JASA ARTICLE

Performance on stochastic figure-ground perception varies with individual differences in speech-in-noise recognition and working memory capacity

Michael A. Johns,[1,a)] Regina C. Calloway,[1] Ian Phillips,[2] Valerie P. Karuzis,[3] Kelsey Dutta,[1] Ed Smith,[4] Shihab A. Shamma,[5] Matthew J. Goupell,[4] and Stefanie E. Kuchinsky[2,b)]

[1]*Institute for Systems Research, University of Maryland, College Park, Maryland 20742, USA*

[2]*Audiology and Speech Pathology Center, Walter Reed National Military Medical Center, Bethesda, Maryland 20889, USA*

[3]*Applied Research Laboratory of Intelligence and Security, University of Maryland, College Park, Maryland 20742, USA*

[4]*Department of Hearing and Speech Sciences, University of Maryland, College Park, Maryland 20742, USA*

[5]*Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742, USA*

ABSTRACT:

Speech recognition in noisy environments can be challenging and requires listeners to accurately segregate a target speaker from irrelevant background noise. Stochastic figure-ground (SFG) tasks in which temporally coherent inharmonic pure-tones must be identified from a background have been used to probe the non-linguistic auditory stream segregation processes important for speech-in-noise processing. However, little is known about the relationship between performance on SFG tasks and speech-in-noise tasks nor the individual differences that may modulate such relationships. In this study, 37 younger normal-hearing adults performed an SFG task with target figure chords consisting of four, six, eight, or ten temporally coherent tones amongst a background of randomly varying tones. Stimuli were designed to be spectrally and temporally flat. An increased number of temporally coherent tones resulted in higher accuracy and faster reaction times (RTs). For ten target tones, faster RTs were associated with better scores on the Quick Speech-in-Noise task. Individual differences in working memory capacity and self-reported musicianship further modulated these relationships. Overall, results demonstrate that the SFG task could serve as an assessment of auditory stream segregation accuracy and RT that is sensitive to individual differences in cognitive and auditory abilities, even among younger normal-hearing adults. https://doi.org/10.1121/10.0016756

(Received 18 March 2022; revised 7 December 2022; accepted 10 December 2022; published online 18 January 2023)

[Editor: James F. Lynch] Pages: 286–303

I. INTRODUCTION

Listening to speech in background noise is a common experience (Burke and Naylor, 2020) that can become challenging, even for younger adults with normal hearing (Zekveld *et al.*, 2010). Speech-in-noise recognition is related to a listener's ability to group similar acoustic elements together while separating dissimilar acoustic elements into distinct sources (called auditory stream segregation; see Moore and Gockel, 2002 for a review).

Tests of speech-in-noise recognition used in the audiology clinic generally involve patients recognizing words or sentences spoken by a target speaker in the presence of a competing sound, such as a competing background speaker, multi-talker babble, or noise (e.g., Killion *et al.*, 2004; Wilson, 2003). Although performance on these tests provides important information about functional hearing abilities, use of speech materials engages language comprehension and occasionally production processes that cannot be readily disentangled from auditory stream segregation processes. In particular, word, phrase, and sentence stimuli can vary substantially in intelligibility due to their phonetic/phonemic, lexical, and semantic properties (Mattys *et al.*, 2012). Indeed, language ability is a major indicator of speech-in-noise recognition performance (Rogers *et al.*, 2006). For example, better lexical access ability is associated with better speech-in-noise recognition for both monolingual and bilingual individuals (Kaandorp *et al.*, 2016). Although early bilinguals who learn their first and second language simultaneously during childhood perform similarly in their first and second language on speech-in-noise recognition tasks, late bilinguals can show worse performance in their second language (Coulter *et al.*, 2021) and bilinguals sometimes perform worse than monolinguals (Kaandorp *et al.*, 2016; Rogers *et al.*, 2006). These linguistic influences on performance further highlight the necessity of developing tasks that do not use speech materials to evaluate auditory stream segregation ability in as wide a range of listeners as possible.

To more precisely assess (and perhaps ultimately treat) the speech-recognition deficits that individuals often report when listening in noise (Pang *et al.*, 2019; Shinn-Cunningham and Best, 2008), identification tasks using stochastic figure-ground

[a)]Electronic mail: maj@umd.edu

[b)]Also at: Department of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742, USA.

https://doi.org/10.1121/10.0016756

(SFG) stimuli (often called "tone clouds") may provide mechanistic insights into the auditory stream segregation processes that underlie speech-recognition abilities with non-linguistic stimuli. SFG tasks have been used to investigate the segregation of background and foreground information in visual processing (Huang *et al.*, 2020; Lamme, 1995), motion perception (Chen *et al.*, 2005), and auditory processing more generally (Teki *et al.*, 2016). When used as a measure of auditory stream segregation, SFG tasks often require listeners to segregate inharmonic temporally synchronous target tones from otherwise spectrally random background tones. Specifically, these stimuli largely consist of consecutive non-repeating chords of pure-tone components selected across a wide range of frequencies (the "ground"). A portion of this is then overlaid with repeating chords consisting of the same set of pure-tone components (the "figure"). Listeners are tasked to indicate when they hear the target figure "pop out" from the ground.

Behavioral and neuroimaging studies have demonstrated that SFG tasks engage segregation processes that underlie speech perception (Teki *et al.*, 2011; Teki *et al.*, 2013). Although SFG and speech-in-noise tasks are both understood to require auditory stream segregation processes, little is known about the relationship between performance on these tasks and the individual difference factors that may modulate such a relationship. The present study takes a next step at better understanding these relationships.

As will be discussed, the present study aimed to do the following: (1) Assess auditory stream segregation ability on an SFG task designed to more closely approximate real-world scenarios and determine its psychometric properties; (2) assess the relationship between this SFG task and a commonly used speech-in-multi-talker-babble recognition test (Quick Speech-in-Noise or QSIN task; Killion *et al.*, 2004); and (3) investigate the extent to which individual difference factors in non-auditory working memory capacity (WMC) and self-reported musicianship affect SFG task performance and its relation with speech-in-noise recognition.

A. Auditory stream segregation and SFG tasks

The temporal coherence model (Shamma *et al.*, 2011) proposes that auditory stream segregation occurs in two steps: (1) feature analysis, in which populations of neurons are tuned to temporal and spectral auditory information, and (2) evaluation of the temporal coherence of auditory stimuli. SFG tasks have demonstrated that this process readily occurs during active and passive listening (e.g., Teki *et al.*, 2011), although the likelihood of its occurrence is affected by temporal coherence of the stimuli (the relationship among auditory components over a period of time; O'Sullivan *et al.*, 2015; Teki *et al.*, 2011; Teki *et al.*, 2013) and perceptual load (i.e., demands on perceptual capacity for incoming information across or within perceptual domains; Molloy *et al.*, 2019). Because of the many ways in which task demands and acoustic properties can be adjusted, SFG tasks provide a rich avenue for examining auditory stream segregation and its relationship to speech-in-noise perception.

SFG tasks often include non-linguistic stimuli in which multiple pure tones at random frequencies are repeated at given intervals over time with a subset of tones forming a figure that can be distinguished from temporally or spectrotemporally incoherent background tones (e.g., Teki *et al.*, 2011). In these tasks, one or more spectrotemporal properties of the figure tones are manipulated to influence the ease of figure detection. For instance, studies have investigated the effect of coherence (i.e., the number of temporally coherent pure-tone components in the figure; O'Sullivan *et al.*, 2015; Teki *et al.*, 2011; Teki *et al.*, 2013; Teki *et al.*, 2016). Detecting temporal coherence within the figure allows for segregation of the background from the figure into two separate auditory streams (Shamma *et al.*, 2011; Teki *et al.*, 2011; Teki *et al.*, 2013). Within SFG tasks, successful segregation is needed for figure detection, which is often behaviorally measured by accuracy (e.g., hit rate or d') of detecting the figure (e.g., O'Sullivan *et al.*, 2015; Teki *et al.*, 2011), or detecting gaps within the figure portion of a stimulus (i.e., figure-gap detection task; Holmes *et al.*, 2021; Holmes and Griffiths, 2019). For example, Teki *et al.* (2011) manipulated coherence (number of pure-tone components) and duration (number of figure chords) of stimuli in a figure detection task, finding that increases in both resulted in increased performance for figure detection. In a figure-gap-detection task, Holmes *et al.* (2021) found that decreased target-to-masker ratio (TMR) thresholds resulted in lower signal detection for three-chord 1200-ms figures containing a 200-ms gap amongst background noise.

B. SFG performance and speech-in-noise perception

Performance on various forms of SFG tasks has been positively associated with speech-in-noise perception ability, both behaviorally (e.g., Holmes and Griffiths, 2019; Teki *et al.*, 2016) and neurally (Holmes *et al.*, 2021; O'Sullivan *et al.*, 2015; Teki *et al.*, 2011). For example, using a figure-gap detection task, Holmes *et al.* (2021) established that d' scores on an SFG task were positively correlated with those for a speech-in-noise task with 16-talker babble. Similarly, Holmes and Griffiths (2019) found that TMR thresholds in a speech-in-noise task were positively correlated with TMR thresholds on a figure-gap discrimination task, both when the three figure frequencies remained the same and when the three figure frequencies changed together over time (i.e., when tones were based on multiples of first formants from a speech-in-noise task). Task difficulty may also affect the relationship between performance on speech-in-noise and SFG tasks; if SFG task demands are too easy or too difficult, the SFG task may fail to show any relationship with speech-in-noise perception. For example, TMR thresholds in a speech-in-noise task were neither significantly correlated with performance on a simpler figure detection task of three-tone chords played, nor in a more difficult figure-gap discrimination task in

which figure frequencies were comprised of the first three formants of the sentences used in the speech-in-noise task (Holmes and Griffiths, 2019).

Neuroimaging studies that have used SFG tasks have observed patterns of neural activity similar to those observed during effortful speech processing (Adank, 2012; Alain *et al.*, 2018; Holmes *et al.*, 2021; Teki *et al.*, 2016), likely because both tasks are subserved by auditory perception, attention, and stream segregation processes. However, even passive SFG paradigms, which do not require an overt response from participants, have been shown to elicit similar patterns of neural activity to active listening tasks (Molloy *et al.*, 2019; Teki *et al.*, 2011), suggesting that auditory stream segregation in SFG tasks can be primarily driven by low-level bottom-up processes.

1. Individual differences

In addition to the acoustic properties of the stimulus, interactions among individual difference factors (e.g., WMC, age, and musical experience) may also contribute to differences in auditory perception, stream segregation, and temporal auditory acuity that underlie speech-in-noise recognition. For example, the relationship between WMC and speech-in-noise recognition appears to be driven by shared domain-general cognitive capacities rather than audibility. The Ease of Language Understanding model (Rönnberg *et al.*, 2008; Rönnberg *et al.*, 2013) states that working memory processes are brought online as listeners try to infer meaning from an acoustic input that mismatches representations in long-term memory. Other perspectives note an even more domain-general role for working memory in speech-in-noise perception; a degraded signal may slow encoding, causing interference as one stimulus (e.g., phoneme, word) is still being perceived while the next one rapidly arrives (Wingfield *et al.*, 2015). Indeed, associations between WMC and the perception of phonemes, syllables, words, and sentences has been observed (see Akeroyd, 2008 for a review), and auditory WMC for speech and non-speech materials has been positively associated with speech-in-noise recognition (Bidelman and Yoo, 2020; Lad *et al.*, 2020). This relationship is not always observed, however, particularly in younger adults (e.g., Füllgrabe and Rosen, 2016a,b; Vermeire *et al.*, 2019) who tend to have better speech-in-noise perception than older adults (Presacco *et al.*, 2016; Vermeire *et al.*, 2019). Nonetheless, the most consistent relationship between WMC and speech-in-noise recognition has been observed in studies that have used the reading span (RSPAN) task (e.g., Akeroyd, 2008; Besser *et al.*, 2013), a visually presented WMC span measure (Daneman and Carpenter, 1980; Rönnberg, 1990).

