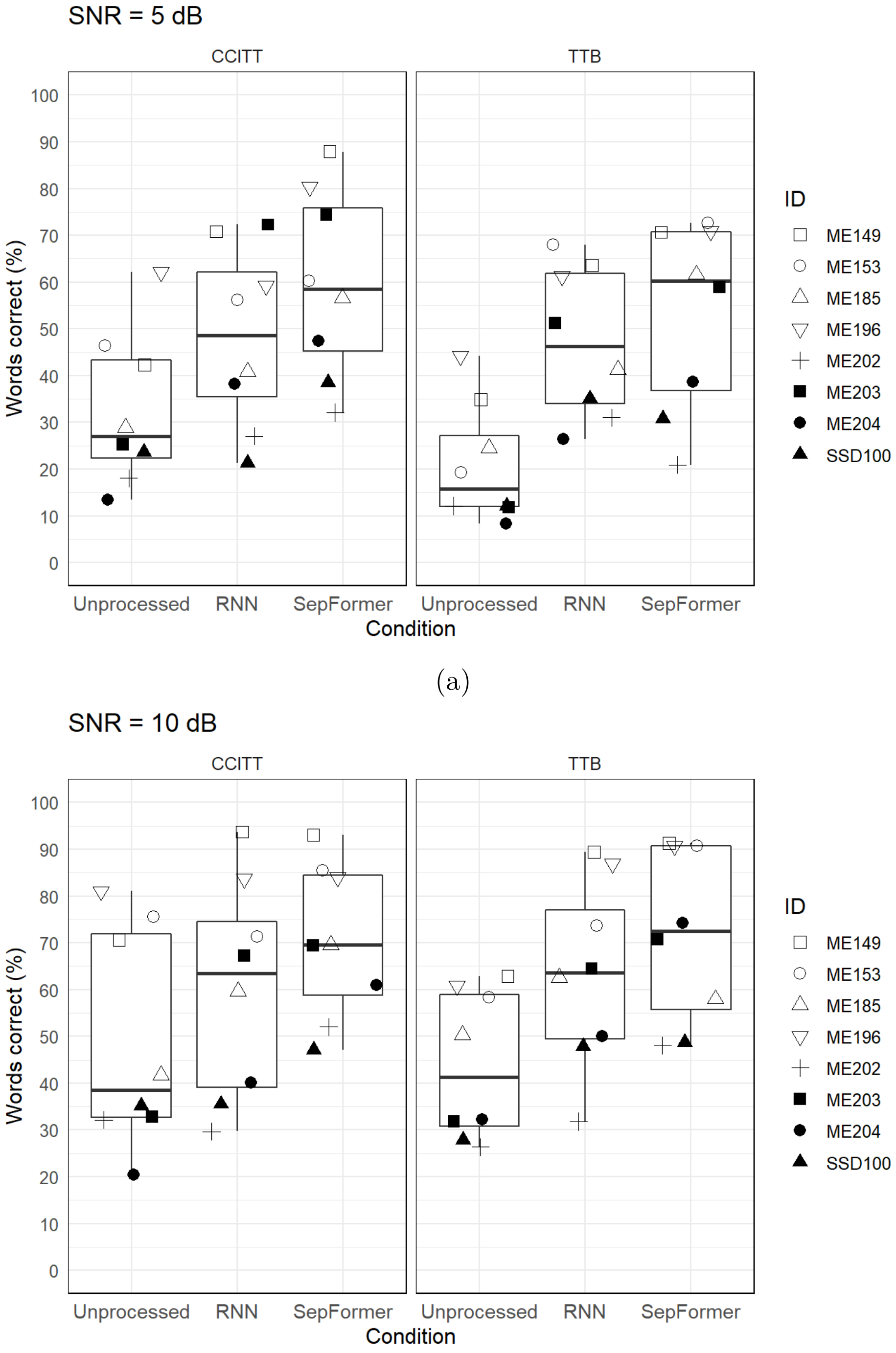
Interestingly, the increase introduced by the models was greater in speech noise than non-speech noise. For RNN model in 5 dB condition, there was a 22% increase in non-speech noise, whereas the increase was 31% in speech noise. An even larger difference was observed with SepFormer: 31% increase in non-speech noise and 45% in speech noise. This is in contrary to traditional signal processing strategies that are designed for removing statistically predictable and relatively stationary noises. These algorithms fall short in more complicated non-stationary speech backgrounds (Boll, 1979; Dawson et al., 2011; Loizou et al., 2005; Mauger et al., 2012; Scalart & Filho, 1996). The better performance in speech noise demonstrated by both RNN and SepFormer is in agreement with the results from many previous studies on machine-earning based noise-reduction models, such as DNNs and Gaussian mixture models (GMMs) (Bramsløw et al., 2018; Chen et al., 2016; Goehring et al., 2017; Healy et al., 2019; Healy et al., 2015; Healy et al., 2013; Hu & Loizou, 2010; G. Kim et al., 2009; Lai et al., 2018; Monaghan et al., 2017). This shows the promise of using machine-learning based models as a complimentary algorithm to current existing signal processing strategies, for noise reduction in more complicated, non-stationary background.

4.6 Discussion

The results show that the speech quality and intelligibility were improved by both RNN and SepFormer models. There were significant increase across all objective evaluation metrics, including source to distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). The increase in the evaluation scores from unprocessed to processed noisy mixtures was significant across all signal-to-noise ratio (SNR) test conditions, in both masker types, in all three metrics, and for both models. The significance was visually apparent for the two quality metrics—SDR and PESQ, but not so much for the intelligibility metric—STOI. It is perhaps because the room for improvement in intelligibility is small in relatively high-SNR conditions we tested, with the worst condition being 1 dB SNR. In general, the SepFormer model, as the current state-of-the-art, performed better than the simple template RNN model. This advantage is especially prominent in lower-SNR conditions. For instance, in the 1 dB SNR condition, SepFormer introduced an improvement of 8.52 in SDR score as compared to 3.35 from RNN. Even in STOI scores, where the benefits from both models were not as great as in the other two quality metrics, SepFormer still outperformed RNN.

