Trends in Hearing
Volume 21: 1–12
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2331216517739427
journals.sagepub.com/home/tia
SAGE

The Effects of Musical Training on Speech Detection in the Presence of Informational and Energetic Masking

Charlotte Morse-Fortier[1,*], Mary M. Parrish[1,#], Jane A. Baran[1], and Richard L. Freyman[1]

Abstract

Recent research has suggested that musicians have an advantage in some speech-in-noise paradigms, but not all. Whether musicians outperform nonmusicians on a given speech-in-noise task may well depend on the type of noise involved. To date, few groups have specifically studied the role that informational masking plays in the observation of a musician advantage. The current study investigated the effect of musicianship on listeners' ability to overcome informational versus energetic masking of speech. Monosyllabic words were presented in four conditions that created similar energetic masking but either high or low informational masking. Two of these conditions used noise-vocoded target and masking stimuli to determine whether the absence of natural fine structure and spectral variations influenced any musician advantage. Forty young normal-hearing listeners (20 musicians and 20 nonmusicians) completed the study. There was a significant overall effect of participant group collapsing across the four conditions; however, planned comparisons showed musicians' thresholds were only significantly better in the high informational masking natural speech condition, where the musician advantage was approximately 3 dB. These results add to the mounting evidence that informational masking plays a role in the presence and amount of musician benefit.

Keywords

musical training, informational masking, speech detection

Date received: 30 April 2017; revised: 11 August 2017; accepted: 1 October 2017

Introduction

The hypothesis that highly trained musicians have special auditory skills has been the topic of much recent research. Musical performance (with the exception of solo performance) requires the ability to maintain attention and distinguish individual sound elements in the presence of competing instruments, voices, rhythms, melodies, and harmonies. Possibly due to the rigorous training and rehearsal that professional musicians engage in on a regular basis, as a group they are able to discriminate tonal sounds more easily than nonmusicians (Oxenham, Fligor, Mason, & Kidd, 2003; Parbery-Clark, Skoe, Lam, & Kraus, 2009b; Zendel & Alain, 2009). Several recent studies have focused on the question of whether these reported advantages extend beyond the domain of music, specifically whether musicians are better at listening to speech in the presence of background noise (e.g., Başkent & Gaudrain, 2016; Clayton et al., 2016; Kraus & Nicol, 2010; Madsen, Whiteford, & Oxenham, 2017; Parbery-Clark, Skoe, & Kraus, 2009a; Parbery-Clark et al., 2009b; Ruggles, Freyman, & Oxenham, 2014; Strait, Kraus, Parbery-Clark, & Ashley, 2010; Swaminathan et al., 2015).

Recently, Coffey, Mogilever, and Zatorre (2017) completed a review of the literature most closely related to the topic of musical training and speech perception

[1] Department of Communication Disorders, University of Massachusetts Amherst, MA, USA

*Charlotte Morse-Fortier is now at Department of Otolaryngology, Harvard Vanguard Medical Associates, Boston, MA, USA.

#Mary M. Parrish is now at Dartmouth-Hitchcock Ear, Nose and Throat Physicians of Southern New Hampshire, Manchester, NH, USA.

Corresponding author:

Charlotte Morse-Fortier, Audiology Department, Harvard Vanguard Medical Associates, 133 Brookline Avenue, Boston, MA 02215, USA.
Email: morsefortier.charlotte@gmail.com

in noise. In all, 29 articles, which included papers on electrophysiology and tonal perception, were analyzed. Of these articles, most of them recent publications, 16 included investigations of behavioral speech-in-noise tasks, measuring sentence, word, or phoneme-level perception. Coffey et al.'s analysis of this research revealed mixed results. Across these 16 papers, the results for 37 conditions were reported, with 20 showing a statistically significant musician advantage for musicians and 17 not showing an advantage, the latter number spread across about half the articles (see their Figure 1). It seems clear that a significant musician advantage is found in some, but not nearly all conditions reported in the literature.

The advantages found for subjects with musical training are often small, as little as 1 dB or less in threshold signal-to-noise ratio (SNR; e.g., Parbery-Clark et al., 2009a). Whether these advantages are found to be statistically significant can be partially dependent on subject sampling in the individual studies (Boebinger et al., 2015; Coffey et al., 2017). These sampling issues may hinder attempts to find patterns of results across studies that explain the most important elements responsible for musician advantages, when they occur.