Although the relationship between WMC and speech-in-noise tasks is sometimes dependent on age, studies investigating musical expertise have found that, for both younger (Parbery-Clark *et al.*, 2009) and older adults (Parbery-Clark *et al.*, 2011), individuals with greater musical experience may perform better on auditory WMC and speech-in-noise tasks (see Coffey *et al.*, 2017 for a review on musician advantages in speech-in-noise tasks). For example, trained musicians appear to show better auditory stream segregation (Marozeau *et al.*, 2010; Zendel and Alain, 2009), better auditory working memory for digits and words (Parbery-Clark *et al.*, 2011), and greater temporal auditory acuity (Parbery-Clark *et al.*, 2011; Rammsayer and Altenmüller, 2006; Rammsayer *et al.*, 2012) than non-musicians (cf., Sherbon, 1975), although speaking a tonal language shows similar benefits (Bidelman *et al.*, 2013). In a study examining speech perception with multiple competing talkers from different spatial locations, Bidelman and Yoo (2020) found that with an increasing number of distractor speakers, musicians were faster at identifying the target speaker, showed less of a decline in speech recognition as the number of distractor speakers increased and were better at identifying the location of the target speaker compared to non-musicians. Similarly, musicians performed better on the QSIN compared to non-musicians. When controlling for WMC, the relationship between musical training and QSIN scores persisted but the relationship between musical training and speech recognition performance in the presence of competing speech was no longer significant. Musicians also demonstrate better auditory frequency discrimination acuity (Parbery-Clark *et al.*, 2009). The possibility of enhanced performance on speech-in-noise tasks for those with higher WMC or musical training could be because (1) individuals with greater WMC may be less impacted by background noise and/or (2) musicians may be able to better perceive more subtle auditory cues (like pitch) than non-musicians, allowing for less effort and more resources to be devoted to the demands of auditory WMC tasks (Parbery-Clark *et al.*, 2009). Of note, although there is evidence that musicians are better than non-musicians at speech-in-noise recognition tasks, this advantage is not always observed and may be related to task difficulty and level of linguistic information needed to complete the task (Coffey *et al.*, 2017).

C. Present study

Based on previous findings of behavioral and neural similarities between speech-in-noise and SFG task performance, the present study aims to characterize the relationship between an SFG task and a standard measure of speech-in-multi-talker babble recognition (QSIN) and to identify potential effects of more sensitive listening abilities, as might occur with musical training (e.g., Bidelman and Yoo, 2020; Parbery-Clark *et al.*, 2009; Parbery-Clark *et al.*, 2011; Yoo and Bidelman, 2019). In addition, the present study seeks to address some potential confounds in the stimulus design of prior SFG tasks. Compared to previous studies employing SFG tasks, the stimuli used in the present study differ in three main aspects. First, previous studies have focused on the detection of figures composed of rapid-rate (> 10 Hz) consecutive chords with an inter-chord interval (ICI) of 0 ms (e.g., Holmes *et al.*, 2021; Teki *et al.*, 2011; though see Teki *et al.*, 2013 and Teki *et al.*, 2016 for

https://doi.org/10.1121/10.0016756

some exceptions) or consecutive figures containing a single gap (e.g., Holmes *et al.*, 2021; Holmes and Griffiths, 2019). The present study instead examined figure detection of non-consecutive and aperiodic figure chords at a slower rate (~4 Hz) with random ICIs. This was done to mitigate the use of periodicity as a potential cue (e.g., Elhilali *et al.*, 2009); rather, to successfully identify a figure, individuals would instead need to rely on the build up of correlated tones (repeated frequencies) over a short period of time (Shamma *et al.*, 2011).

Second, we applied a random jitter to the pure-tone components comprising the background chords such that the onsets of the corresponding tones were asynchronous. This differs from previous studies where the background consisted of coherent, albeit random, chords of tones with identical onsets. The jittered background, we argue, is a more accurate depiction of what listeners might encounter in real-world scenarios: a background that is temporally uncorrelated from an internally consistent figure.

Third, almost all prior SFG tasks reviewed here employed a randomly selected number of pure-tone components in the background chords (as few as 5 to as many as 21), with the exception of O'Sullivan *et al.* (2015) who employed 15 tones per chord. This meant that the intensity of the background chords would vary while the intensity of the figure chords would remain constant (and, possibly, higher than the background). The SFG stimuli in the present study were designed to be relatively temporally and spectrally flat such that the background tones—despite being jittered—were evenly distributed across both time and frequencies as a way to negate potential confounds related to changes in intensity that may alert listeners to the presence of a figure.

Also, unlike prior SFG studies, a measure of reaction time (RT) to figure detection was included rather than manipulating figure duration explicitly. RT in the present study serves as a measure of processing speed. Specifically, RT is a continuous measure that reflects the amount of time needed for auditory stream segregation and decision execution. As such, RTs add further insight into how auditory segregation processes are influenced by stimulus features and individual differences in non-auditory WMC and musicianship. Trial-level accuracy and RT data were analyzed using generalized linear mixed-effects regression, which has been shown to better account for subject- and item-level variability and produce smaller type I error rates than analyses performed on aggregated data (Murayama *et al.*, 2014).

We generated three principal hypotheses to address the three study aims. In line with previous work (e.g., Teki *et al.*, 2011), it was hypothesized that (1) increasing temporal coherence (i.e., the number of tones in each figure chord) will facilitate figure detection in the SFG task reflected in increased accuracy and decreased RTs. Furthermore, given the neural and behavioral associations between speech-in-noise recognition and SFG task performance, it was also hypothesized that (2) individuals exhibiting better performance in the SFG task will also exhibit better speech-in-multi-talker babble recognition scores on the QSIN. Finally, given that individual differences in WMC and musicianship have been linked to speech-in-noise recognition (e.g., Parbery-Clark *et al.*, 2009; cf., Madsen *et al.*, 2019), it was also hypothesized that (3) individual differences in WMC and self-reported musicianship (a) would be related to SFG task performance examined in hypothesis 1 and (b) would explain variability in the association between SFG and QSIN performance examined in hypothesis 2.

II. METHOD

This study was approved by the University of Maryland's Institutional Review Board (IRB) and the U.S. Department of Navy Human Research Protection Program. Due to the COVID-19 pandemic, all tasks were administered remotely with proctors supervising testing sessions via Zoom (Zoom Video Communications, Inc., San Jose, CA). PsychoPy (Peirce *et al.*, 2019) was used to create the SFG, Reading Span (v2020.2.10), and QSIN tasks (v2020.2.4) administered on Google Chrome (Google, LLC, Mountain View, CA) via Pavlovia, an online platform for remote data collection. A recent study demonstrated that PsychoPy is able to achieve a mean RT measurement precision of 1.36 ms (Windows 10, Chrome) when administered online, despite a relatively long—although consistent—lag of 43.95 ms (see Bridges *et al.*, 2020 for further discussion). In the present study, sensible RT data were successfully obtained for the SFG task using this online platform.

All tasks were completed using wired or Bluetooth-connected headphones on either a desktop computer or a laptop. Participants were asked to confirm that their laptop was fully charged or plugged in, and that their Bluetooth-connected headphones were fully charged (if applicable). All responses were collected via button press (e.g., the spacebar) only.

A. Participants

Thirty-seven participants, with an average age of 21.8 years (standard deviation, $SD = 3.9$, *range*: 18 to 35), were recruited from the University of Maryland's Paid Psychology Sona Systems research platform, from listserv emails sent in the University community, and from individuals who were previously interested in studies conducted in-person under the same IRB protocol. Participants were provided monetary compensation ($16/h in the form of an online gift card) for their participation. All participants were native speakers of American English, reported having normal or corrected-to-normal vision, normal hearing and no issues with ears/hearing, no personal history of neurological, neuropsychiatric, or psychiatric disorders or learning disabilities, and no impairments of their dominant hand that would affect making rapid button-press responses. Of the 37 participants, five were excluded from analyses: three participants were excluded due to language experience before age 12 (see the following section), and a further two participants were excluded due to technical issues.

B. Materials

1. Demographics, language, and music history questionnaire

Participants were asked to provide information about their demographic background, language history, and musical experience. Demographic information collected included gender, age, highest level of education, major/minor, and race/ethnicity. Language history background questions included native language, languages learned before age 12 years (as well as settings in which these languages were learned), if they spent time with friends or relatives who spoke a language other than English (including at home), and the native languages of their parents/guardians. Additionally, information about all languages other than English with which they had any experience was collected, including self-ratings of reading, writing, listening, and speaking abilities. Participants who had some experience with languages other than English before age 12 years (e.g., through non-immersion programs in elementary or middle school) were eligible, but participants with tonal language experience (given the link between musical ability and use of a tone language, e.g., Bidelman *et al.*, 2013) or significant experience with a language other than English before age 12 years were ineligible.

Participants were asked about their musical experience, including one question from the Ollen Musical Sophistication Index (Ollen, 2006; Zhang and Schubert, 2019) that asked participants to choose a title that best described them: non-musician, music-loving non-musician, amateur musician, serious amateur musician, semiprofessional musician, or professional musician. Participants also answered items constructed by the research team that measured overall musical sophistication and self-rated pitch ability (see the Appendix). For *overall musical sophistication*, participants rated how frequently they engage in the following activities: (1) Listen to music (radio, YouTube, Pandora, Spotify, etc.), (2) Attend concerts, (3) Play/sing music, and (4) Compose music. Ratings were made on a five-point Likert scale (Rarely/Never [Yearly] to Very frequently [Daily or more than daily]). For *self-reported pitch*, participants rated themselves on the following abilities: (1) Singing in tune, (2) Noticing when someone else is out of tune, and (3) Noticing when a wrong note is played in a song. Ratings were made on a five-point Likert scale (Poor to Excellent). These latter two measures were used to verify differences between musicians and non-musicians (see Sec. III).

2. SFG task

The SFG stimuli consisted of 50-ms tones that either made up the background or figure portions of the stimulus. The background consisted of random tones across a range of frequencies. The figure portion of the stimulus consisted of tones that cohere via repeating frequencies with similar onset times (Fig. 1). The tones that formed the stimuli were selected from a six-octave range, with frequencies ranging from 100 to 6400 Hz. All stimuli consisted of tones from 30 frequency bins selected based on log-scale spacing.

All stimuli were 6000 ms in duration. Stimuli containing the target figure had a maximum 3000-ms figure duration, with a figure onset that randomly varied from 1000 to 2500 ms post stimulus onset. Across stimuli, the average figure onset was 1797 ms ($SD = 453$, *range*: 1004 to 2549 ms). The average figure duration was 2820 ms ($SD = 55$, *range*: 2714 to 2930 ms). All stimuli were normalized to 70 dB sound pressure level (SPL) using Praat (Boersma and Weenink, 2021; v.6.1.39). After normalization, a linear model suggested that there was no difference in intensity between stimuli containing a figure and distractor stimuli without a figure ($t = -0.81$, $p = 0.42$). All stimuli were presented diotically.

(a) *Coherence*. The aspect of figure coherence that was manipulated in this study was the number of tones in each figure chord (i.e., the number of tones that occurred at specific repeating frequencies with the same temporal onset). The coherence level varied from 4, 6, 8, or 10 tones. A total of 160 experimental stimuli were created, with 20 target and 20 control stimuli for each of the four levels of coherence. All participants encountered the same stimuli. Stimuli were created using MATLAB (R 2019b) at a sampling rate of 44.1 kHz. Each 6000-ms stimulus was constructed by first creating a control stimulus containing only background chords. Each background chord had a duration of 50 ms and consisted of 24 tones that were randomly selected from the 30 frequencies. Thus, each chord consisted of a different random combination of tones at different frequencies. These chords occurred at 200-ms intervals, indicating that each chord occurred five times per second. Additionally, each frequency bin was presented an average of four times per second; in other words, they occurred an average of 24 times across the entire 6000-ms stimulus (*range*: 23 to 25). Next, a corresponding target stimulus was created by adding figure tones to the background stimulus. For each tone added to the figure chord, a tone of the same frequency was removed from a different temporal location in the background, such that the control, background-only stimulus contained the same number of tones in each frequency bin as the target stimulus with the figure. Thus, half of the trials contained a target stimulus with figure chords, and half contained a control stimulus with only the background. Beyond the tone substitutions, each target and control stimulus pair had the same spectrotemporal makeup.