The benefits observed from both DNN models with CI users, mirrored the improvements in speech quality and intelligibility shown in the objective evaluation metrics. For example, in the 5 dB condition, the CI listeners' median intelligibility score averaged across speech and non-speech conditions (not shown) increased from 24% to 46% due to RNN processing and further increased to 60% in the SepFormer processing conditions. Similarly, the RNN and SepFormer models increased the median intelligibility scores from 38% to 64% and 70% in the 10 dB SNR condition, respectively. The reduced benefit of SepFormer over RNN under this condition might be due to the fact that the 10 dB SNR condition is an easier listening condition where "de-noising" or enhancement of the stimulus is much less necessary. These results are consistent with previous studies on the application of machine learning-based algorithms to "de-noising" strategies towards better processing in cochlear implants (Goehring et al., 2017; Goehring et al., 2019; Hu & Loizou, 2010; Lai et al., 2018). Our study further confirms the promise of using DNN-based algorithms to solve speech-in-noise hearing problems among CI listeners. It serves as an important extension and optimization of previous work. While Hu and Loizou, 2010 and Lai et al., 2018 used the same speaker for training and testing, our study further investigated the models' capability of generalization by using different speech and non-speech materials for testing. We also used a much larger and diverse training dataset. The training dataset is of approximately 30 hours in total duration with 4 distinct target-masker configurations. Within each configuration, 10 different SNR conditions in a 1 dB step were implemented with equal representation. The models were also trained more extensively with 100 learning cycles or epochs, whereas some other studies trained their models with a much less number of learning cycles. For instance, Goehring et al., 2019 trained their model for only one epoch. Although the successful training with such a small number of learning cycles helps demonstrate the model's promise of being continuously trained "on the go" with smaller resources, we wanted to explore the full capacity of the models for noise reduction through a more thorough training process. In addition to the basic RNN model, we also implemented and tested the current best-performing model in the field of speech separation—the SepFormer, to explore the limit to which a DNN-based strategy could suppress the noise for CI devices. While, not surprisingly, the SepFormer model outperformed RNN in every test, it is a very complicated model containing over 26 million parameters. The processing time of such a heavy model turned out to be almost 5 times the duration of the input signal on average, which renders it not suitable for a real-time application such as a CI speech processor. The time constraint for the processing delay in a real-time device such as CI should be below about 10-20 ms to avoid disturbance in speech production and audio-visual integration (Bramsløw et al., 2018; Goehring et al., 2018; Goehring et al., 2019; Stone & Moore, 1999). Another limitation comes from the high demand for computational power and memory and it is unrealistic to run these highly complex state-of-the-art models within the CI processors. RNN model, on the other hand, is a much lighter model and only takes around 3% of the incoming signal duration for processing on average, which makes it promising to apply the RNN model to CI applications.

Improving the speech-in-noise hearing for CI listeners is the most challenging problem for CI research and the development of future CI technologies. The results from this study demonstrate the promise of using DNN-based technologies for noise reduction in more complicated and unpredictable, but more common everyday listening environments, where current signal processing strategies for noise reduction fail to produce satisfactory performance. Both models evaluated here in this study are single-channel models. Dual-channel models with spatial-hearing sensitivity might further improve the performance. Future implementation of models that consider more aspects of the "cocktail-party listening", combined with current signal processing strategies, will hopefully be integrated and adopted by CI devices for better speech-in-noise hearing.



(a)

(b)

Figure 4.3. Individual sentence recognition performance plotted as a function of each processing condition. The boxes depict the values between the 25th and 75th percentiles, and the whiskers represent minimum and maximum values. Medians are shown as horizontal lines. (a) SNR = 5 dB (b) SNR = 10 dB; CCITT: speech-shaped noise, TTB: two-talker babble.

5. CONCLUSION

This thesis aimed to investigate the perceptual role of temporal fine structure (TFS) in everyday hearing, through the individual-difference approach. Although the results suggest that TFS might not be critical for providing listeners with more release from masking, TFS enables a listener to have a more resilient hearing that is less prone to corruption from reverberation. Furthermore, TFS processing was shown to be able to reduce an individual's reaction time during the speech-in-noise listening task, suggesting lessened listening effort. In addition to studying the TFS, this thesis also implemented two state-of-the-art deep neural network (DNN) models of speech enhancement and separation to explore their potential to be used for de-noising algorithms in hearing assistive devices, such as cochlear implants (CIs). Both models introduced significant improvements in all speech quality and intelligibility metrics tested. The models also significantly improved speech-in-noise listening performance among CI listeners.

Chapter 2 laid the foundation for this thesis to study the role of TFS in speech-in-noise listening, by establishing an array of TFS-sensitivity measurements at the individual level. Pinning down the perceptual significance of TFS has been elusive due to the limitation of the most commonly used approach—vocoding. It became very important that the vocoding strategy was avoided throughout this thesis. Therefore, we aimed to use individual-difference approach as an alternative. With the exploration and evaluation of many classic measurements of TFS sensitivity both behaviorally and electrophysiologically, we found that "adjusted" binaural assays may be well-suited for quantifying individual TFS processing. It should be emphasized that extraneous variables dominate both behavioral scores and EEG amplitude metrics, rendering them ineffective. The raw metrics have to be adjusted to account for the factors that are irrelevant to TFS processing.

With the available tools for assessing individual differences in TFS sensitivity, chapter 3 proceeded to investigate the perceptual role of TFS in noise by adding comprehensive measurements of speech intelligibility under various types of listening conditions. In particular, we examined most of the major aspects of speech-in-noise listening where TFS has been believed to play a role, including pitch perception, spatial localization, informational masking, listening in reverberation, and listening efforts. The comparison between the individual differences in TFS sensitivity and speech-in-noise intelligibility revealed that better TFS sensitivity does not necessarily introduce more benefits of release from masking for speech-in-noise listening, which is consistent with existing literature. However, this study adds to the previous knowledge that better TFS sensitivity serves as a protective shield against corruption due to reverberation and reduces reaction time suggesting lessened listening effort in everyday hearing.

Chapter 4 explored the potential of DNN technology for modeling speech separation and enhancement. Current state-of-the-art signal processing strategies provide limited benefits of noise reductions for hearing assistive technologies. Despite the wide adoption of DNN technologies for advanced speech applications such as speech recognition and synthesis, the application of DNN for noise-reduction problems of hearing assistive devices has been scarce. To fill this gap, we implemented two state-of-the-art DNN models. Both models introduced significant improvements in speech quality and intelligibility across all acoustic evaluation metrics tested. The models also introduced significant improvements in speech intelligibility scores among CI listeners in both speech and non-speech noise interference. This work serves as a proof of concept that DNN technology has the potential to be incorporated into the “front-end” noise-reduction algorithms in hearing assistive devices, as well as to complement other approaches by serving as a research tool to help generate and rapidly pilot-test hypotheses about various questions regarding speech-in-noise hearing before moving onto the formal experiments.