To sort out some of these issues, Coffey et al. (2017) called for targeted manipulations that could probe the key factors that determine the conditions under which musicians have an advantage in speech-in-noise perception. One of these potentially important factors is the type of noise that is used to explore masked speech perception performance. When the noise is competing speech, there can be masking elements beyond traditional ''energetic'' masking, in some cases creating confusion between target and masker (Brungart, 2001; Kidd, Mason, & Arbogast, 2002). Several investigators (e.g., Başkent & Gaudrain, 2016; Boebinger et al., 2015; Clayton et al., 2016; Swaminathan et al., 2015) have proposed that musicians might demonstrate substantial advantages in situations where this type of masking, often called informational masking, dominates. Indeed, the few studies published so far on informational masking conditions appear to show greater benefits of musical training than have been observed in most other studies. The purpose of the current article was to expand the literature on musician advantages in speech perception with informational maskers.

Findings from studies with nonspeech stimuli support the premise that musicians might well have an advantage with informational masking in speech perception. Differences in fundamental frequency can help the listener resolve potential confusion between the target speaker and competing speech. Because musicians are trained to focus on pitch, one might expect them to use this cue more efficiently than nonmusicians (Kishon-Rabin, Amir, Vexler, & Zaltz, 2001; Micheyl, Delhommeau, Perrot, & Oxenham, 2006). For tonal stimuli, at least one study (Oxenham et al., 2003) has reported a particularly large advantage for musicians in conditions where informational masking has been shown to be a dominant feature (Kidd, Mason, Deliwala, Woods, & Colburn, 1994). In the 2003 paper, two different multi-tone maskers were used to mask a 1000-Hz tone burst that differed markedly in the amount of informational masking produced. Masker intensity was varied adaptively to determine the SNR at which a listener could detect the target 1-kHz tone in the presence of each of these maskers. Musicians and nonmusicians performed similarly in the low informational masking condition. However, musicians had a 9-dB average difference in their masked detection thresholds between the high and low informational masking conditions, whereas nonmusicians had nearly a 25-dB average difference in performance. Thus, for this task, the results indicate that many of the musicians were able to largely overcome informational masking, whereas nonmusicians were able to do so much less consistently.

Only a few studies in the literature have addressed whether these kinds of results for nonspeech stimuli translate to musician advantages in speech perception in the presence of informational masking. Many of the speech perception studies on the effects of musical training have included single- or multitalker maskers, but it cannot always be clearly determined whether there was a significant informational masking component. For just one or two interfering talkers, there are noticeable spectrotemporal dips in the masker that young normal-hearing listeners can often exploit. For this reason, speech is sometimes a less effective masker than steady-state noise (e.g., Duquesnoy, 1983; Edmonds & Culling, 2005), and when this occurs, informational masking is not strongly implicated. Conversely, when speech maskers produce greater amounts of masking relative to noise maskers, or if threshold SNRs are elevated, this is suggestive of a substantial ''perceptual'' or informational masking component (Carhart, Tillman, & Greetis, 1969). As more masking talkers are added, it can be more difficult to identify the role of informational masking. For example, it is unclear how much informational masking is presented by the four-talker babble from the QuickSIN test, which has been used in some previous studies on the effects of musical training (Parbery-Clark et al., 2009a, 2009b; Ruggles et al., 2014).

A second indication of significant informational masking is when the result of spatially separating target and masker is larger than predicted based on head shadow and binaural interaction effects, either in anechoic or reverberant space (Freyman, Helfer, McCall, & Clifton, 1999; Kidd, Mason, Brughera, & Hartmann, 2005). Informational masking has also been quantified by ideal time-frequency segregation, where time-frequency units in the target-masker mixture that contain poor SNRs are removed. This processing has substantially more benefit in informational than energetic masking (Brungart, Chang, Simpson, & Wang, 2006; Kidd et al., 2016).

Among the articles reviewed by Coffey et al. (2017), there are only a handful of conditions in which one can be reasonably certain of a large informational masking component, as evaluated on the indications discussed above. Swaminathan et al. (2015) compared the performance of 12 musicians and 12 nonmusicians on a sentence recognition task with a strong informational masking component. The stimuli were semantically correct, unpredictable five-word sentences (corpus developed by Kidd, Best, & Mason, 2008). The target sentence and two simultaneous masking sentences were spoken by different female talkers in co-located and spatial conditions, and with natural and time-reversed maskers.