(b) *Background jitter*. To reduce the temporal coherence between background tones and figure tones within target stimuli, a jitter was applied to the background, such that each tone within the background was randomly assigned to have jitter from a range of zero to 150 ms sampled from a Gaussian distribution. Thus, the background was designed to lack spectrotemporal coherence (i.e., within a given time window there was

https://doi.org/10.1121/10.0016756

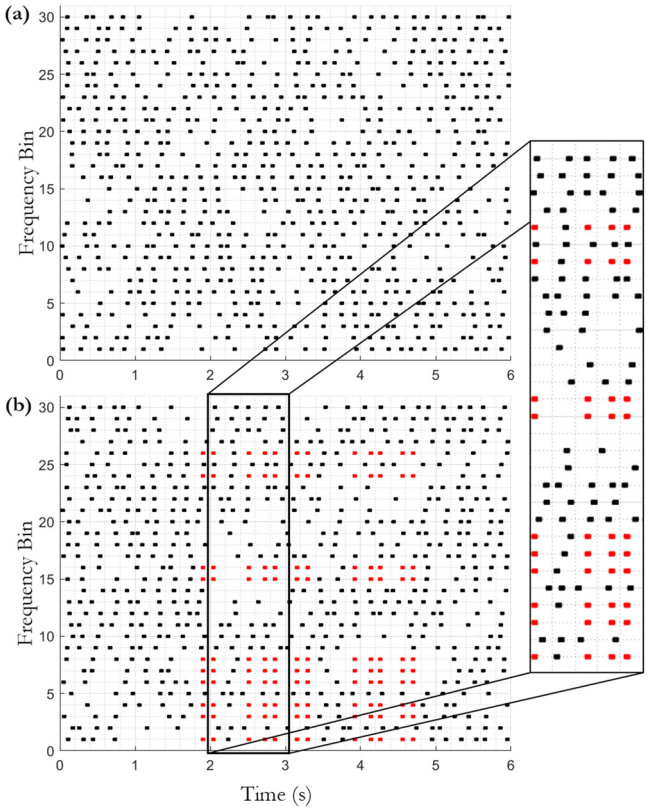

FIG. 1. (Color online) Example stimulus with a background duration of 6 s (i.e., black bars) and a figure duration of ∼3 s (i.e., red bars). Each colored bar indicates an instance in which a 50-ms tone is present. (a) Control stimulus with background tones only (black bars) and no figure. (b) The target stimulus with a figure consisting of a series of 10-tone chords repeated at an average rate of four chords per second.

a random subset of frequencies that did not share a temporal pattern).

(c) *Figure jitter*. Each figure chord repeated four times per second. For example, in Fig. 1, the 1.1–3.9 s (∼3 s total) figure portion of the stimulus contained 12 repeating figure chords, adhering to the four figure chords per second criterion. A coherent figure jitter was introduced to create random onsets of each figure chord. Each figure chord was assigned a random coherent jitter value between zero and 150 ms. While applying jitter, temporally adjacent tones in the same frequency bin were separated by at least 2.5 ms to avoid temporal overlap between figure and background tones.

As mentioned in Sec. I, our stimuli differ from those of previous SFG tasks in three key ways. First, the figure chords in the present study were presented nonconsecutively at a rate of 4 Hz with variable ICIs. Second, we employed background jitter such that the background tones were temporally uncorrelated with the figure chords. Third, background tones were evenly distributed across both time and frequencies to avoid fluctuations in intensity. These measures were taken as a means to mitigate potential confounds in prior SFG stimuli (e.g., the use of periodicity or intensity as alternative cues) and to have the stimuli more adequately reflect real-world listening conditions (a correlated foreground in an uncorrelated background).

3. QSIN task

The QSIN is a brief speech-in-noise task that measures signal-to-noise ratio (SNR) loss—defined here as the increase in SNR needed for a listener to correctly recognize 50% of the words in a sentence presented over multi-talker babble (Killion *et al.*, 2004). The task consisted of four lists of six sentences presented in varying levels of noise, with five key words each. Within each list, the first sentence was played at 25-dB SNR, with SNR decreasing by 5 dB for each successive sentence, with the final sentence presented at 0-dB SNR. Stimuli were root-mean-square (rms) normalized to 70 dB SPL in Praat (Boersma and Weenink, 2021; v.6.1.39). Scores were based on the total number of key words that participants correctly repeated. For each list, the SNR score was calculated as 25.5 minus the total correct for that list. The final score was the average across the four lists. The task took approximately 5 min to complete.

4. RSPAN

RSPAN is a complex span WMC measure and was included in the present study because it is the most commonly used measure in investigations of working memory associations with speech recognition in noise (e.g., Akeroyd, 2008; Besser *et al.*, 2013). A non-auditorily presented measure was selected in order to prevent variability in audibility from driving associations between WMC and SFG performance. The RSPAN task used in the present study is based on the RSPAN implemented by Rönnberg *et al.* (1989), which was adapted from Baddeley *et al.* (1985). In this task, 54 sentences were visually presented. Sentences were presented one word at a time, 800 ms per word. Half of the sentences were semantically anomalous, and half were not. Participants were instructed to read each sentence aloud and to indicate whether the sentence made sense after reading the last word. Sentences were presented in set sizes of three to six, with three of each set size, and at the end of each set the last word of each sentence was verbally recalled. RSPAN scores were calculated as the percentage of correctly recalled sentence-final words. The task took approximately 15 min to complete.

C. Procedure

All stimuli were presented to participants via headphones at a comfortable audio level. Participants first completed the demographics and language history questionnaire. Next, because participants completed the experiment with their own devices and headphones, the Online Hearing Test and Audiogram was administered as a basic hearing and sound check to identify any participants with either grossly abnormal hearing or problematic sound presentation hardware (e.g., sound card or headphones) across tested frequencies. In the calibration portion of the task, participants heard a calibrated audio file of the sound of two hands rubbing together (Torres-Russotto *et al.*, 2009). Participants were

instructed to rub their own hands together, close to their nose, and asked to adjust the volume level of their computer to try to match the volume of the calibration audio file to the sound of their hands. After participants completed the calibration portion, they completed the full Online Hearing Test and Audiogram, which included one-third octave-band warble tones from 2.5 to 8 kHz. Intensity had been calibrated to range from –5 to 80 dB hearing level (HL) relative to the calibration file. From –5 to 20 dB HL, intensity increased in increments of 5 dB HL. From 30 to 80 dB HL, intensity increased in increments of 10 dB HL. A researcher asked participants to indicate when they could hear a tone, by raising their hand. Starting with a 1-kHz tone at 60 dB HL, participants were asked to complete the task with their eyes closed so that they could not see the sound level of the tones they heard. The experimenter decreased the intensity level by two levels (e.g., from 60 to 40 dB HL) when participants reported hearing a tone and increased by one when participants did not report hearing a tone (e.g., from 40 to 50 dB HL) until the lowest threshold was reported. This procedure continued for 2, 4, 8, 0.25, and 0.5 kHz, in that order. Although the full range of tones was presented, only 0.25 – 4 kHz were used to exclude participants with abnormal values (above 40 dB HL). No participants were excluded due to abnormal values.

After the hearing screening task, participants completed the SFG task. To ensure participants understood the task and what constituted a figure, they listened to several example stimuli and completed practice trials with feedback before beginning the experimental portion of the task. First, to ensure participants knew what a figure should sound like, they were provided with example stimuli with fewer overall tones than the practice and experimental stimuli (Fig. 2). The first example stimulus contained 12 background tones only. The next example was the corresponding target stimulus, with 12 tones in the figure, meaning that no background tones played during the figure chords. As the figure chord was played, visual cues indicated when the figure was present; "Figure present" appeared on the screen as soon as the first figure chord was presented, and a green rectangle was displayed around the text during each figure chord and disappeared during the intervals in which the figure chord was not present.

Next, participants were presented with four sets of example stimuli that increased in difficulty, with three stimuli in each set. For set one, in the first of three stimuli, participants were presented with a target stimulus with 14 tones overall and were asked if they could detect the figure. For the second stimulus in this set, participants heard the target again with the corresponding visual cues while each figure chord played auditorily. Finally, the corresponding control stimulus was played, without the figure. This sequence was repeated for triads of stimuli with 12, 10, and 8 tones each in the figure.

After listening to all example SFG stimuli, participants then completed three blocks of the practice task. Participants were instructed to press a button if they heard a "figure," or a group of tones that repeated together, and to do so as soon as they detected the target figure. Responses were scored as correct if the participant accurately identified the presence of a target figure within the time window starting 120 ms after the onset of the initial figure through the end of the trial (Schröter *et al.*, 2007). If participants pressed the button multiple times during the trial, only the first button press was used. Each practice trial consisted of a 6000-ms stimulus with a fixation cross, followed by 1000-ms feedback that either displayed "Correct!" or "Incorrect." The next trial began after the stimulus sound offset. Regardless of whether participants responded to detecting the figure, the entire stimulus was presented. Practice trials were presented in pairs, and participants were informed of this; however, the order of the pairs, and the order of the sounds within each pair (target-control or control-target) was re-randomized at each repeat of the block. The first block contained 12 tones in each figure chord, with 30 total tones in the overall stimulus. After six correct responses, participants completed the second practice block, in which the number of tones in the figure decreased to 10. After six consecutive correct responses, the coherence level varied randomly between eight, six, and four tones in a figure. Participants completed 12 trials at this level, half target stimuli and half control stimuli, with feedback, with no accuracy criterion.

Stimuli for the experimental trials were chosen such that figure onsets and durations were balanced across coherence levels (four, six, eight, and ten tones). Experimental trials were presented in four blocks, with the different

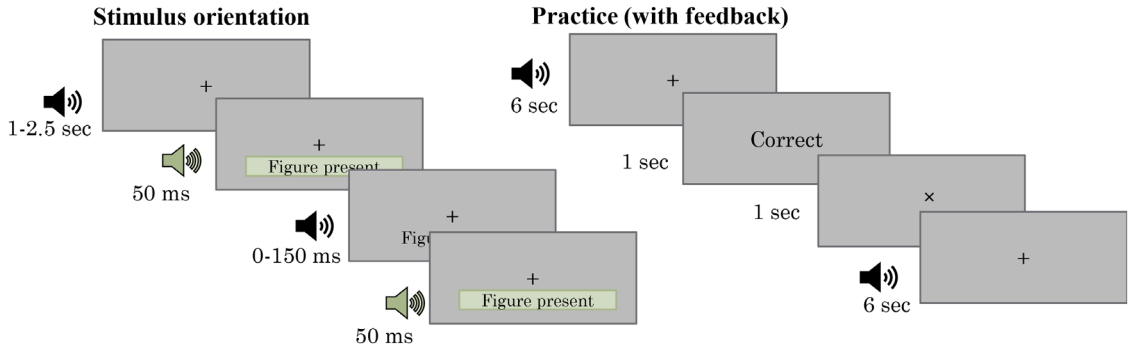

FIG. 2. (Color online) Example screen of SFG stimulus orientation and practice trials.

https://doi.org/10.1121/10.0016756

coherence levels intermixed in each block. Each block contained 40 trials: 20 control trials and 20 target figure trials. Stimuli were also assigned to blocks to balance figure onsets and durations across blocks. Participants were offered a short break between each block. Trials within blocks were pseudorandomized, such that no more than three of one trial type (control/target) or coherence level (e.g., ten tones in a figure) occurred in a row, and such that members of each stimulus pair (control and corresponding target) were not presented within the same block, or in consecutive trials across blocks. No feedback was provided during the experimental trials and trials were separated by an intertrial interval (ITI) of 1000 ms. To clearly separate trials, an "x" was displayed during the ITI. Once each trial began, the "x" changed to a fixation cross (+) that remained on the screen for the duration of the 6000-ms stimulus. The dependent variables are accuracy and RT for each stimulus. Only responses that registered after stimulus onset but before onset of the next stimulus were recorded.