To my knowledge, this thesis is the first to have studied the role of TFS in the everyday hearing through an individual-difference approach by systematically measuring TFS sensitivity and speech-in-noise intelligibility under various listening conditions on a large typical-hearing population (n=200). To ameliorate COVID-related restrictions on in-person measurement, chapter 3 was fully carried out via remote testing, which also makes this study one of the first few studies that have successfully achieved this enterprise in the auditory field. The findings regarding the possible role of TFS in reducing the impact of reverberation and lessening listening effort raise the possibility that cochlear implant sound coding strategies could be improved by attempting to provide usable TFS information, and that the individualized TFS assays can also help predict listening outcomes in reverberant, real-world listening environments. The success with online experiments also serves as a template for future applications of remote testing in both research and clinics. Future work could extend from this associational study to look into the exact mechanisms underlying the relationship between the TFS processing and listening in reverberation, and listening effort. Finally, the DNN study is one of the first attempts to incorporate artificial intelligence technology into the field of hearing research. The remarkable improvements in speech quality and intelligibility brought by the DNN models show us a new direction for tackling the challenging noise-reduction problem in both research and clinic. The two models implemented in this study are both a single-channel monaural model of speech-in-noise hearing. Future work should focus on building more sophisticated yet realistic models of speech-in-noise listening, by integrating many other aspects of everyday hearing, such as spatial hearing, selective attention, and even audio-visual integration.

REFERENCES

Ahissar, E., & Vaadia, E. (1990). Oscillatory activity of single units in a somatosensory cortex of an awake monkey and their possible role in texture analysis. [Publisher: National Academy of Sciences Section: Research Article]. *Proceedings of the National Academy of Sciences*, *87*(22), 8935–8939. https://doi.org/10.1073/pnas.87.22.8935

Apoux, F., Crouzet, O., & Lorenzi, C. (2001). Temporal envelope expansion of speech in noise for normal-hearing and hearing-impaired listeners: Effects on identification performance and response times. *Hearing Research*, *153*(1), 123–131. https://doi.org/10.1016/S0378-5955(00)00265-3

Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *112*(5), 2086–2098. https://doi.org/10.1121/1.1510141

Ardoint, M., & Lorenzi, C. (2010). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hearing Research*, *260*(1), 89–95. https://doi.org/10.1016/j.heares.2009.12.002

Attneave, F., & Olson, R. K. (1971). Pitch as a Medium: A New Approach to Psychophysical Scaling [Publisher: University of Illinois Press]. *The American Journal of Psychology*, *84*(2), 147–166. https://doi.org/10.2307/1421351

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Royal Statistical Society.*

Bentsen, T., May, T., Kressner, A. A., & Dau, T. (2018). The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility [Publisher: Public Library of Science]. *PLOS ONE*, *13*(5), e0196924. https://doi.org/10.1371/journal.pone.0196924

Bernstein, J. G., & Oxenham, A. J. (2003). Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *113*(6), 3323–3334. https://doi.org/10.1121/1.1572146

Bernstein, J. G. W., & Brungart, D. S. (2011). Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *130*(1), 473–488. https://doi.org/10.1121/1.3589440

Bernstein, J. G. W., & Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *125*(5), 3358–3372. https://doi.org/10.1121/1.3110132

Bernstein, L. R., & Trahiotis, C. (2002). Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli". *The Journal of the Acoustical Society of America*, *112*(3), 1026–1036. https://doi.org/10.1121/1.1497620

Best, V., Ozmeral, E., Gallun, F. J., Sen, K., & Shinn-Cunningham, B. G. (2005). Spatial unmasking of birdsong in human listeners: Energetic and informational factors. *The Journal of the Acoustical Society of America*, *118*(6), 3766–3773. https://doi.org/10.1121/1.2130949

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. *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2019.02.031

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015). Individual Differences Reveal Correlates of Hidden Hearing Deficits. *Journal of Neuroscience*, *35*(5), 2161–2172. https://doi.org/10.1523/JNEUROSCI.3915-14.2015

Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2014). Rapid acquisition of auditory subcortical steady state responses using multichannel recordings. *Clinical Neurophysiology*, *125*(9), 1878–1888. https://doi.org/10.1016/j.clinph.2014.01.011

Bird, J., & Darwin, C. J. (1997). Effects of a difference in fundamental frequency in separating two sentences.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction [Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209

Borjigin, A., Hustedt-Mai, A. R., & Bharadwaj, H. M. (2021). *Individualized Assays of Temporal Coding in the Ascending Human Auditory System* (tech. rep.) [Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article]. https://doi.org/10.1101/2021.09.13.460174

Bramsløw, L., Naithani, G., Hafez, A., Barker, T., Pontoppidan, N. H., & Virtanen, T. (2018). Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *144*(1), 172–185. https://doi.org/10.1121/1.5045322

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*(1), 23–36. https://doi.org/10.1016/S0095-4470(19)30909-X

Brughera, A., Dunai, L., & Hartmann, W. M. (2013). Human interaural time difference thresholds for sine tones: The high-frequency limit. *The Journal of the Acoustical Society of America*, *133*(5), 2839–2855. https://doi.org/10.1121/1.4795778

Buss, E., Hall, J. W. I., & Grose, J. H. (2004). Temporal Fine-Structure Cues to Speech and Pure Tone Modulation in Observers with Sensorineural Hearing Loss. *Ear and Hearing*, *25*(3), 242–250. https://doi.org/10.1097/01.AUD.0000130796.73809.09

Cameron, S., & Dillon, H. (2007). Development of the Listening in Spatialized Noise-Sentences Test (LISN-S). *Ear and Hearing*, *28*(2), 196–211.

Chen, J., & Wang, D. (2017). Long short-term memory for speaker generalization in supervised speech separation [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *141*(6), 4705–4714. https://doi.org/10.1121/1.4986931

Chen, J., Wang, Y., Yoho, S. E., Wang, D., & Healy, E. W. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *139*(5), 2604–2612. https://doi.org/10.1121/1.4948445

Churchill, T. H., Kan, A., Goupell, M. J., & Litovsky, R. Y. (2014). Spatial hearing benefits demonstrated with presentation of acoustic temporal fine structure cues in bilateral cochlear implant listeners [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *136*(3), 1246–1256. https://doi.org/10.1121/1.4892764

Culling, J. F., & Stone, M. A. (2017). Energetic Masking and Masking Release. *The Auditory System at the Cocktail Party*.

Culling, J. F., Summerfield, Q., & Marshall, D. H. (1994). Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. *Speech Communication*, *14*(1), 71–95. https://doi.org/10.1016/0167-6393(94)90058-2

Cullington, H. E., & Zeng, F.-G. (2008). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *123*(1), 450–461. https://doi.org/10.1121/1.2805617

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*(9), 327–333. https://doi.org/10.1016/S1364-6613(97)01097-8

Darwin, C. J. (2005). Pitch and Auditory Grouping. In C. J. Plack, R. R. Fay, A. J. Oxenham, & A. N. Popper (Eds.), *Pitch: Neural Coding and Perception* (pp. 278–305). Springer. https://doi.org/10.1007/0-387-28958-5_8

Dawson, P. W., Mauger, S. J., & Hersbach, A. A. (2011). Clinical Evaluation of Signal-to-Noise Ratio–Based Noise Reduction in Nucleus® Cochlear Implant Recipients. *Ear and Hearing*, *32*(3), 382–390. https://doi.org/10.1097/AUD.0b013e318201c200