Results indicated that threshold performance for both groups in the co-located natural speech condition required SNRs that were between 0 and +5 dB. These are high values that strongly suggest significant informational masking, especially when compared with the results observed for the reversed nonspatial masker condition where thresholds for both groups were below −10 dB SNR. The nonspatial natural masker conditions (where the maskers and target were similar sentences that began synchronously) seemed to be so difficult that the task could not be solved unless the target was louder than the maskers. The difficulty of this task could have obscured any potential musician advantages in the co-located condition and, indeed, differences between the groups in this high informational masking condition were small and nonsignificant. Musicians outperformed nonmusicians by an average of 6.6 dB when the natural maskers were separated by ±15°, indicating that the spatial cues were more likely effective on average in releasing informational masking for the musician group. Similar results were observed in a second study with 17 additional subjects per group (Clayton et al., 2016), using identical target stimuli with the natural maskers only.

Başkent and Gaudrain (2016) asked musicians and nonmusicians to repeat sentences spoken by a target male talker in the presence of masker sentences spoken by the same male talker at an SNR of −6 dB. Differences in fundamental frequency (F_0) and vocal tract length between the target and masker were manipulated via signal processing in different conditions. Results showed that both groups benefitted from greater F_0 and vocal tract length differences between target and masker, but musicians significantly outperformed nonmusicians in all conditions, by almost 20 percentage points in the unprocessed condition. While these data cannot be directly compared with decibel differences in threshold SNRs reported in many other studies, the magnitude of the effect is sizeable. The extent of informational masking in the conditions tested is not easy to assess, but using the same talker as both target and masker increases the potential for informational masking (Brungart, 2001).

Boebinger et al. (2015) also compared musicians and nonmusicians on a speech-in-noise task, using the Bamford–Kowal–Bench sentences spoken by a female talker. They tested speech-in-noise performance in the presence of four different maskers derived from a male voice. The maskers were intended to vary the amount of informational masking and consisted of the following (from most to least): clear speech, spectrally rotated speech, amplitude-modulated speech noise, and speech-spectrum steady-state noise. The results indicated no significant performance differences between the musician and nonmusician groups for any of the four masker conditions. However, even in the putative highest informational masking condition, it is possible that informational masking was not very substantial due to the use of different-sex target and masking talkers (Brungart, 2001). Indeed, subjects achieved the best threshold SNRs in the clear speech masker condition, which is not strongly indicative of informational masking.

Based upon the extant literature on the effects of musical training, there are only a few conditions where it is obvious that informational masking was a dominant stimulus feature. There is clearly much more work to be done to understand how musical training affects informational masking in speech recognition. The current study used a basic word detection task in four conditions to investigate the effects of musical training on informational and energetic masking. A two-talker masker recorded by speakers of the same sex as the target speaker was used to maximize informational masking. The paradigm was identical in many respects to a subset of experiments conducted by Freyman, Balakrishnan, and Helfer (2008) and Balakrishnan and Freyman (2008), but in those studies, subjects were not asked about musical training. The experiment was run both with and without target-masker spatial separation to vary the amount of informational masking.

The current paradigm used unprocessed targets and maskers as well as with stimuli processed with noise-excited vocoding, which effectively eliminated pitch cues. Based on previous research with tonal stimuli (Oxenham et al., 2003) and speech stimuli (Başkent & Gaudrain, 2016; Clayton et al., 2016; Swaminathan et al., 2015), we hypothesized that musicians would have a significant advantage in the natural speech nonspatial high informational masking condition. For this condition, Balakrishnan and Freyman (2008) reported that some subjects appeared to be able to partially resolve informational masking in some adaptive runs. The current study evaluated whether such subjects were more likely to be musicians than nonmusicians. For the nonspatial vocoded condition where better processing of pitch cues would not afford an advantage, as well as the spatial (primarily energetic masking) conditions, the accumulated research reviewed by Coffey et al. (2017) suggests that there may be a small musician advantage that may or may not be statistically significant.

Methods

Participants

Forty college students with normal hearing (thresholds of 25 dB HL or better at octave frequencies from 250 to 8000 Hz, as measured on the day of participation) completed the study. Twenty subjects were trained musicians and 20 were nonmusicians, using criteria similar to those used by Oxenham et al. (2003). Subjects in the musician group consisted of 13 females and 7 males with an average age of 20.1 years. Nonmusician subjects consisted of 19 females and 1 male with an average age of 22.5 years (range: 20 to 28 years). As Oxenham et al. (2003) found no significant effect of gender on performance in their study, this subject variable was not controlled in the present study. Table 1 presents detailed subject information for the musician participants in the current study.