After completing the SFG task, participants completed the QSIN task. Participants were instructed to listen to the sound of a woman talking with other speakers in the background. Participants were informed that the woman's voice will be easy to detect at first but will become more difficult as the task progresses when the background speakers become louder. The task was to repeat each sentence that the woman spoke. For the current study, participants completed a practice list and four experimental lists (QSIN lists 1–4).

Finally, participants completed the RSPAN task. Participants were asked to read each sentence aloud and judge whether the sentence made sense. Set sizes increased as the task progressed. At the end of every set, participants were asked to recall the last word of each sentence and were given a 2-min time limit to recall the words. After participants finished recalling the words, they pressed a key on the keyboard to move to the next set, a fixation cross was presented for one second, and the next set began. Participants were encouraged to recall the words in order of sentence presentation; however, RSPAN scores were calculated as the number of correctly recalled words regardless of recall order (Rönnberg *et al.*, 1989).

A linear model predicting QSIN scores by musicianship while controlling for WMC did not find any differences between musicians and non-musicians on the QSIN ($p = 0.67$). A second model predicting WMC by musicianship while controlling for QSIN scores likewise did not find any differences between musicians and non-musicians on the RSPAN task ($p = 0.54$).

III. ANALYSES

All data were analyzed in R (R Core Team, 2021; v.4.1.0). Two dependent variables were analyzed: (1) accuracy, based on whether participants either correctly indicated that they heard the figure or correctly withheld a response when no figure was present; and (2) RT (for target trials only) measured in milliseconds, adjusted from the onset of the first figure chord until the participant made a button press and subsequently log-normalized. For RT, trial exclusions included incorrect responses ($N = 561/2560$, 21.9%) and trials in which the participant indicated that they heard the figure before the onset of the initial figure chord or within 120 ms of the onset of the initial figure chord (Schröter *et al.*, 2007; $N = 56/2560$, 2.1%). Given that there were pre-determined lower and upper bounds for RT, outlier detection was not performed. Each dependent variable was predicted by the figure coherence level (i.e., the number of tones in a figure chord, henceforth referred to simply as Coherence), as well as three other participant-level covariates of (1) QSIN performance, (2) RSPAN performance, and (3) musicianship.

RSPAN score was calculated as the z-scored number of sentence-final words correctly recalled in the RSPAN task (henceforth, RSPAN). QSIN scores were based on the average SNR loss scores across the four trials in the QSIN task (henceforth, QSIN). QSIN scores were not z-scored given that the reference level (0-dB SNR loss) is meaningful.

Musicianship was based on participants' self-identification: non-musicians or music-loving non-musicians were grouped together as non-musicians ($N = 16$ participants) and all other categories (amateur musician, serious amateur musician, semiprofessional musician, or professional musician) were grouped together as musicians ($N = 16$ participants). Musicians rated themselves as having significantly higher self-reported musical sophistication ($M = 2.81/5$, $SD = 0.65$) than non-musicians ($M = 2.36/5$, $SD = 0.35$; two-sample t-test: $t = 2.45$, $p = 0.02$) as well as significantly higher self-reported pitch ability ($M = 3.85/5$, $SD = 0.63$) than non-musicians ($M = 3.02/5$, $SD = 1.04$; two-sample t-test: $t = 2.73$, $p = 0.01$). See supplementary material for additional information on participants' musical experience and instruments played.[1]

A linear mixed-effects model with a Gaussian family was used to analyze log-transformed RT, and a binomial logistic mixed-effects model was used to analyze accuracy. Linear/logistic mixed-effects regression (LMER) has many advantages over traditional analyses using analysis of variance (ANOVA), primarily because it allows for the inclusion of both by-participant and by-item random intercepts (rather than performing separate analyses on aggregated participant and item data) and for the specification of random slopes which allow the by-participant and by-item random intercepts to vary (e.g., across levels of a factorized condition or by a continuous variable).

All models were created using the buildmer function in the buildmer package (Voeten, 2021; v.1.9). All models were originally specified with all fixed effects and interactions and fully specified random intercepts and random slopes and used the "bobyqa" optimizer with the default number of iterations. The buildmer function was then used to simplify the random effects, given that (1) overly complicated random effects often prevent model convergence and thus need to be simplified, and (2) random effects should be specified based only on what the data can support (Matuschek *et al.*, 2017). In addition,

fixed effects that did not significantly contribute to model fit, tested via backwards elimination with likelihood ratio tests, were removed from the model. This was beneficial here because it limited the number of complex interactions, thus simplifying interpretation of the final model. The maximal models submitted to buildmer predicted the dependent variable by Coherence, Musicianship, RSPAN, and QSIN as well as the pertinent random effects for each variable. The two, three-way interactions between Coherence, RSPAN, and Musicianship and Coherence, QSIN, and Musicianship were specified along with all lower-order interactions. The four-way interaction between Coherence, QSIN, RSPAN, and Musicianship was not included as there were not *a priori* theoretical predictions of the interaction between RSPAN and QSIN. Coherence and Musicianship were both dummy-coded; each level of Coherence (four, six, eight, and ten) served as a reference level in re-leveled models to ensure all adequate comparisons were made. When there was a significant interaction with Musicianship, the reference level was changed to non-musicians to fully investigate the interaction. Last, random effects by participant and by item were included with the random slopes initially maximally specified.

The final models selected by buildmer, including the final random effects specification, are presented at the top of Tables II (accuracy) and III (RT). The buildmer function provides *p*-values for all fixed effects calculated by the lmerTest package (Kuznetsova *et al.*, 2017; v.3.1–3), which uses *t*-tests that employ the Satterthwaite method (Satterthwaite, 1941) to approximate degrees of freedom. These *p*-values indicate differences between a given level in the model and the reference level; for example, when comparing a figure coherence level of 6 to a level of 4. When visualizing the fixed effects, model-predicted values were obtained using the ggemmeans function in the ggeffects package (Lüdecke, 2018; v.1.1.1), providing estimated marginal means as calculated by the emmeans function in the emmeans package (Russell, 2021; v.1.7.1–1).

IV. RESULTS

Table I provides the means and standard deviations across the four coherence levels for: (1) adjusted RT (relative to the onset of the first figure), (2) accuracy, (3) the number of figure chords heard before a correct response was given, (4) the number of hits, (5) the number of misses, (6) the number of false alarms, and (7) the number of correct rejections.

A. Figure detection accuracy

Model estimates for accuracy[2] are presented in Fig. 3, which shows that performance improves as the coherence level increases. The final best-fitting model predicted accuracy by the interaction between Coherence and RSPAN plus the interaction between Musicianship and QSIN (Table II). The model summaries revealed that the likelihood of accuracy was greater for 6 than 4 tones in a figure ($b = 0.15$, $t = 8.54$, $p < 0.001$) and for 8 than 6 tones in a figure ($b = 0.08$, $t = 4.84$, $p < 0.001$). Likelihood of accuracy did not significantly differ between 8 and 10 tones in a figure ($b = -0.01$, $t = -0.59$, $p = 0.55$; see Fig. 3). For the interaction between Coherence and RSPAN, the model summaries suggested that while there was no effect of RSPAN on the likelihood of accuracy for 4 ($b = 0.02$, $t = 1.07$, $p = 0.29$), 6 ($b = 0.02$, $t = 0.97$, $p = 0.34$), and 8 ($b = -0.02$, $t = -0.96$, $p = 0.34$) tones in a figure, there was a significant effect for 10 tones in a figure ($b = -0.04$, $t = -2.43$, $p = 0.02$), such that increasing RSPAN scores (better performance) were associated with decreasing likelihood of accuracy (Fig. 4). For the interaction between Musicianship and QSIN, the model summaries suggested that there was no effect of QSIN on likelihood of accuracy for non-musicians ($b = 0.01$, $t = 0.82$, $p = 0.42$), but there was a significant effect for musicians ($b = 0.02$, $t = -2.77$, $p = 0.01$), such that increasing QSIN scores (worse performance) were associated with decreasing likelihood of accuracy (Fig. 5).

TABLE I. Descriptive statistics of performance on SFG task. Standard deviations are given in parentheses.

	4 tones	6 tones	8 tones	10 tones
Mean adjusted RT (ms)	2033.91 (1043.18)	1680.76 (846.22)	1298.63 (742.44)	1190.86 (613.85)
Mean accuracy (%)	57.03 (49.52)	71.64 (45.09)	79.92 (40.07)	78.91 (40.81)
Mean no. of figures to RT	7.87 (3.16)	6.66 (2.97)	5.45 (2.71)	4.99 (2.44)
Mean no. of hits	9.34 (3.17)	15.16 (2.36)	17.91 (1.84)	18.31 (1.64)
Mean no. of misses	10.66 (3.17)	4.84 (2.36)	2.09 (1.84)	1.69 (1.64)
Mean no. of false alarms	6.53 (4.23)	6.50 (4.14)	5.94 (4.25)	6.75 (3.70)
Mean no. of correct rejections	13.47 (4.23)	13.50 (4.14)	14.06 (4.25)	13.25 (3.70)

1. Post hoc WMC and accuracy analysis

The finding that better WMC, as measured by the RSPAN task, predicted worse SFG performance for accuracy in the easiest condition (ten tones in a figure) was unexpected. One explanation (see Sec. III) could be that individuals with better WMC were more susceptible to interference from auditory memory traces formed when perceiving figures in previous trials. If this were the case, then it might be expected that this interference accumulates as the task progresses, such that its effects are stronger towards the end of the task. To examine this, a *post hoc* analysis was performed on the accuracy data for ten tones in a figure across the four blocks (1, 2, 3, and 4) of the SFG task. A binomial LMER was performed, using the same procedures outlined in Sec. III. The model predicted accuracy by the interaction between Block and RSPAN, with the inclusion of by-participant and by-item random intercepts. Each block served as the reference level in re-leveled models to ensure that all adequate comparisons were made. Because this was a *post hoc* analysis, the Holm correction for multiple comparisons was applied to all terms (Holm, 1979). The models

https://doi.org/10.1121/10.0016756

TABLE II. Accuracy model summary. Four figure tones and musicians as reference levels.

Formula: Accuracy ~ Coherence * RSPAN + Musicianship * QSIN + (1|Participant)					
		$n = 5120$			
Fixed effects	*b*	Std. error	*df*	*t*	*p*
Intercept	0.59	0.02	57.9	28.78	< 0.001
6 figure tones	0.15	0.02	5088	8.54	< 0.001
8 figure tones	0.23	0.02	5088	13.39	< 0.001
10 figure tones	0.22	0.02	5088	12.79	< 0.001
RSPAN	0.02	0.02	86.1	1.07	0.29
QSIN	–0.04	0.02	32.0	–2.77	< 0.01
Musicianship	–0.05	0.03	32.0	–1.84	0.08
6 figure tones by RSPAN	–0.002	0.02	5088	–0.10	0.91
8 figure tones by RSPAN	–0.03	0.02	5088	–1.99	0.05
10 figure tones by RSPAN	–0.06	0.02	5088	–3.43	< 0.001
Musicianship by QSIN	0.06	0.02	32.0	2.52	0.02
Random effects		Variance		Std. dev.	
Participant		0.004		0.06	

suggested that, while there were no significant effects of RSPAN on Accuracy for blocks 1 (corrected $p = 0.37$), 2 (corrected $p = 0.37$), or 3 (corrected $p = 0.06$), there was a significant effect of RSPAN on Accuracy in block 4 (corrected $p < 0.01$). It was only in this final block that better performance on the RSPAN task was significantly negatively associated with Accuracy, suggesting that the effect was strongest towards the end of the task.