Deroche, M. L. D., Culling, J. F., Lavandier, M., & Gracco, V. L. (2017). Reverberation limits the release from informational masking obtained in the harmonic and binaural domains. *Attention, Perception, & Psychophysics*, *79*(1), 363–379. https://doi.org/10.3758/s13414-016-1207-3

Dreyer, A., & Delgutte, B. (2006). Phase Locking of Auditory-Nerve Fibers to the Envelopes of High-Frequency Sounds: Implications for Sound Localization [Publisher: American Physiological Society]. *Journal of Neurophysiology*, *96*(5), 2327–2341. https://doi.org/10.1152/jn.00326.2006

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *97*(1), 585–592. https://doi.org/10.1121/1.413112

Dye, R. H. (1990). The combination of interaural information across frequencies: Lateralization on the basis of interaural delay. *The Journal of the Acoustical Society of America*, *88*(5), 2159–2170. https://doi.org/10.1121/1.400113

Efron, B., & Stein, C. (1981). The Jackknife Estimate of Variance [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *9*(3), 586–596. Retrieved February 28, 2022, from https://www.jstor.org/stable/2240822

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., & Scollie, S. (2015). Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools [Conference Name: IEEE Signal Processing Magazine]. *IEEE Signal Processing Magazine*, *32*(2), 114–124. https://doi.org/10.1109/MSP.2014.2358871

Fay, R. R. (2005). *Sound Source Localization* (1st ed. 2005.). Springer New York.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *The Journal of the Acoustical Society of America*, *124*(3), 1627–1637. https://doi.org/10.1121/1.2951964

Freyman, R. L., Griffin, A. M., & Oxenham, A. J. (2012). Intelligibility of whispered speech in stationary and modulated noise maskers [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *132*(4), 2514–2523. https://doi.org/10.1121/1.4747614

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *104*(6), 3586–3596. https://doi.org/10.1121/1.423941

Gallun, F. J., McMillan, G. P., Kampel, S. D., Jakien, K. M., Srinivasan, N. K., Stansell, M. M., & Gordon, S. Y. (2015). Verification of an automated headphone-based test of spatial release from masking [Publisher: Acoustical Society of America]. *Proceedings of Meetings on Acoustics*, *25*(1), 050001. https://doi.org/10.1121/2.0000165

Gallun, F. J., Diedesch, A. C., Kampel, S. D., & Jakien, K. M. (2013). Independent impacts of age and hearing loss on spatial release in a complex auditory environment [Publisher: Frontiers]. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00252

Gilbert, G., & Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *The Journal of the Acoustical Society of America*, *119*(4), 2438–2444. https://doi.org/10.1121/1.2173522

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Goehring, T., Bolner, F., Monaghan, J. J. M., van Dijk, B., Zarowski, A., & Bleeck, S. (2017). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research*, *344*, 183–194. https://doi.org/10.1016/j.heares.2016.11.012

Goehring, T., Chapman, J. L., Bleeck, S., & Monaghan*, J. J. M. (2018). Tolerable delay for speech production and perception: Effects of hearing ability and experience with hearing aids [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14992027.2017.1367848]. *International Journal of Audiology*, *57*(1), 61–68. https://doi.org/10.1080/14992027.2017.1367848

Goehring, T., Keshavarzi, M., Carlyon, R. P., & Moore, B. C. J. (2019). Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *146*(1), 705–718. https://doi.org/10.1121/1.5119226

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks [ISSN: 2379-190X]. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Grose, J. H., & Mamo, S. K. (2010). Processing of temporal fine structure as a function of age. *Ear and hearing*, *31*(6), 755–760. https://doi.org/10.1097/AUD.0b013e3181e627e7

Grose, J. H., & Mamo, S. K. (2012). Frequency modulation detection as a measure of temporal processing: Age-related monaural and binaural effects. *Hearing Research*, *294*(1), 49–54. https://doi.org/10.1016/j.heares.2012.09.007

Harris, J. D. (1952). Pitch Discrimination [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *24*(6), 750–755. https://doi.org/10.1121/1.1906970

Hartmann, W. M., & Pumplin, J. (1988). Noise power fluctuations and the masking of sine signals [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *83*(6), 2277–2289. https://doi.org/10.1121/1.396358

Hasenstaub, A., Otte, S., Callaway, E., & Sejnowski, T. J. (2010). Metabolic cost as a unifying principle governing neuronal biophysics [Publisher: National Academy of Sciences Section: Biological Sciences]. *Proceedings of the National Academy of Sciences*, *107*(27), 12329–12334. https://doi.org/10.1073/pnas.0914886107

He, N.-j., Dubno, J. R., & Mills, J. H. (1998). Frequency and intensity discrimination measured in a maximum-likelihood procedure from young and aged normal-hearing subjects [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *103*(1), 553–565. https://doi.org/10.1121/1.421127

Healy, E. W., Delfarah, M., Johnson, E. M., & Wang, D. (2019). A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *145*(3), 1378–1388. https://doi.org/10.1121/1.5093547

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., & Wang, D. (2015). An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *138*(3), 1660–1669. https://doi.org/10.1121/1.4929493

Healy, E. W., Yoho, S. E., Wang, Y., & Wang, D. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *134*(4), 3029–3038. https://doi.org/10.1121/1.4820893

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *92*(2), 490–499. https://doi.org/10.1037/0033-2909.92.2.490

Henning, G. B. (1966). Frequency Discrimination of Random-Amplitude Tones [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *39*(2), 336–339. https://doi.org/10.1121/1.1909894

Henning, G. B. (1983). Lateralization of low-frequency transients. *Hearing Research*, *9*(2), 153–172. https://doi.org/10.1016/0378-5955(83)90025-4

Henry, K. R. (1995). Auditory nerve neurophonic recorded from the round window of the Mongolian gerbil. *Hearing Research*, *90*(1), 176–184. https://doi.org/10.1016/0378-5955(95)00162-6

Hersbach, A. A., Arora, K., Mauger, S. J., & Dawson, P. W. (2012). Combining Directional Microphone and Single-Channel Noise Reduction Algorithms: A Clinical Evaluation in Difficult Listening Conditions With Cochlear Implant Users. *Ear and Hearing*, *33*(4), e13. https://doi.org/10.1097/AUD.0b013e31824b9e21

Hershkowitz, R. M., & Durlach, N. I. (1969). Interaural Time and Amplitude jnds for a 500-Hz Tone. *The Journal of the Acoustical Society of America*, *46*(6B), 1464–1467. https://doi.org/10.1121/1.1911887

Hilbert, D. (1906). *Grundz?ge einer allgemeinen theorie der linearen integralgleichungen* [Google-Books-ID: caARAwAAQBAJ]. .