Subjects who were classified as musicians reported daily musical practice and were currently enrolled in a music program at the university level. All musicians had completed two or more years of private lessons prior to or during their college enrollments. Most of the subjects in the musician group played more than one instrument. Subjects in the nonmusician group did not have any history of formal musical training. Twelve nonmusicians were undergraduate students in the Department of Communication Disorders, and eight were undergraduate students from other departments. Those recruited from the Department of Communication Disorders were given extra credit for participation in this study; the remaining subjects were compensated financially for their time. The University of Massachusetts Institutional Review Board approved all procedures, and all subjects provided written consent.

Table 1. Subject Demographics for Musicians.

Musicians	Age	Gender	Years played	Instrument(s)
1	21	F	13	Clarinet
2	18	F	5	Trumpet
3	20	F	10	Clarinet
4	18	F	8, 2.5,3	Piano, Viola, Clarinet
5	22	M	14	Trumpet
6	21	F	8,12,6	Piccolo, Flute, Sax
7	20	F	11	Clarinet
8	18	M	9	Trombone
9	21	F	7,2,2	Bassoon, Clarinet, Piano
10	18	F	13	Piano, Flute, Cello
11	20	F	11	Piano, Voice, Clarinet
12	20	F	14,10	Piano, Trombone
13	23	F	15	Piano
14	19	M	2,10,3,1,5,1	Voice, Trumpet, Piano, Percussion, Flute
15	20	F	16,11,8,3	Piano, Clarinet, Saxophone, Oboe
16	19	M	3,10,2,1	Euphonium, Drums, Piano, Guitar
17	19	F	10	Piano
18	19	M	7	Trombone, Voice, Piano
19	24	M	15,8,1,4,2,2,2	Saxophone, Voice, Trombone, Trumpet, Flute, Clarinet, Piano
20	21	M	12, 18, 1	Percussion, Drums, Ocarina

Stimuli

The target stimuli for the two-talker natural speech masker conditions consisted of 20 consonant–vowel–consonant words excised from a list of nonsense sentences developed by Helfer (1997) and spoken by an adult female. The masker consisted of a mixture of the speech of two other recorded female talkers who recited nonsense sentences from the same corpus (but not the sentences from which the target words were drawn). The sentences from each of the two masking talkers were concatenated to form a steady stream of speech 35 s long. The two streams (one from each talker) were mixed at equal overall root mean square (RMS) levels and played in a continuous loop. The choice of a two-talker paradigm in the present study was based upon previous findings where spatial release from informational masking was largest for two-talker babble and diminished as additional talkers were added to the masker (Freyman, Balakrishnan, & Helfer, 2004).

For the vocoded conditions, the target words and masker described earlier were processed using six-channel vocoding with a noise carrier using the same algorithm employed by Qin and Oxenham (2003). The frequency range of 80 to 6000 Hz was divided into six channels of equal bandwidth according to the equivalent rectangular bandwidth scale (Glasberg & Moore, 1990), using digital sixth-order Butterworth bandpass filters. Envelopes were extracted from the filter outputs by digitally low-pass filtering rectified signals with a cutoff frequency of either 300 Hz or half the bandwidth (whichever was lower), using a second-order Butterworth filter. White noise filtered to have the same bandwidth as the filtered signals was multiplied by the corresponding envelope in the time domain to create noises that matched the temporal envelopes in each channel. The six modulated noises were summed to create a broadband six-channel speech-envelope-modulated noise for each of the 20 target words and the two-talker masker. This vocoding process removed the natural fine structure of the speech signal and minimized the pitch and intonation contours existing in natural speech.

Equipment

Testing took place in an anechoic chamber (4.9 × 4.1 × 3.12 m) where 0.72-m foam wedges covered the walls, ceiling, and floor. Subjects were seated at the center of the chamber on a chair supported by a wire grid, facing a foam-covered semicircular arc on which the loudspeakers were placed. Stimuli were presented from two loudspeakers, placed at ear level to the height of an average adult's head. The front speaker was positioned at 0° horizontal azimuth from the subject, and the right loudspeaker was positioned at 60° to the right of the subject, and both were located 1.9 m from the approximate center of the subject's head.