B. Figure detection RT

The final best fitting model predicted log RT by the interaction between Coherence and QSIN. The model summaries revealed that log RTs were not significantly different between 6 and 4 tones in a figure ($b = –0.14$, $t = –1.79$, $p = 0.08$; Table III) and significantly faster for 8 than 6 tones in a figure ($b = –0.27$, $t = –3.74$, $p < 0.001$). Log RTs did not significantly differ between 8 and 10 tones in a figure ($b = –0.08$, $t = –1.15$, $p = 0.26$; see Fig. 6). For the interaction between Coherence and QSIN, the model summaries suggested that there were no effects of QSIN on log RTs for 4 ($b = 0.002$, $t = 0.06$, $p = 0.95$) or 6 ($b = 0.05$, $t = 1.45$, $p = 0.14$) tones in a figure, there were significant effects for 8 ($b = 0.09$, $t = 2.68$, $p = 0.01$) and 10 ($b = 0.07$, $t = 2.12$, $p = 0.04$) tones in a figure. For both, increasing QSIN scores (worse performance) were associated with slower log RTs (Fig. 7).

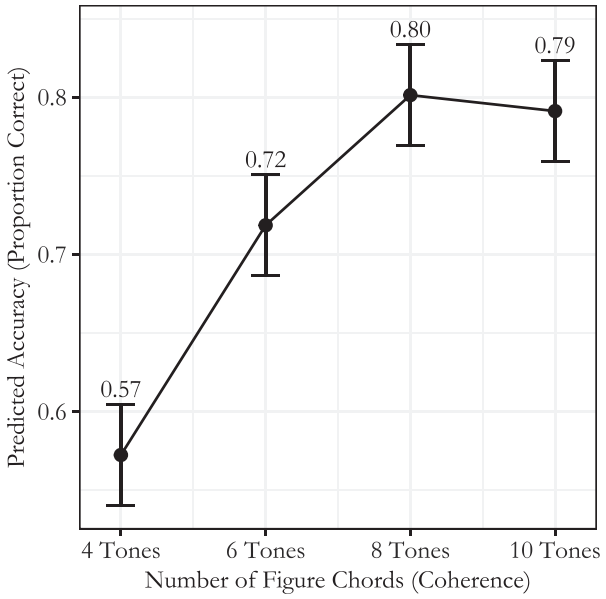

FIG. 3. Predicted mean likelihood of accuracy by the number of tones in a figure. Error bars represent the 95% confidence interval.

V. DISCUSSION

The goals of the present study were to assess the effect of temporal coherence on auditory stream segregation, indicated by performance on an SFG task involving non-consecutive figure chords, measure its relationship to speech-in-noise recognition, and examine potential individual differences that may modulate the relationship between auditory stream segregation and speech-in-noise recognition. In the current study's paradigm manipulating temporal coherence—which included non-consecutive figure chords with four, six, eight, or ten tones in each figure chord—both accuracy and speed of figure detection associated with auditory stream segregation were measured. Overall, findings demonstrated that temporal coherence and individual differences in speech-in-noise recognition and musicianship influenced the accuracy and speed of figure detection.

A. Auditory stream segregation in the discontinuous SFG task

Aligning with previous studies that have investigated auditory stream segregation with SFG tasks, the results of the present study demonstrated that increasing coherence, manipulated by increasing the number of tones per figure

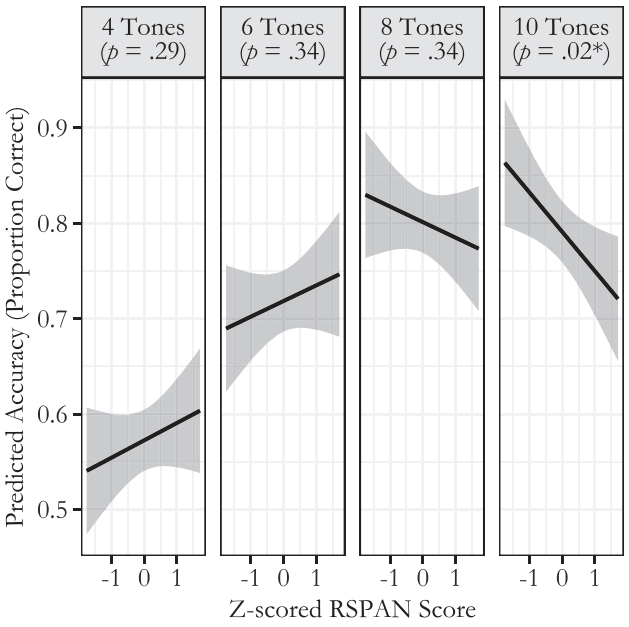

FIG. 4. Interaction between RSPAN and Coherence level on likelihood of accuracy. Shaded regions represent the 95% confidence interval.

chord, resulted in improved performance on the SFG task (Fig. 3). Accuracy increased as the number of tones in a figure chord increased, reaching asymptote around eight tones in each figure chord. These findings align with other research on auditory stream segregation using detection of consecutive figure chords (O'Sullivan *et al.*, 2015; Teki *et al.*, 2011, 2016) and detection of figure gaps (Holmes *et al.*, 2021; Holmes and Griffiths, 2019). Increased temporal coherence (i.e., the number of synchronized tones) resulted in better auditory stream segregation. The extent to which the current SFG performance effects are weaker than in previous studies (e.g., Teki *et al.*, 2013) may depend

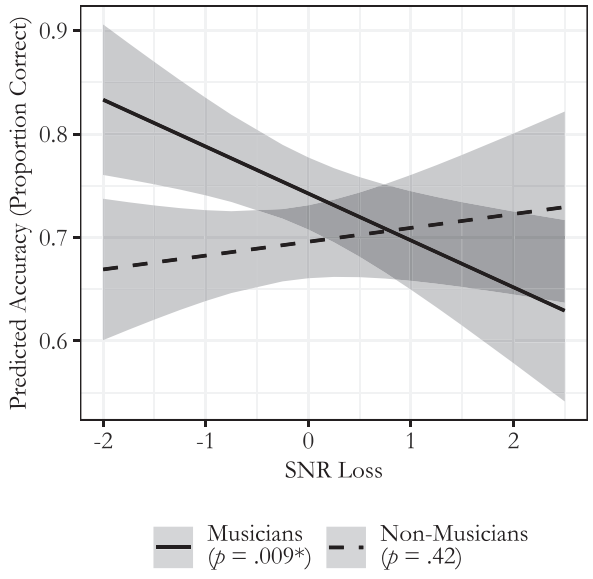

FIG. 5. Interaction between Mean QSIN SNR Loss and Musicianship on likelihood of accuracy. Shaded regions represent the 95% confidence interval.

partly on the alterations made to the stimuli to be spectrally and temporally flat, reducing the availability of these additional cues. However, differences may have also arisen as a result of having presented the experiment remotely, which limited the extent of control over the acoustic environment that was used for testing. Nonetheless, the changes employed in the present SFG task—namely, background jitter and non-consecutive figure chords with variable ICIs—both resulted in findings that are congruent with previous studies using SFG task while using stimuli that more accurately reflect real-word listening conditions.

Auditory stream segregation for SFG stimuli may involve bottom-up (Molloy *et al.*, 2019; O'Sullivan *et al.*, 2015; Teki *et al.*, 2011; Teki *et al.*, 2016) and top-down processes (O'Sullivan *et al.*, 2015). Because the current study's SFG task required active figure detection, following the two-step temporal coherence model (Shamma *et al.*, 2011), auditory stream segregation of the figures from background likely involved participants first identifying frequency features of the stimuli tones, then segregating the figure chords from the background tones based on the temporal coherence of the figure chords. This positions the current SFG task as relying more on low-level features, in contrast to speech-in-noise tasks that depend more on high-level (e.g., lexical and semantic) properties (Mattys *et al.*, 2012).

Similar to the accuracy findings, increased temporal coherence (i.e., number of tones in each figure chord) resulted in faster RTs for figure detection, reflective of faster auditory stream segregation (Fig. 6). Similar to the accuracy results, RTs reached an asymptote around eight tones in each figure chord, with no difference between eight and ten tones in a figure. These findings demonstrate that temporal coherence can drive how quickly listeners can segregate the figure from the background, providing a rough measure of how much information (i.e., number of coherent figure chords) is needed before behavioral effects of stream segregation can be observed. Similar results were found in other studies that directly manipulated the number of consecutive

TABLE III. Log RT model summary. Four figure tones as reference level.

Formula: Log RT ~ Coherence * QSIN + (1|Participant) + (1|Item)					
	$n = 1943$				
Fixed effects	*b*	Std. error	*df*	*t*	*p*
Intercept	7.47	0.06	116.31	116.57	< 0.001
6 figure tones	–0.14	0.08	82.65	–1.79	0.08
8 figure tones	–0.41	0.08	80.37	–5.42	< 0.001
10 figure tones	–0.49	0.08	80.14	–6.51	< 0.001
QSIN	< 0.01	0.04	85.27	0.06	0.95
6 figure tones by QSIN	0.05	0.03	1853.75	1.48	0.14
8 figure tones by QSIN	0.09	0.03	1854.33	2.79	0.01
10 figure tones by QSIN	0.07	0.03	1856.24	2.21	0.03
Random effects	Variance		Std. dev.		
Participant	0.03		0.18		
Item	0.05		0.21		

https://doi.org/10.1121/10.0016756

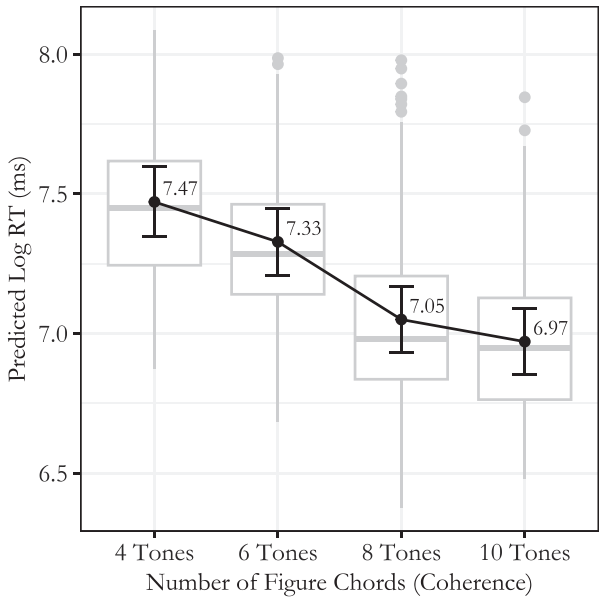

FIG. 6. Predicted Log Adjusted RT by the number of tones in a figure chord with predicted means. Error bars represent the 95% confidence interval; boxplots represent distribution of raw data.

figure chords (i.e., figure duration); participants were more accurate in detecting figures with more chords than figures with fewer chords (Teki *et al.*, 2011; Teki *et al.*, 2013; Teki *et al.*, 2016). Teki *et al.* (2013) showed that increased number of figure chords (figure duration) and temporal coherence (number of tones per figure chord) resulted in increased accuracy, but chord duration did not influence accuracy. Thus, the build-up of information across incoming temporally coherent chords facilitates auditory stream segregation necessary for figure detection.

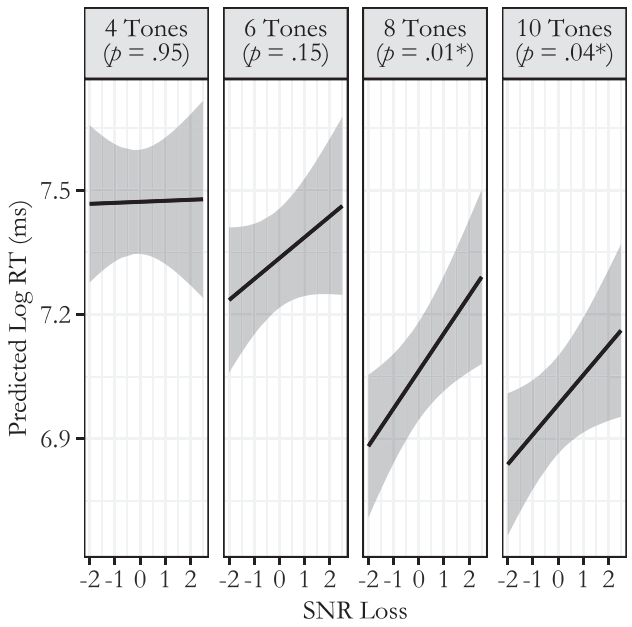

FIG. 7. Interaction between Mean QSIN SNR Loss and Coherence level on predicted log adjusted RT. Shaded regions represent the 95% confidence interval.