Hochmair, I., Hochmair, E., Nopp, P., Waller, M., & Jolly, C. (2015). Deep electrode insertion and sound coding in cochlear implants. *Hearing Research*, *322*, 14–23. https://doi.org/10.1016/j.heares.2014.10.006

Hopkins, K., & Moore, B. C. J. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *125*(1), 442–446. https://doi.org/10.1121/1.3037233

Hopkins, K., & Moore, B. C. J. (2010). Development of a fast method for measuring sensitivity to temporal fine structure information at low frequencies. *International Journal of Audiology*, *49*(12), 940–946. https://doi.org/10.3109/14992027.2010.512613

Hopkins, K., Moore, B. C. J., & Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *123*(2), 1140–1153. https://doi.org/10.1121/1.2824018

Houtsma, A. J. M., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *87*(1), 304–310. https://doi.org/10.1121/1.399297

Hu, Y., & Loizou, P. C. (2008). Evaluation of Objective Quality Measures for Speech Enhancement [Conference Name: IEEE Transactions on Audio, Speech, and Language Processing]. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238. https://doi.org/10.1109/TASL.2007.911054

Hu, Y., & Loizou, P. C. (2010). Environment-specific noise suppression for improved speech intelligibility by cochlear implant users [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *127*(6), 3689–3695. https://doi.org/10.1121/1.3365256

Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a divided speech identification task [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *123*(6), 4380–4392. https://doi.org/10.1121/1.2904825

Ihlefeld, A., & Shinn-Cunningham, B. G. (2011). Effect of source spectrum on sound localization in an everyday reverberant room. *The Journal of the Acoustical Society of America*, *130*(1), 324–333. https://doi.org/10.1121/1.3596476

Jakien Kasey M., & Gallun Frederick J. (2018). Normative Data for a Rapid, Automated Test of Spatial Release From Masking. *American Journal of Audiology*, *27*(4), 529–538. https://doi.org/10.1044/2018_AJA-17-0069

Jakien Kasey M., Kampel Sean D., Stansell Meghan M., & Gallun Frederick J. (2017). Validating a Rapid, Automated Test of Spatial Release From Masking [Publisher: American Speech-Language-Hearing Association]. *American Journal of Audiology*, *26*(4), 507–518. https://doi.org/10.1044/2017_AJA-17-0013

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *68*(4), 1115–1122. https://doi.org/10.1121/1.384982

Joris, P. X., Schreiner, C. E., & Rees, A. (2004). Neural Processing of Amplitude-Modulated Sounds [Publisher: American Physiological Society]. *Physiological Reviews*, *84*(2), 541–577. https://doi.org/10.1152/physrev.00029.2003

Joris, P. X., & Yin, T. C. T. (1992). Responses to amplitude-modulated tones in the auditory nerve of the cat [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *91*(1), 215–232. https://doi.org/10.1121/1.402757

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., & Zhang, W. (2019). A Comparative Study on Transformer vs RNN in Speech Applications. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 449–456. https://doi.org/10.1109/ASRU46091.2019.9003750

Keshavarzi, M., Goehring, T., Turner, R. E., & Moore, B. C. J. (2019). Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *145*(3), 1493–1503. https://doi.org/10.1121/1.5094765

Keshavarzi, M., Goehring, T., Zakis, J., Turner, R. E., & Moore, B. C. J. (2018). Use of a Deep Recurrent Neural Network to Reduce Wind Noise: Effects on Judged Speech Intelligibility and Sound Quality [Publisher: SAGE Publications Inc]. *Trends in Hearing*, *22*, 2331216518770964. https://doi.org/10.1177/2331216518770964

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418–435. https://doi.org/10.1121/1.2743154

Kidd, G., & Colburn, H. S. (2017). Informational Masking in Speech Recognition. *The Auditory System at the Cocktail Party*.

Kim, G., Lu, Y., Hu, Y., & Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *126*(3), 1486–1494. https://doi.org/10.1121/1.3184603

Kim, J., El-Khamy, M., & Lee, J. (2020). T-GSA: Transformer with Gaussian-Weighted Self-Attention for Speech Enhancement [ISSN: 2379-190X]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6649–6653. https://doi.org/10.1109/ICASSP40776.2020.9053591

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization [arXiv: 1412.6980]. *arXiv:1412.6980 [cs]*. Retrieved March 25, 2022, from http://arxiv.org/abs/1412.6980

Klumpp, R. G., & Eady, H. R. (1956). Some Measurements of Interaural Time Difference Thresholds. *The Journal of the Acoustical Society of America*, *28*(5), 859–860. https://doi.org/10.1121/1.1908493

Koerner, T. K., Muralimanohar, R. K., Gallun, F. J., & Billings, C. J. (2020). Age-Related Deficits in Electrophysiological and Behavioral Measures of Binaural Temporal Processing. *Frontiers in Neuroscience*, *14*, 1105. https://doi.org/10.3389/fnins.2020.578566

Kohlrausch, A., Fassel, R., van der Heijden, M., Kortekaas, R., van de Par, S., Oxenham, A. J., & Püschel, D. (1997). Detection of Tones in Low-noise Noise: Further Evidence for the Role of Envelope Fluctuations [Library Catalog: www.ingentaconnect.com Publisher: S. Hirzel Verlag].

Kokkinakis, K., & Loizou, P. C. (2011). Evaluation of objective measures for quality assessment of reverberant speech [ISSN: 2379-190X]. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2420–2423. https://doi.org/10.1109/ICASSP.2011.5946972

Kolbæk, M., Yu, D., Tan, Z.-H., & Jensen, J. (2017). Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks [Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing], *25*(10), 1901–1913. https://doi.org/10.1109/TASLP.2017.2726762

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions [Publisher: The Association for Research in Vision and Ophthalmology]. *Journal of Vision*, *5*(5), 8–8. https://doi.org/10.1167/5.5.8

Lai, Y.-H., Tsao, Y., Lu, X., Chen, F., Su, Y.-T., Chen, K.-C., Chen, Y.-H., Chen, L.-C., Po-Hung Li, L., & Lee, C.-H. (2018). Deep Learning–Based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients. *Ear and Hearing*, *39*(4), 795–809. https://doi.org/10.1097/AUD.0000000000000537

Laughlin, S. B., de Ruyter van Steveninck, R. R., & Anderson, J. C. (1998). The metabolic cost of neural information [Number: 1 Publisher: Nature Publishing Group]. *Nature Neuroscience*, *1*(1), 36–41. https://doi.org/10.1038/236

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning [Number: 7553 Publisher: Nature Publishing Group]. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lelo de Larrea-Mancera, E. S., Stavropoulos, T., Hoover, E. C., Eddins, D. A., Gallun, F. J., & Seitz, A. R. (2020). Portable Automated Rapid Testing (PART) for auditory assessment: Validation in a young adult normal-hearing population [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *148*(4), 1831–1851. https://doi.org/10.1121/10.0002108