Separate computers were used to deliver the target and masker. The masker was delivered by a Windows-based personal computer (Dell Dimension XPD 333), using the computer sound card with Cool Edit Pro software. A second computer was used to deliver the stimuli containing the target words via a digital-to-analog converter (TDT DA1) that was sent through a low-pass filter at 8.5 kHz (TDT), attenuated (TDT PA3), and mixed in with the masker (TDT SUM3). The final output was delivered through a Crown D40 amplifier and Realistic Minimus 7 loudspeakers.

Calibration of stimulus levels was performed prior to the beginning of testing each day. A sawtooth wave ($F_0 = 100$ Hz) having the same RMS level as the target words was used to calibrate the target. The masker was calibrated using a speech-spectrum noise with the same RMS as the masker. A B&K 2204 sound level meter was used to measure the microphone output using the C-scale and Fast meter response. The microphone was positioned at the approximate center of the subject's head.

Procedures

The procedures employed in the present study were similar to those used by Balakrishnan and Freyman (2008). Subjects were seated in the middle of the anechoic chamber facing the front loudspeaker and were given a response box with light-emitting diodes and four buttons that lit up sequentially, marking four temporal intervals. One target word was chosen randomly for each trial and was presented during one of the four temporal intervals. Subjects were familiarized with the 20 target words by reading a list of the words before testing. This list was removed before testing began. Subjects were instructed to select the button for the interval in which they heard the target word in a four-alternative-forced-choice paradigm. After they responded with a button press, the light-emitting diode that corresponded to the correct interval was illuminated to provide feedback.

The masker was turned on and set to play on loop mode before beginning each adaptive track. The masker was set at a fixed level of 50 dBC in each masker channel, while the level of the target stimulus was adapted. A two-hits-down one-miss-up stepping rule was used to estimate the 70.7% criterion performance for detection of the target word (Levitt, 1971). An individual adaptive track (a ''run'') consisted of 10 reversals of the adaptive tracking. The subject's threshold was calculated as the arithmetic mean of the last six reversals. The initial step size was 16 dB, which was gradually reduced as reversals progressed to a final step size of 2 dB.

For all subjects, data were collected for four conditions: two listening configurations (spatial and nonspatial) and two speech types (two-talker natural speech and two-talker vocoded speech). In the vocoded conditions, both the target and masker were vocoded. For the spatial conditions, the masker was delivered from both front and right loudspeakers, while the target was delivered through the front loudspeaker only. Zero padding was used to delay the masker delivered through the front channel by 4 ms relative to the right channel. Due to the precedence effect, the masker is localized well to right of the target when using this particular presentation setup (Freyman et al., 1999). This configuration has been shown to give listeners 17 to 20 dB of release from informational masking in this detection task, with little to no release from energetic masking (Balakrishnan & Freyman, 2008; Freyman et al., 2008). For the nonspatial conditions, the right loudspeaker was turned off, and both the target and masker were presented through the front loudspeaker only.

The final threshold estimate for a subject was calculated from the arithmetic mean of thresholds obtained from four runs for each of four conditions. Sixteen runs completed a session, divided into two sets of eight runs for each spatial configuration. Within the set of eight runs, the speech type (natural or vocoded) was alternated for each pair of two runs. Subjects were given a short break, if needed, between the sets of eight runs. In each group, 10 subjects completed the spatial conditions first, and 10 completed the nonspatial conditions first. The order of speech type alternation was also counterbalanced across subjects, equally for each group. At the beginning of the listening session, subjects were given a practice run with the masker in the nonspatial loudspeaker condition (where the target and masker were presented from the front loudspeaker). The study took each participant about 2 hr to complete.

Results

Individual subject means were calculated from the average of each subject's four runs in each condition. Any run with a standard deviation of more than 5 dB across the last six reversals was discarded from data analysis; thus, some means were based on the average of three runs. Twenty-one runs were discarded from the total 640 runs (5 from musicians and 16 from nonmusicians). A three-way analysis of variance was conducted (musician/nonmusician × natural/vocoded × spatial/nonspatial). There was a significant main effect of musicianship, $F(1, 38) = 4.206$, $p = .047$, where musicians performed better than nonmusicians across all conditions. Significant main effects were also found for speech type, $F(1, 38) = 131.741$, $p < .001$, and spatial configuration, $F(1, 38) = 1906.85$, $p < .001$. Detection thresholds for natural speech were 2.5 to 6 dB better than for vocoded speech, depending on the condition (Table 2), consistent with previous findings (Freyman et al., 2008). Similarly, detection thresholds for speech in spatially separated maskers were much lower than for nonspatial conditions, with subjects displaying an average spatial release from masking of approximately 20 dB for natural speech and 21 dB for vocoded speech. This substantial improvement is considered to be due to release from informational masking because this spatial configuration has not been found to lead to any release from continuous noise masking (Balakrishnan & Freyman, 2008; Freyman et al., 2008). The three-way interaction (musician/nonmusician × spatial/co-located ×natural/vocoded) did not reach significance, $F(1, 38) =3.465$, $p = .070$.