In designing the SFG stimuli in the current study to be spectrally and temporally flat, jitter was applied to the background tones so that the onset of the figure would not result in an intensity cue. While the consistency of the timing of the figure tone onsets is important for establishing temporal coherence over time, it could in theory provide a cue for the onset of the figure. In other words, it is possible that a temporal coherence mechanism was not needed to successfully identify the figure; instead, participants simply indicated when they heard any group of tones with perceptually simultaneous onsets. However, there are several reasons to believe that participants did not use this as a cue to perform the SFG task. One reason is that previous studies have shown that a build-up of coherent tones across time (rather than a single chord presentation) is important for figure detection in SFG stimuli (e.g., O'Sullivan *et al.*, 2015) and that longer figure durations result in improved detection (e.g., Teki *et al.*, 2011). In the present study, the mean number of figure chords that it took to correctly identify the presence of the figure was 6.24 across all four conditions (Table I). Even in the 10-tones condition with the highest accuracy/fastest RTs, it took an average of five repetitions for participants to respond. Out of a total of 1943 trials, only 29 comprised a correct response after the first chord of the figure (nine such responses in the 4-tones condition, seven in the 6-tones condition, six in the 8-tones condition, and seven in the 10-tones condition). Although coherent onsets could be a reliable cue, participants did not appear to make effective use of this information in their responses.

Finally, the literature on perceptual simultaneity judgements have reported the likelihood of judging two asynchronous pure tones of different frequencies to be synchronous is greater than 75% when onset asynchronies are approximately 25 ms or longer (Parker, 1988), depending on the relative frequency of the tones (Okazaki and Ichikawa, 2017). Thus, even with background jitter in the current study (0 to 150 ms), participants would likely have perceived coherent tones in the background by chance. Responses to these apparent chords would have led to false alarms, with potentially more in the 4-chord condition, in which there was the most opportunity for background tones to cohere by chance. However, false alarm rates were low and there were no systematic differences across conditions (Table I). Given that (1) in our own data, several repetitions of the figure chord were required before a correct response was made, and (2) participants did not appear to respond only when multiple tones co-occurred with similar onsets, the best explanation for the pattern of results in the present study is that of a temporal coherence mechanism. In other words, participants required multiple repetitions of coherent tones across time in order to successfully segregate the figure from the background.

B. Auditory stream segregation and speech-in-noise recognition

At higher coherence levels in the SFG task (i.e., eight and ten tones in a figure), faster RTs were associated with better performance on the QSIN (Fig. 7), indicating that a

participant's ability to recognize speech in noise was related to how quickly they could segregate auditory streams. In noisy environments, listeners must quickly segregate the target speaker from background noise to understand what the speaker is saying. Slower auditory stream segregation would likely result in missed words or phrases while listening to speech. At lower coherence levels (and hence increased task difficulty), this relationship between SFG task RTs and speech-in-noise recognition (i.e., QSIN) diminishes as RTs become less sensitive to temporal coherence (Fig. 7). The current results therefore align with findings from Holmes and Griffiths (2019), who also found that the relationship between performance on SFG tasks and speech-in-noise perception is diminished at increased levels of task difficulty (i.e., figure frequencies changed at differing rates).

Although QSIN scores and SFG task RTs were related in the present study, there are fundamental noteworthy differences between the properties and demands of each of these two tasks. On a stimulus level, temporal coherence is typically manipulated in SFG tasks with no direct manipulation of SNR (Molloy *et al.*, 2019; Teki *et al.*, 2011; Teki *et al.*, 2016), whereas the SNR for target vs background speakers is the main manipulation in speech-in-noise tasks (e.g., Alain *et al.*, 2018; Killion *et al.*, 2004). More importantly, whereas language processing is absent for SFG tasks, it is vital for speech-in-noise recognition. For speech-in-noise tasks that involve recalling or identifying spoken words embedded in sentences, in addition to low-level acoustic perception processes, listeners engage in higher-level comprehension processes that include lexical access (Carroll *et al.*, 2016) and to some extent syntactic knowledge (Kidd *et al.*, 2014). For instance, Coulter *et al.* (2021) found that higher contextual constraint had a larger positive effect on speech recognition in noise compared to quiet conditions. Given the differences in task demands on the QSIN and SFG tasks, the fact that a relationship was observed between the two suggests that low-level auditory stream segregation drove the positive relationship between SFG task RTs and QSIN scores; better QSIN scores were associated with faster RTs. Importantly, this finding further suggests that the SFG task is a potentially viable measure of language-independent auditory stream segregation processes.

C. Individual differences and auditory stream segregation

1. WMC and the highest SFG coherence level

Individuals with better WMC were predicted to perform better on the SFG task. Interestingly, for the highest coherence level (ten tones in a figure), individuals with higher WMC showed decreased accuracy (Fig. 4), which was most pronounced at the end of the task. Several explanations may account for this finding.

Although this study demonstrates potential negative effects of higher WMC on performance on the SFG task under easy conditions, findings on the relationship between WMC and speech-in-noise perception in younger normal-hearing adults have been mixed (Füllgrabe and Rosen, 2016a,b). In one meta-analysis, Füllgrabe and Rosen (2016b) demonstrated that correlations between speech-in-noise perception and RSPAN ranged from negative to positive (–0.29 to 0.64). Positive correlations were more likely to be observed in older adults, for whom working-memory-related processes may compensate for degraded speech representations (Füllgrabe and Rosen, 2016b). Furthermore, Carroll *et al.* (2016) suggest that WMC may mediate the effect of vocabulary breadth and speed of lexical access on speech-in-noise recognition. In the present study, rather than linking WMC to speech-in-noise recognition—which relies on high-level processes associated with lexical access from long-term memory—its link to figure detection in the SFG was examined instead. The reliance on lower-level and bottom-up processes for maintaining the representation of unfamiliar figure chords may affect how WMC relates to figure detection. Moderate increases in working memory task demands have been linked to reduced distractibility (SanMiguel *et al.*, 2008). However, high WMC individuals may need more extreme increases in task demands to see this reduction. Thus, the SFG task in the present study may not have been demanding enough for participants with higher WMC. Perhaps because stimuli with the highest coherence level provide the least chance for improvement, participants who became more distracted or less engaged showed the greatest reduction in performance on this level.

An alternative explanation for the decreased accuracy for participants with higher WMC could be a result of interference from auditory memory traces formed when perceiving figure chords in previous trials. Because easily perceived and highly salient stimulus features are perceived more quickly (Pooresmaeili *et al.*, 2014), recalled more accurately (Fine and Minnery, 2009), and leave longer memory traces (Thiel *et al.*, 2016) compared to less salient stimulus features, it is possible that high-WMC individuals received more interference from these memory traces. Support for this interference account comes from a WMC training paradigm that used a visual n-back task, in which participants indicated whether a current stimulus matched a stimulus presented n trials previously (Harbinson *et al.*, 2011). Within a training session, performance on the n-back task waned as participants encountered more trials, suggesting that earlier stimuli had active memory traces that interfered with task performance (Harbinson *et al.*, 2011). In the present study, high-WMC individuals—who are better able to maintain more memory activations (Thiel *et al.*, 2016)—may have had more interference because the earlier, more salient ten tones-in-a-figure stimuli interfered with processing the current stimulus. Because these auditory stimuli in the lower coherence conditions were less salient, they likely did not cause the same level of interference. Furthermore, interference may have been most easily detected in the easiest condition where there were more opportunities to observe decreases in performance.

The fact that only the final block of the SFG task had a negative relationship to WMC (as measured by RSPAN)

suggests: (1) a shift in auditory perception throughout the task, which has been observed in SFG tasks examining auditory perceptual learning (e.g., Agus and Pressnitzer, 2021), (2) a build-up of interference as more stimuli were encountered, or (3) some combination of the two. To better understand how individual differences might modulate the effect of temporal coherence on auditory stream segregation, more research is needed on other factors likely to influence performance, such as motivation and attention. Furthermore, because this study's sample consisted of younger adults with average WMC scores (*range*: 19 – 41, $M = 30.16$, maximum possible score = 54)), administering this SFG task to older adults, who tend to show declines in WMC (Bopp and Verhaeghen, 2005; Wang *et al.*, 2011), may provide a clearer understanding of factors that influence auditory stream segregation.

2. Musicianship

Regarding accuracy of figure detection (i.e., successful auditory stream segregation), musicians showed a positive relationship between accuracy on the SFG task and speech-in-noise recognition, whereas non-musicians did not (Fig. 5). This suggests that self-identified musicians who have good speech-in-noise recognition also show enhanced auditory stream segregation ability, complementing previous findings that musicians have better auditory stream segregation (Marozeau *et al.*, 2010; Zendel and Alain, 2009) and temporal auditory acuity (Parbery-Clark *et al.*, 2011; Rammsayer and Altenmüller, 2006) than non-musicians. Perhaps musicians, compared to non-musicians, relied more on auditory stream segregation processes to complete the QSIN task in the present study, whereas non-musicians may have relied more heavily on other cues, such as syntactic and semantic context, which is relevant because the QSIN sentences do contain some linguistic context (Wilson *et al.*, 2007). This explanation is consistent with previous research that has suggested that individuals with better musical ability may better perceive subtler auditory cues (Parbery-Clark *et al.*, 2009). In other words, musicians may make use of lower-level acoustic information (Zendel *et al.*, 2015), while non-musicians may rely more on higher-level linguistic contextual information to identify words. As previously mentioned, given that low-level auditory stream segregation may have driven the relationship between the SFG task RTs and QSIN scores, this would also explain why musicians—but not non-musicians—showed a positive relationship between SFG task accuracy and the QSIN. Notably, the present study did not specifically recruit expert or novice musicians, or determine musicianship based on experience, but instead categorized participants as musicians or non-musicians based on self-reports. Finding effects of self-reported musicianship in a typical sample of younger adults shows that the SFG task has promise for further exploring individual variation in auditory stream segregation for individuals within a mid-range of musicianship. Furthermore, differences in musicianship suggest that SFG performance may be modifiable, at least with extensive experience or training. Future work may benefit from exploring the extent to which SFG-based training interventions would lead to better speech-in-noise recognition more broadly.

Because a variety of factors relate to speech-in-noise task performance (e.g., musical ability, cognitive ability, language ability; Bidelman and Yoo, 2020; Yoo and Bidelman, 2019), the degree to which QSIN task performance reflects auditory stream segregation is likely to depend on how individuals differentially apply these skills to complete the task. Group differences based on age (Helfer and Jesse, 2015; McAuliffe *et al.*, 2013; Pichora-Fuller, 2008), language background (Skoe and Karayanidi, 2019), and musicianship (Kaplan *et al.*, 2021; Zendel *et al.*, 2015; Zendel and Alain, 2012), have been found to influence the strategies and skills individuals use to complete speech-in-noise tasks, which is often also dependent on task difficulty. For example, Zendel *et al.* (2015) found that, unlike musicians, non-musicians showed increased N400 event-related potential responses (a marker of lexical-semantic access) as the SNR for words in noise decreased, suggesting that they relied more on lexical processing to overcome difficult listening situations. Higher vocabulary knowledge is associated with better speech recognition in difficult listening environments (Bernard *et al.*, 2014; Carroll *et al.*, 2016; McAuliffe *et al.*, 2013), further exemplifying how linguistic ability might influence speech-in-noise task performance. In the present study, the SFG task is more attuned to auditory stream segregation processes without the interference of linguistic factors, although cognitive factors such as WMC and attention may also contribute to task performance.