Levitt, H., & Rabiner, L. R. (1967). Binaural Release From Masking for Speech and Gain in Intelligibility [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *42*(3), 601–608. https://doi.org/10.1121/1.1910629

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019). Neural Speech Synthesis with Transformer Network [Number: 01]. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6706–6713. https://doi.org/10.1609/aaai.v33i01.33016706

Licklider, J. C. R. (1948). The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *20*(2), 150–159. https://doi.org/10.1121/1.1906358

Lin, F. R., Niparko, J. K., & Ferrucci, L. (2011). Hearing Loss Prevalence in the United States. *Archives of Internal Medicine*, *171*(20), 1851–1853. https://doi.org/10.1001/archinternmed.2011.506

Litovsky, R. Y. (2005). Speech intelligibility and spatial release from masking in young children [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *117*(5), 3091–3099. https://doi.org/10.1121/1.1873913

Loizou, P. C., Lobo, A., & Hu, Y. (2005). Subspace algorithms for noise reduction in cochlear implants [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *118*(5), 2791–2793. https://doi.org/10.1121/1.2065847

Lorenzi, C., Debruille, L., Garnier, S., Fleuriot, P., & Moore, B. C. J. (2009). Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *125*(1), 27–30. https://doi.org/10.1121/1.2939125

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, *103*(49), 18866–18869. https://doi.org/10.1073/pnas.0607364103

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation [Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

Macpherson, E. A., & Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *111*(5), 2219–2236. https://doi.org/10.1121/1.1471898

Madhu, N., Spriet, A., Jansen, S., Koning, R., & Wouters, J. (2013). The Potential for Speech Intelligibility Improvement Using the Ideal Binary Mask and the Ideal Wiener Filter in Single Channel Noise Reduction Systems: Application to Auditory Prostheses [Conference Name: IEEE Transactions on Audio, Speech, and Language Processing]. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(1), 63–72. https://doi.org/10.1109/TASL.2012.2213248

Mauger, S. J., Dawson, P. W., & Hersbach, A. A. (2012). Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *131*(1), 327–336. https://doi.org/10.1121/1.3665990

May, T., & Dau, T. (2014). Requirements for the evaluation of computational speech segregation systems [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *136*(6), EL398–EL404. https://doi.org/10.1121/1.4901133

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual Differences Reveal the Basis of Consonance. *Current Biology*, *20*(11), 1035–1041. https://doi.org/10.1016/j.cub.2010.04.019

Mener, D. J., Betz, J., Genther, D. J., Chen, D., & Lin, F. R. (2013). Hearing Loss and Depression in Older Adults. *Journal of the American Geriatrics Society*, *61*(9), 1627–1629. https://doi.org/10.1111/jgs.12429

Metting van Rijn, A. C., Peper, A., & Grimbergen, C. A. (1990). High-quality recording of bioelectric events. *Medical and Biological Engineering and Computing*, *28*(5), 389–397. https://doi.org/10.1007/BF02441961

Mick, P., Kawachi, I., & Lin, F. R. (2014). The Association between Hearing Loss and Social Isolation in Older Adults [Publisher: SAGE Publications Inc]. *Otolaryngology–Head and Neck Surgery*, *150*(3), 378–384. https://doi.org/10.1177/0194599813518021

Mok, B. A., Viswanathan, V., Borjigin, A., Singh, R., & Bharadwaj, H. (2020). Anonymous multipart web-based psychoacoustics: Infrastructure, hearing screening, and comparison with lab-based studies [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *148*(4), 2713–2714. https://doi.org/10.1121/1.5147521

Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., & Bleeck, S. (2017). Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *141*(3), 1985–1998. https://doi.org/10.1121/1.4977197

Moore, B. C. J. (1973). Frequency difference limens for short-duration tones [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *54*(3), 610–619. https://doi.org/10.1121/1.1913640

Moore, B. C. J. (1968). Parallels betwen frequency selectivity measured psychophysically and in cochlear mechanics.

Moore, B. C. J., & Ernst, S. M. A. (2012). Frequency difference limens at high frequencies: Evidence for a transition from a temporal to a place code [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *132*(3), 1542–1547. https://doi.org/10.1121/1.4739444

Moore, B. C. J., & Sek, A. (1996). Detection of frequency modulation at low modulation rates: Evidence for a mechanism based on phase locking. *The Journal of the Acoustical Society of America*, *100*(4), 2320–2331. https://doi.org/10.1121/1.417941

Moore, B. C. J., & Sek, A. (2009). Development of a fast method for determining sensitivity to temporal fine structure. *International Journal of Audiology*, *48*(4), 161–171. https://doi.org/10.1080/14992020802475235

Murphy, J., Summerfield, A. Q., O'Donoghue, G. M., & Moore, D. R. (2011). Spatial hearing of normally hearing and cochlear implanted children. *International Journal of Pediatric Otorhinolaryngology*, *75*(4), 489–494. https://doi.org/10.1016/j.ijporl.2011.01.002

Ochoa, J., & Torebjörk, E. (1983). Sensations evoked by intraneural microstimulation of single mechanoreceptor units innervating the human hand. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1983.sp014873]. *The Journal of Physiology*, *342*(1), 633–654. https://doi.org/10.1113/jphysiol.1983.sp014873

Ohm, G. S. (1843). Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.18431350802]. *Annalen der Physik*, *135*(8), 513–565. https://doi.org/10.1002/andp.18431350802

Oxenham, A. J. (2012). Pitch Perception [Publisher: Society for Neuroscience Section: Mini-Reviews]. *Journal of Neuroscience*, *32*(39), 13335–13338. Retrieved April 5, 2022, from https://www.jneurosci.org/content/32/39/13335

Oxenham, A. J. (2013). Revisiting place and temporal theories of pitch. *Acoustical Science and Technology*, *34*(6), 388–396. https://doi.org/10.1250/ast.34.388

Oxenham, A. J., Bernstein, J. G. W., & Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception. Retrieved March 5, 2022, from https://www.pnas.org/doi/abs/10.1073/pnas.0306958101

Oxenham, A. J., Micheyl, C., Keebler, M. V., Loper, A., & Santurette, S. (2011). Pitch perception beyond the traditional existence region of pitch. Retrieved March 5, 2022, from https://www.pnas.org/doi/abs/10.1073/pnas.1015291108

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. *The Journal of the Acoustical Society of America*, *125*(1), 457–468. https://doi.org/10.1121/1.3021299

Palmer, A. R., & Russell, I. J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research*, *24*(1), 1–15. https://doi.org/10.1016/0378-5955(86)90002-X

Palomäki, K. J., Brown, G. J., & Wang, D. (2004). A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, *43*(4), 361–378. https://doi.org/10.1016/j.specom.2004.03.005

Papesh, M. A., Folmer, R. L., & Gallun, F. J. (2017). Cortical Measures of Binaural Processing Predict Spatial Release from Masking Performance. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00124

Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V., & Polley, D. B. (2020). Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing (B. G. Shinn-Cunningham, H. Luo, F.-G. Zeng, & C. Lorenzi, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *9*, e51419. https://doi.org/10.7554/eLife.51419

Picton, T. W., John, M. S., Dimitrijevic, A., & Purcell, D. (2003). Human auditory steady-state responses: Respuestas auditivas de estado estable en humanos. *International Journal of Audiology*, *42*(4), 177–219. https://doi.org/10.3109/14992020309101316

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *114*(1), 446–454. https://doi.org/10.1121/1.1579009

Richards, D. G., & Wiley, R. H. (1980). Reverberations and Amplitude Fluctuations in the Propagation of Sound in a Forest: Implications for Animal Communication [Publisher: The University of Chicago Press]. *The American Naturalist*, *115*(3), 381–399. https://doi.org/10.1086/283568

Romo, R., & Salinas, E. (1999). Sensing and deciding in the somatosensory system. *Current Opinion in Neurobiology*, *9*(4), 487–493. https://doi.org/10.1016/S0959-4388(99)80073-7

Rose, J. E., Brugge, J. F., Anderson, D. J., & Hind, J. E. (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. [Publisher: American Physiological Society]. *Journal of Neurophysiology*, *30*(4), 769–793. https://doi.org/10.1152/jn.1967.30.4.769

Rosen, S., & Olin, P. (1965). Hearing loss and coronary heart disease. *Bulletin of the New York Academy of Medicine*, *41*(10), 1052–1068. Retrieved April 12, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1750791/

Rothauser, E. H. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. on Audio and Electroacoustics*, *AU-17,3*, 225–246. https://doi.org/10.1109/TAU.1969.1162058

Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences*, *108*(37), 15516–15521. https://doi.org/10.1073/pnas.1108912108

Russo, F. A., & Pichora-Fuller, M. K. (2008). Tune In or Tune Out: Age-Related Differences in Listening to Speech in Music. *Ear and Hearing*, *29*(5), 746–760. https://doi.org/10.1097/AUD.0b013e31817bdd1f

Santarelli, R., del Castillo, I., Rodríguez-Ballesteros, M., Scimemi, P., Cama, E., Arslan, E., & Starr, A. (2009). Abnormal Cochlear Potentials from Deaf Patients with Mutations in the Otoferlin Gene. *Journal of the Association for Research in Otolaryngology*, *10*(4), 545. https://doi.org/10.1007/s10162-009-0181-z

Scalart, P., & Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation [ISSN: 1520-6149]. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, *2*, 629–632 vol. 2. https://doi.org/10.1109/ICASSP.1996.543199

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123. https://doi.org/10.1016/j.visres.2016.02.002

Seebeck, A. (1841). Beobachtungen über einige Bedingungen der Entstehung von Tönen [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.18411290702]. *Annalen der Physik*, *129*(7), 417–436. https://doi.org/10.1002/andp.18411290702

Sęk, A., & Moore, B. C. J. (2012). Implementation of two tests for measuring sensitivity to temporal fine structure. *International Journal of Audiology*, *51*(1), 58–63. https://doi.org/10.3109/14992027.2011.605808

Shadlen, M. N., & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, *4*(4), 569–579. https://doi.org/10.1016/0959-4388(94)90059-0

Shinn-Cunningham, B., Best, V., & Lee, A. K. C. (2017). Auditory Object Formation and Selection. *The Auditory System at the Cocktail Party*.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Shower, E. G., & Biddulph, R. (1931). DIFFERENTIAL PITCH SENSITIVITY OF THE EAR, 14.

Singer, W., & Gray, C. M. (1995). Visual Feature Integration and the Temporal Correlation Hypothesis [_eprint: https://doi.org/10.1146/annurev.ne.18.030195.003011]. *Annual Review of Neuroscience*, *18*(1), 555–586. https://doi.org/10.1146/annurev.ne.18.030195.003011

Sinha, B. K., Hartung, J., & Knapp, G. (2011). *Statistical Meta-Analysis with Applications* [Google-Books-ID: JEoNB_2NONQC]. John Wiley & Sons.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87. https://doi.org/10.1038/416087a

Snyder, R. L., & Schreiner, C. E. (1985). Forward masking of the auditory nerve neurophonic (ANN) and the frequency following response (FFR). *Hearing Research*, *20*(1), 45–62. https://doi.org/10.1016/0378-5955(85)90058-9

Srinivasan, N. K., Jakien, K. M., & Gallun, F. J. (2016). Release from masking for small spatial separations: Effects of age and hearing loss [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *140*(1), EL73–EL78. https://doi.org/10.1121/1.4954386

Stecker, C., & Gallun, F. (2012). Binaural Hearing, Sound Localization, and Spatial hearing. Translational perspectives in Auditory Neuroscience: Normal Aspects of Hearing. *Translational Perspectives in Auditory Neuroscience*. Plural Publishing, Incorporated.

Stickney, G. S., Assmann, P. F., Chang, J., & Zeng, F.-G. (2007). Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *122*(2), 1069–1078. https://doi.org/10.1121/1.2750159

Stone, M. A., & Moore, B. C. J. (1999). Tolerable Hearing Aid Delays. I. Estimation of Limits Imposed by the Auditory Path Alone Using Simulated Hearing Losses. *Ear and Hearing*, *20*(3), 182–192. Retrieved March 22, 2022, from https://journals.lww.com/ear-hearing/Fulltext/1999/06000/Tolerable_Hearing_Aid_Delays__I__Estimation_of.2.aspx?casa_token=yreyDi0_9qIAAAAA:Iihb31R3n_Z_NBPd9MPB7lodfbL4rQI4-7XEIdi5b_R2hGOmL6QUTiXGOxy6ME9BXoI4nu66HB82hfM1blELt7k

Stone, M. A., & Moore, B. C. J. (2014). On the near non-existence of “pure” energetic masking release for speech. *The Journal of the Acoustical Society of America*, *135*(4), 1967–1977. https://doi.org/10.1121/1.4868392

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *The Journal of the Acoustical Society of America*, *125*(5), 3328–3345. https://doi.org/10.1121/1.3097469

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is All You Need in Speech Separation [arXiv: 2010.13154]. *arXiv:2010.13154 [cs, eess]*. Retrieved March 18, 2021, from http://arxiv.org/abs/2010.13154

Summers Van, & Leek Marjorie R. (1998). F0 Processing and the Seperation of Competing Speech Signals by Listeners With Normal Hearing and With Hearing Loss [Publisher: American Speech-Language-Hearing Association]. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1294–1306. https://doi.org/10.1044/jslhr.4106.1294

Swaminathan, J., & Heinz, M. G. (2012). Psychophysiological Analyses Demonstrate the Importance of Neural Envelope Coding for Speech Perception in Noise. *Journal of Neuroscience*, *32*(5), 1747–1756. https://doi.org/10.1523/JNEUROSCI.4493-11.2012

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech [ISSN: 2379-190X]. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4214–4217. https://doi.org/10.1109/ICASSP.2010.5495701

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech [Conference Name: IEEE Transactions on Audio, Speech, and Language Processing]. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125–2136. https://doi.org/10.1109/TASL.2011.2114881

Talbot, W. H., Darian-Smith, I., Kornhuber, H. H., & Mountcastle, V. B. (1968). The sense of flutter-vibration: Comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. [Publisher: American Physiological Society]. *Journal of Neurophysiology*, *31*(2), 301–334. https://doi.org/10.1152/jn.1968.31.2.301

Tang, Y., Arnold, C., & Cox, T. (2017). A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners. *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, *1*, 5. https://doi.org/10.3390/ohbm1010005

Tollin, D. J., Koka, K., & Tsai, J. J. (2008). Interaural Level Difference Discrimination Thresholds for Single Neurons in the Lateral Superior Olive [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *28*(19), 4848–4860. https://doi.org/10.1523/JNEUROSCI.5421-07.2008

Undurraga, J. A., Haywood, N. R., Marquardt, T., & McAlpine, D. (2016). Neural Representation of Interaural Time Differences in Humans—an Objective Measure that Matches Behavioural Performance. *Journal of the Association for Research in Otolaryngology*, *17*(6), 591–607. https://doi.org/10.1007/s10162-016-0584-6

Uusitalo, M. A., & Ilmoniemi, R. J. (1997). Signal-space projection method for separating MEG or EEG into components. *Medical and Biological Engineering and Computing*, *35*(2), 135–140. https://doi.org/10.1007/BF02534144

Vaillancourt, V., Laroche, C., Giguère, C., & Soli, S. D. (2008). Establishment of Age-Specific Normative Data for the Canadian French Version of the Hearing in Noise Test for Children. *Ear and Hearing*, *29*(3), 453–466. https://doi.org/10.1097/01.aud.0000310792.55221.0c

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. Retrieved March 22, 2022, from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Verschooten, E., Desloovere, C., & Joris, P. X. (2018). High-resolution frequency tuning but not temporal coding in the human cochlea. *PLOS Biology*, *16*(10), e2005164. https://doi.org/10.1371/journal.pbio.2005164

Verschooten, E., & Joris, P. X. (2014). Estimation of Neural Phase Locking from Stimulus-Evoked Potentials. *Journal of the Association for Research in Otolaryngology*, *15*(5), 767–787. https://doi.org/10.1007/s10162-014-0465-9

Verschooten, E., Robles, L., & Joris, P. X. (2015). Assessment of the Limits of Neural Phase-Locking Using Mass Potentials [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *35*(5), 2255–2268. https://doi.org/10.1523/JNEUROSCI.2979-14.2015

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C. J., Joris, P. X., Heinz, M. G., & Plack, C. J. (2019). The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. *Hearing Research*, *377*, 109–121. https://doi.org/10.1016/j.heares.2019.03.011

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *66*(5), 1364–1380. https://doi.org/10.1121/1.383531

Vincent, E., Gribonval, R., & Fevotte, C. (2006). Performance measurement in blind audio source separation [Conference Name: IEEE Transactions on Audio, Speech, and Language Processing]. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1462–1469. https://doi.org/10.1109/TSA.2005.858005

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 [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *150*(3), 2230–2244. https://doi.org/10.1121/10.0006385

Viswanathan, V., Shinn-Cunningham, B. G., & Heinz, M. G. (2021). Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *150*(4), 2664–2676. https://doi.org/10.1121/10.0006527

Wallace, M. N., Rutkowski, R. G., Shackleton, T. M., & Palmer, A. R. (2000). Phase-locked responses to pure tones in guinea pig auditory cortex. *NeuroReport*, *11*(18), 3989–3993. Retrieved February 20, 2020, from https://journals.lww.com/neuroreport/Fulltext/2000/12180/Phase_locked_responses_to_pure_tones_in_guinea_pig.17.aspx

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR. In E. Vincent, A. Yeredor, Z. Koldovský, & P. Tichavský (Eds.), *Latent Variable Analysis and Signal Separation* (pp. 91–99). Springer International Publishing. https://doi.org/10.1007/978-3-319-22482-4_11

Werner, L. (2017). Infants and Children at the Cocktail Party. *The Auditory System at the Cocktail Party*.

Whiteford, K. L., Kreft, H. A., & Oxenham, A. J. (2017). Assessing the Role of Place and Timing Cues in Coding Frequency and Amplitude Modulation as a Function of Age. *Journal of the Association for Research in Otolaryngology*, *18*(4), 619–633. https://doi.org/10.1007/s10162-017-0624-x

Whiteford, K. L., Kreft, H. A., & Oxenham, A. J. (2020). The role of cochlear place coding in the perception of frequency modulation (A. J. King, T. Reichenbach, T. Reichenbach, & E. A. Lopez-Poveda, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *9*, e58468. https://doi.org/10.7554/eLife.58468

Whiteford, K. L., & Oxenham, A. J. (2015). Using individual differences to test the role of temporal and place cues in coding frequency modulation. *The Journal of the Acoustical Society of America*, *138*(5), 3093–3104. https://doi.org/10.1121/1.4935018

WHO. (2020). Deafness and Hearing Loss - World Health Organization.

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America*, *85*(2), 868–878. https://doi.org/10.1121/1.397558

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *91*(3), 1648–1661. https://doi.org/10.1121/1.402445

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Wouters, J., & Vanden Berghe, J. (2001). Speech Recognition in Noise for Cochlear Implantees with a Two-Microphone Monaural Adaptive Noise Reduction System. *Ear and Hearing*, *22*(5), 420–430.

Yin, T. C., & Chan, J. C. (1990). Interaural time sensitivity in medial superior olive of cat. *Journal of Neurophysiology*, *64*(2), 465–488. https://doi.org/10.1152/jn.1990.64.2.465

Yuen Kevin C. P., & Yuan Meng. (2014). Development of Spatial Release From Masking in Mandarin-Speaking Children With Normal Hearing [Publisher: American Speech-Language-Hearing Association]. *Journal of Speech, Language, and Hearing Research*, *57*(5), 2005–2023. https://doi.org/10.1044/2014_JSLHR-H-13-0060

Zwicker, E. (1956). Die elementaren Grundlagen zur Bestimmung der Informationskapazität des Gehörs. Retrieved December 9, 2019, from https://www.ingentaconnect.com/content/dav/aaua/1956/00000006/00000004/art00008#

Zwislocki, J., & Feldman, R. S. (1956). Just Noticeable Differences in Dichotic Phase. *The Journal of the Acoustical Society of America*, *28*(5), 860–864. https://doi.org/10.1121/1.1908495