D. Limitations and future directions

Although stimuli were normalized to 70 dB SPL across the figure and background portions of each stimulus, stimuli containing figures consisted of tones shifted to have the same onset time, which resulted in reduced background tone density when the figure chord was played. Thus, figure detection could have been driven by a combination of temporal coherence and fewer overlapping background tones. However, because the study findings align with results from other SFG tasks in which the number of frequency components was identical in each figure and background chord (e.g., Teki *et al.*, 2011; Teki *et al.*, 2016), it is more likely that figure detection was driven by temporal coherence of the figure chords. Additionally, whereas other studies used consecutive chords with interchord intervals of 0 ms (e.g., Holmes *et al.*, 2021; Holmes and Griffiths, 2019; Molloy *et al.*, 2019; Teki *et al.*, 2011; Teki *et al.*, 2016), the current study had variable interchord intervals ranging from 2.5 to 150 ms for the figure chord and the intervening background tones. Thus, participants' use of gaps would likely not have been a useful strategy for figure detection.

On an analytical note, the present study sought to apply mixed-effects regression to the analysis of trial-level behavioral data from SFG tasks, an analytical approach that prior

SFG studies have not commonly used. There are many benefits of mixed-effects regression over other common analytical methods (such as fixed effects regression and ANOVA; see Cunnings, 2012; Linck and Cunnings, 2015), including those used in signal detection theory such as *d*' (Murayama *et al.*, 2014). The primary benefit of mixed-effects regressions of trial-level data over these other analytical methods is that they are able to simultaneously capture both by-participant and by-item random variance, eschewing the traditional approach where data were aggregated and analyzed across participants and items separately (Gordon, 2019). This allows for robust, generalizable findings from only a single model.

Although the current study focused on younger normal-hearing adults, future work would benefit from recruiting individuals with a wider range of auditory and cognitive abilities and examining the impact of adjusting different parameters of SFG stimuli (e.g., temporal jitter, figure duration) on auditory stream segregation in these populations. In particular, aging has been associated with poorer auditory temporal processing (e.g., Fitzgibbons and Gordon-Salant, 2001) and WMC (Bopp and Verhaeghen, 2005; Wang *et al.*, 2011), which may underlie declines in speech-in-noise recognition. Altering the temporal properties of SFG stimuli might then be predicted to have an even larger impact on older adults, who are less able to rely on WMC to compensate and who have generally exhibited more positive correlations between WMC and speech-in-noise than younger adults (Füllgrabe and Rosen, 2016b). Using the SFG stimuli to investigate the varied mechanisms that underlie speech-in-noise deficits may thus help to advance the development of training-based interventions to improve communication among older adults. In addition, the current SFG task is also likely to be useful in studies where it is desirable to assess the mechanisms underlying speech-in-noise processing, irrespective of differences in linguistic processing. For example, a SFG task could be used to identify differences in stream segregation abilities across monolingual and bilingual speakers or to draw parallels about stream segregation across animal and human models.

Future work would especially benefit from examining the relationship between SFG and other measures of WMC (or executive function more broadly), speech-in-noise recognition, and musicianship, particularly with larger sample sizes. As this was the first study using these particular SFG stimuli, the experiment design included some of most commonly used WMC and speech-in-noise measures reported in the literature or used in the clinic that were deployable through remote testing. We focused on a common sentence-in-noise test, for example, because SFG performance depends upon the build-up of temporal coherence over time and because the repetition of chords in our study occurred at a slow aperiodic rate. However, SFG properties could be adjusted to examine the extent to which they better capture the properties of syllable, word, phrase, or sentence perception in noise, and whether that alters the strength or direction of the relationship between SFG performance and WMC. Last, WMC is commonly associated with other domain-general executive functions (e.g., McCabe *et al.*, 2010). Future studies should examine the extent to which the observed relationship between performance and WMC in the easiest SFG condition depends upon using a complex span measure like the RSPAN vs other working memory measures, like the n-back (Redick and Lindsey, 2013).

Furthermore, it is worth considering the limitations and future directions of the relationship between musicianship and performance on SFG tasks found here. Musician and non-musician groups were formed based on self-rated musicianship classifications from one question of the Ollen Musical Sophistication Index, which is notably different from other definitions of musicians and non-musician groups in the literature. Such definitions typically have a more pronounced contrast between groups and are often defined by length of private music lessons, which could be highly related to socioeconomic status or other resources, rather than musical ability. Of note, 2 out of 16 participants in the self-classified musician group in this study indicated that they had not taken private lessons. However, examining the variability of SFG effect sizes across participants revealed that these two individuals fell well inside the range of scores for other self-rated musicians and were not outliers. Future work may further benefit from using full musical sophistication indices, such as the Ollen Musical Sophistication Index and Goldsmiths Musical Sophistication Index (Müllensiefen *et al.*, 2014), as a more granular and continuous measure of musicianship, rather than limiting analyses to a group comparison with a cut point. This may also be particularly interesting in the context of SFG task performance, as Lad *et al.* (2022) found that scores on the Goldsmiths Musical Sophistication Index were correlated with a sound frequency subconstruct of working memory, but not other subconstructs.

VI. CONCLUSION

The results of the present study demonstrated that a non-linguistic SFG task, designed to mimic some aspects of speech stimuli, was related to performance on a standardized measure of speech-in-noise recognition among younger normal-hearing adults. The findings suggest that this SFG task was able to assess lower-level processes of auditory stream segregation. The present study included measures of RT, demonstrating the importance of estimating how quickly auditory stream segregation occurred. Individual differences in WMC and musicianship were both found to modulate performance on this SFG task as well as the relationship between the SFG and QSIN tasks. Together, these findings position this discontinuous SFG task as a potentially viable assessment for measuring such processes without being contaminated by individual linguistic differences, such as strength or speed of lexical processing. The measure may thus be especially suitable for use among

individuals with different language backgrounds, for whom speech-in-noise tasks may not be available in their native language, and in investigating issues of auditory stream segregation across animal and human models.

ACKNOWLEDGMENTS

We would like to thank Matthew Turner and Nick Pandža for their assistance with data collection. We also thank Daniel Stolzberg and Ali Mohammed for discussions regarding auditory stream segregation tasks and SFG stimuli parameters. This study was supported by the Naval Information Warfare Center and Defense Advanced Research Projects Agency under Cooperative Agreement N66001-17-2–4009, NIH/NIA P01AG055365, NIH R01 DC016119, and partial support from AFOSR (FA9550-19-1–0408) and training Grant No. DC-00046 from the National Institute of Deafness and Communicative Disorders of the NIH. The identification of specific products or scientific instrumentation is considered an integral part of the scientific endeavor and does not constitute endorsement or implied endorsement on the part of the author, DoD, or any component agency. The views expressed in this article are those of the author and do not reflect the official policy of the Department of Army/Navy/Air Force, Department of Defense, or U.S. Government.

APPENDIX

Self-report measure of overall musical sophistication and pitch ability. Ratings are based on a five-point Likert scale.

Overall musical sophistication
How frequently do you engage in the following activities?
Rarely/Never (Yearly), Infrequently (Monthly), Sometimes (Weekly), Frequently (Multiple times per week), or Very Frequently (Daily or more than daily)
Listen to music (radio, YouTube, Pandora, Spotify, etc.)
Attend concerts
Play/sing music
Compose music
Pitch ability
How would you rate yourself on the following abilities?
1 – Poor, 2, 3 – Average, 4, 5 – Excellent
Singing in tune
Noticing when someone else is out of tune
Noticing when a wrong note is played in a song

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0016756 for additional information on participants' musical experience and instruments played.
[2]An additional model using d' as the dependent variable revealed identical effects to those listed here for accuracy.

Adank, P. (**2012**). "The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses," Brain Lang. **122**, 42–54.
Agus, T. R., and Pressnitzer, D. (**2021**). "Repetition detection and rapid auditory learning for stochastic tone clouds," J. Acoust. Soc. Am. **150**, 1735–1749.
Akeroyd, M. A. (**2008**). "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults," Int. J. Audiol. **47**, S53–S71.
Alain, C., Du, Y., Bernstein, L. J., Barten, T., and Banai, K. (**2018**). "Listening under difficult conditions: An activation likelihood estimation meta-analysis," Hum. Brain Mapp. **39**, 2695–2709.
Baddeley, A., Logie, R., and Nimmo-Smith, I. (**1985**). "Components of fluent reading," J. Mem. Lang. **24**, 119–131.
Benard, M. R., Susanne Mensink, J., and Başkent, D. (**2014**). "Individual differences in top-down restoration of interrupted speech: Links to linguistic and cognitive abilities," J. Acoust. Soc. Am. **135**, EL88–EL94.
Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2013**). "How linguistic closure and verbal working memory relate to speech recognition in noise—A review," Trends Amplif. **17**, 75–93.
Bidelman, G. M., Hutka, S., and Moreno, S. (**2013**). "Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music," PLoS One **8**, e60676.
Bidelman, G. M., and Yoo, J. (**2020**). "Musicians show improved speech segregation in competitive, multi-talker cocktail party scenarios," Front. Psychol. **11**, 1927.
Boersma, P., and Weenink, D. (**2021**). Praat: Doing phonetics by computer (version 6.1.39) [computer program], http://www.praat.org/ (Last viewed July 10, 2021).
Bopp, K. L., and Verhaeghen, P. (**2005**). "Aging and verbal memory span: A meta-analysis," J. Gerontol., Ser. B: Psychol. Sci. Soc. Sci. **60**, P223–233.
Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (**2020**). "The timing mega-study: Comparing a range of experiment generators, both lab-based and online," PeerJ **8**, e9414.
Burke, L. A., and Naylor, G. (**2020**). "Daily-life fatigue in mild to moderate hearing impairment: An ecological momentary assessment study," Ear Hear. **41**, 1518–1532.
Carroll, R., Warzybok, A., Kollmeier, B., and Ruigendijk, E. (**2016**). "Age-related differences in lexical access relate to speech recognition in noise," Front. Psychol. **7**, 990.
Chen, Y., Bidwell, L. C., and Holzman, P. S. (**2005**). "Visual motion integration in schizophrenia patients, their first-degree relatives, and patients with bipolar disorder," Schizophr. Res. **74**, 271–281.
Coffey, E. B., Mogilever, N. B., and Zatorre, R. J. (**2017**). "Speech-in-noise perception in musicians: A review," Hear. Res. **352**, 49–69.
Coulter, K., Gilbert, A. C., Kousaie, S., Baum, S., Gracco, V. L., Klein, D., Titone, D., and Phillips, N. A. (**2021**). "Bilinguals benefit from semantic context while perceiving speech in noise in both of their languages: Electrophysiological evidence from the N400 ERP," Bilingualism **24**, 344–357.
Cunnings, I. (**2012**). "An overview of mixed-effects statistical models for second language researchers," Second Lang. Res. **28**, 369–382.
Daneman, M., and Carpenter, P. A. (**1980**). "Individual differences in working memory and reading," J. Verbal Learn. Verbal Behav. **19**, 450–466.
Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (**2009**). "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," PLoS Biol. **7**, e1000129.
Fine, M. S., and Minnery, B. S. (**2009**). "Visual salience affects performance in a working memory task," J. Neurosci. **29**, 8016–8021.
Fitzgibbons, P. J., and Gordon-Salant, S. (**2001**). "Aging and temporal discrimination in auditory sequences," J. Acoust. Soc. Am. **109**, 2955–2963.
Füllgrabe, C., and Rosen, S. (**2016a**). "Investigating the role of working memory in speech-in-noise identification for listeners with normal hearing," in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*, edited by P. van Dijk, D. Başkent, E. Gaudrain, E. de Kleine, A. Wagner, and C. Lanting (Springer, Cham), pp. 29–36.
Füllgrabe, C., and Rosen, S. (**2016b**). "On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds," Front. Psychol. **7**, 1268.
Gordon, K. R. (**2019**). "How mixed-effects modeling can advance our understanding of learning and memory and improve clinical and educational practice," J. Speech. Lang. Hear. Res. **62**, 507–524.

Harbinson, I. J., Atkins, S. M., and Dougherty, M. R. (**2011**). "N-back training task performance: Analysis and model," Proc. Ann. Meet. Cog. Sci. Soc. **33**, 120–125, available at https://escholarship.org/uc/item/37c3c9hq.
Helfer, K. S., and Jesse, A. (**2015**). "Lexical influences on competing speech perception in younger, middle-aged, and older adults," J. Acoust. Soc. Am. **138**, 363–376.
Holm, S. (**1979**). "A simple sequentially rejective multiple test procedure," Scand. J. Stat. **6**, 65–70.
Holmes, E., and Griffiths, T. D. (**2019**). " 'Normal' hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception," Sci. Rep. **9**, 16771.
Holmes, E., Zeidman, P., Friston, K. J., and Griffiths, T. D. (**2021**). "Difficulties with speech-in-noise perception related to fundamental grouping processes in auditory cortex," Cereb. Cortex **31**, 1582–1596.
Huang, L., Wang, L., Shen, W., Li, M., Wang, S., Wang, X., Ungerleider, L. G., and Zhang, X. (**2020**). "A source for awareness-dependent figure–ground segregation in human prefrontal cortex," Proc. Natl. Acad. Sci. U.S.A. **117**, 30836–30847.
Kaandorp, M. W., De Groot, A. M. B., Festen, J. M., Smits, C., and Goverts, S. T. (**2016**). "The influence of lexical-access ability and vocabulary knowledge on measures of speech recognition in noise," Int. J. Audiol. **55**, 157–167.
Kaplan, E. C., Wagner, A. E., Toffanin, P., and Başkent, D. (**2021**). "Do musicians and non-musicians differ in speech-on-speech processing?," Front. Psychol. **12**, 623787.
Kidd, G., Mason, C. R., and Best, V. (**2014**). "The role of syntax in maintaining the integrity of streams of speech," J. Acoust. Soc. Am. **135**, 766–777.
Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (**2004**). "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **116**, 2395–2405.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest Package: Tests in linear mixed effects models," J. Stat. Softw. **82**, 1–26.
Lad, M., Billig, A. J., Kumar, S., and Griffiths, T. D. (**2022**). "A specific relationship between musical sophistication and auditory working memory," Sci. Rep. **12**, 3517.
Lad, M., Holmes, E., Chu, A., and Griffiths, T. D. (**2020**). "Speech-in-noise detection is related to auditory working memory precision for frequency," Sci. Rep. **10**, 1.
Lamme, V. (**1995**). "The neurophysiology of figure-ground segregation in primary visual cortex," J. Neurosci. **15**, 1605–1615.
Linck, J. A., and Cunnings, I. (**2015**). "The utility and application of mixed-effects models in second language research," Lang. Learn. **65**, 185–207.
Lüdecke, D. (**2018**). "ggeffects: Tidy data frames of marginal effects from regression models," JOSS **3**, 772.
Madsen, S. M. K., Marschall, M., Dau, T., and Oxenham, A. J. (**2019**). "Speech perception is similar for musicians and non-musicians across a wide range of conditions," Sci. Rep. **9**, 10404.
Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N., and Blamey, P. J. (**2010**). "The effect of visual cues on auditory stream segregation in musicians and non-musicians," PLoS One **5**, e11297.
Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cogn. Process. **27**, 953–978.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (**2017**). "Balancing Type I error and power in linear mixed models," J. Mem. Lang. **94**, 305–315.
McAuliffe, M. J., Gibson, E. M. R., Kerr, S. E., Anderson, T., and LaShell, P. J. (**2013**). "Vocabulary influences older and younger listeners' processing of dysarthric speech," J. Acoust. Soc. Am. **134**, 1358–1368.
McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., and Hambrick, D. Z. (**2010**). "The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct," Neuropsychology **24**, 222–243.
Molloy, K., Lavie, N., and Chait, M. (**2019**). "Auditory figure-ground segregation is impaired by high visual load," J. Neurosci. **39**, 1699–1708.
Moore, B. C. J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acta Acust. united Ac. **88**, 320–332.
Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (**2014**). "The musicality of non-musicians: An index for assessing musical sophistication in the general population," PLoS One **9**, e89642.
Murayama, K., Sakaki, M., Yan, V. X., and Smith, G. M. (**2014**). "Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective," J. Exp. Psychol. Learn. Mem. Cogn. **40**, 1287–1306.
Okazaki, S., and Ichikawa, M. (**2017**). "Perceptual simultaneity range as a function of frequency separation for two pure tones," Acoust. Sci. Tech. **38**, 185–192.
Ollen, J. E. (**2006**). "A criterion-related validity test of selected indicators of musical sophistication using expert ratings," Ph.D. thesis, Ohio State University, Colubmus, OH.
O'Sullivan, J. A., Shamma, S. A., and Lalor, E. C. (**2015**). "Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening," J. Neurosci. **35**, 7256–7263.
Pang, J., Beach, E. F., Gilliver, M., Yeend, I., Pang, J., Beach, E. F., Gilliver, M., and Yeend, I. (**2019**). "Adults who report difficulty hearing speech in noise: An exploration of experiences, impacts and coping strategies," Int. J. Audiol. **58**, 851–860.
Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009**). "Musician enhancement for speech-in-noise," Ear Hear. **30**, 653–661.
Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., and Kraus, N. (**2011**). "Musical experience and the aging auditory system: Implications for cognitive abilities and hearing speech in noise," PLoS One **6**, e18082.
Parker, E. M. (**1988**). "Auditory constraints on the perception of voice-onset time: The influence of lower tone frequency on judgments of tone-onset simultaneity," J. Acoust. Soc. Am. **83**, 1597–1607.
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (**2019**). "PsychoPy2: Experiments in behavior made easy," Behav. Res. **51**, 195–203.
Pichora-Fuller, M. K. (**2008**). "Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing," Int. J. Audiol. **47**, S72–S82.
Pooresmaeili, A., Bach, D. R., and Dolan, R. J. (**2014**). "The effect of visual salience on memory-based choices," J. Neurophysiol. **111**, 481–487.
Presacco, A., Simon, J. Z., and Anderson, S. (**2016**). "Evidence of degraded representation of speech in noise in the aging midbrain and cortex," J. Neurophysiol. **116**, 2346–2355.
Rammsayer, T., and Altenmüller, E. (**2006**). "Temporal information processing in musicians and nonmusicians," Music Percept. **24**, 37–48.
Rammsayer, T. H., Buttkus, F., and Altenmüller, E. (**2012**). "Musicians do better than nonmusicians in both auditory and visual timing tasks," Music Percept. **30**, 85–96.
R Core Team (**2021**). "R: A Language and Environment for Statistical Computing," https://www.R-project.org/ (Last viewed August 15, 2021).
Redick, T. S., and Lindsey, D. R. B. (**2013**). "Complex span and *n*-back measures of working memory: A meta-analysis," Psychon. Bull. Rev. **20**, 1102–1113.
Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (**2006**). "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," Appl. Psycholinguist. **27**, 465–485.
Rönnberg, J. (**1990**). "Cognitive and communicative function: The effects of chronological age and 'handicap age'," Eur. J. Cogn. Psychol. **2**, 253–273.
Rönnberg, J., Arlinger, S., Lyxell, B., and Kinnefors, C. (**1989**). "Visual evoked potentials: Relation to adult speechreading and cognitive function," J. Speech. Lang. Hear. Res. **32**, 725–735.
Rönnberg, J., Luner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., and Rudner, M. (**2013**). "The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances," Front. Syst. Neurosci. **7**, 31.
Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (**2008**). "Cognition counts: A working memory system for ease of language understanding," Int. J. Audiol. **47**, S99–S105.
Russell, L. (**2021**). "emmeans: Estimated marginal means, aka least-square means," R package version 1.7.1-1, https://CRAN.R-project.org/package-emmeans (Last viewed August 23, 2021).
SanMiguel, I., Corral, M. J., and Escera, C. (**2008**). "When loading working memory reduces distraction: Behavioral and electrophysiological evidence from an auditory-visual distraction paradigm," J. Cogn. Neurosci. **20**, 1131–1145.
Satterthwaite, F. E. (**1941**). "Synthesis of variance," Psychometrika **6**, 309–316.

Schröter, H., Ulrich, R., and Miller, J. (**2007**). "Effects of redundant auditory stimuli on reaction time," Psychon. Bull. Rev. **14**, 39–44.
Shamma, S. A., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123.
Sherbon, J. W. (**1975**). "The association of hearing acuity, diplacusis, and discrimination with music performance," J. Res. Music Educ. **23**, 249–257.
Shinn-Cunningham, B. G., and Best, V. (**2008**). "Selective attention in normal and impaired hearing," Trends Amplif. **12**, 283–299.
Skoe, E., and Karayanidi, K. (**2019**). "Bilingualism and speech understanding in noise: Auditory and linguistic factors," J. Am. Acad. Audiol. **30**, 115–130.
Teki, S., Barascud, N., Picard, S., Payne, C., Griffiths, T. D., and Chait, M. (**2016**). "Neural correlates of auditory figure-ground segregation based on temporal coherence," Cereb. Cortex **26**, 3669–3680.
Teki, S., Chait, M., Kumar, S., Shamma, S., and Griffiths, T. D. (**2013**). "Segregation of complex acoustic scenes based on temporal coherence," ELife **2**, e00699.
Teki, S., Chait, M., Kumar, S., Von Kriegstein, K., and Griffiths, T. D. (**2011**). "Brain bases for auditory stimulus-driven figure-ground segregation," J. Neurosci. **31**, 164–171.
Thiel, C. M., Özyurt, J., Nogueira, W., and Puschmann, S. (**2016**). "Effects of age on long term memory for degraded speech," Front. Hum. Neurosci. **10**, 473.
Torres-Russotto, D., Landau, W. M., Harding, G. W., Bohne, B. A., Sun, K., and Sinatra, P. M. (**2009**). "Calibrated finger rub auditory screening test (CALFRAST)," Neurology **72**, 1595–1600.
Vermeire, K., Knoop, A., De Sloovere, M., Bosch, P., and van den Noort, M. (**2019**). "Relationship between working memory and speech-in-noise recognition in young and older adult listeners with age-appropriate hearing," J. Speech. Lang. Hear. Res. **62**, 3545–3553.
Voeten, C. C. (**2021**). "Buildmer: Stepwise elimination and term reordering for mixed-effects regression (R package version 1.9)," https://cran.r-project.org/package=buildmer (Last viewed October 19, 2021).
Wang, M., Gamo, N. J., Yang, Y., Jin, L. E., Wang, X.-J., Laubach, M., Mazer, J. A., Lee, D., and Arnsten, A. F. T. (**2011**). "Neuronal basis of age-related working memory decline," Nature **476**, 210–213.
Wilson, R. H. (**2003**). "Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance," J. Am. Acad. Audiol. **14**, 453–470.
Wilson, R. H., McArdle, R. A., and Smith, S. L. (**2007**). "An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss," J. Speech Lang. Hear. Res. **50**, 844–856.
Wingfield, A., Amichetti, N. M., and Lash, A. (**2015**). "Cognitive aging and hearing acuity: Modeling spoken language comprehension," Front. Psychol. **6**, 00684.
Yoo, J., and Bidelman, G. M. (**2019**). "Linguistic, perceptual, and cognitive factors underlying musicians' benefits in noise-degraded speech perception," Hear. Res. **377**, 189–195.
Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2010**). "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," Ear Hear. **31**, 480–490.
Zendel, B. R., and Alain, C. (**2009**). "Concurrent sound segregation is enhanced in musicians," J. Cogn. Neurosci. **21**, 1488–1498.
Zendel, B. R., and Alain, C. (**2012**). "Musicians experience less age-related decline in central auditory processing," Psychol. Aging **27**, 410–417.
Zendel, B. R., Tremblay, C. D., Belleville, S., and Peretz, I. (**2015**). "The impact of musicianship on the cortical mechanisms related to separating speech from background noise," J. Cogn. Neurosci. **27**, 1044–1059.
Zhang, J. D., and Schubert, E. (**2019**). "A single item measure for identifying musician and nonmusician categories based on measures of musical sophistication," Music Percept. **36**, 457–467.