Based on the hypotheses guiding the study design, planned comparisons were conducted for musician effect on the four experimental conditions. There was a significant effect of group, $F(1, 38) = 5.565$, $p = .024$, for the natural nonspatial condition only, wherein the musicians (mean SNR threshold = −7.75 dB) outperformed the nonmusicians (mean SNR = −4.56) by 3.19 dB. Descriptive statistics for all four conditions are presented in Figure 1 and Table 2. Inspection of these data shows that musicians also had better thresholds than nonmusicians in the remaining three conditions, although these differences did not reach statistical significance.

The nature of the differences between musicians and nonmusicians for the natural nonspatial condition is visible in Figure 2 (upper left panel), which plots subject thresholds for each individual adaptive run, ranked according to threshold. Runs are plotted individually due to the large threshold variability between subject runs, especially in the nonspatial conditions. Several subjects achieved substantial release from informational masking in those conditions, but often on only one or two of the four runs. It is difficult to determine an exact threshold SNR below which a subject has overcome or "broken through" informational masking. However, using a criterion of approximately −10 dB SNR, it is apparent that many musician runs (17/79 runs or 21.5%) showed detection of the target at thresholds equal to or better than this criterion, while very few nonmusician runs revealed breakthroughs (3/74 runs or 4%).

In the most difficult condition, where the stimuli were vocoded and the target and masker were co-located (upper right panel of Figure 2), musicians had lower average thresholds, but the two best runs were made by a nonmusician (NM8). This brought the nonmusician average down considerably. Despite this one exceptional nonmusician, musicians still had lower average thresholds than nonmusicians by 0.77 dB in this condition. Recalculating the average SNR without

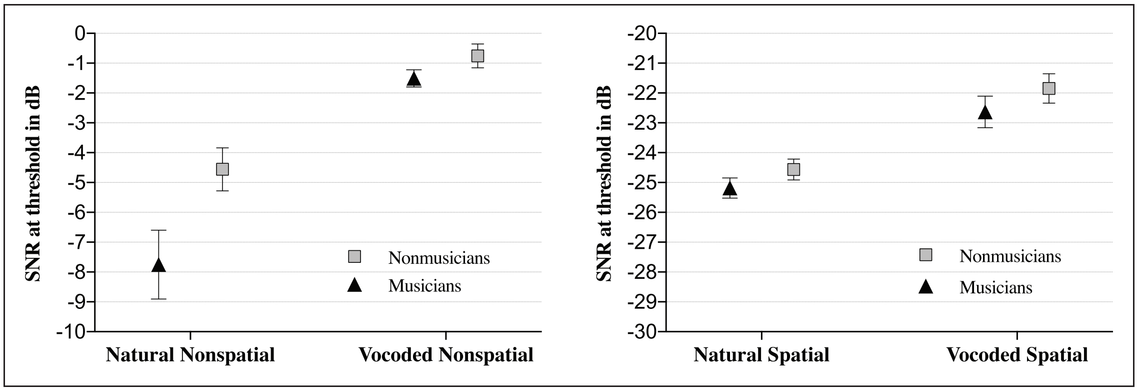

Figure 1. Left panel: Mean ± one standard error for nonspatial conditions. Right panel: Mean ± one standard error for spatial conditions. Note the 20-dB shift of the *y* axis between left and right panels.
SNR = signal-to-noise ratio.

Table 2. Mean SNR Thresholds (Standard Error) by Subject Group and Condition.

	Natural nonspatial	Vocoded nonspatial	Natural spatial	Vocoded spatial
Musician mean SNR threshold (*SE*)	−7.75 (1.16)	−1.51 (0.37)	−25.18 (0.37)	−22.63 (0.46)
Nonmusician mean SNR threshold (*SE*)	−4.56 (0.68)	−0.75 (0.67)	−24.56 (0.38)	−21.84 (0.43)

Note. SNR = signal-to-noise ratio.

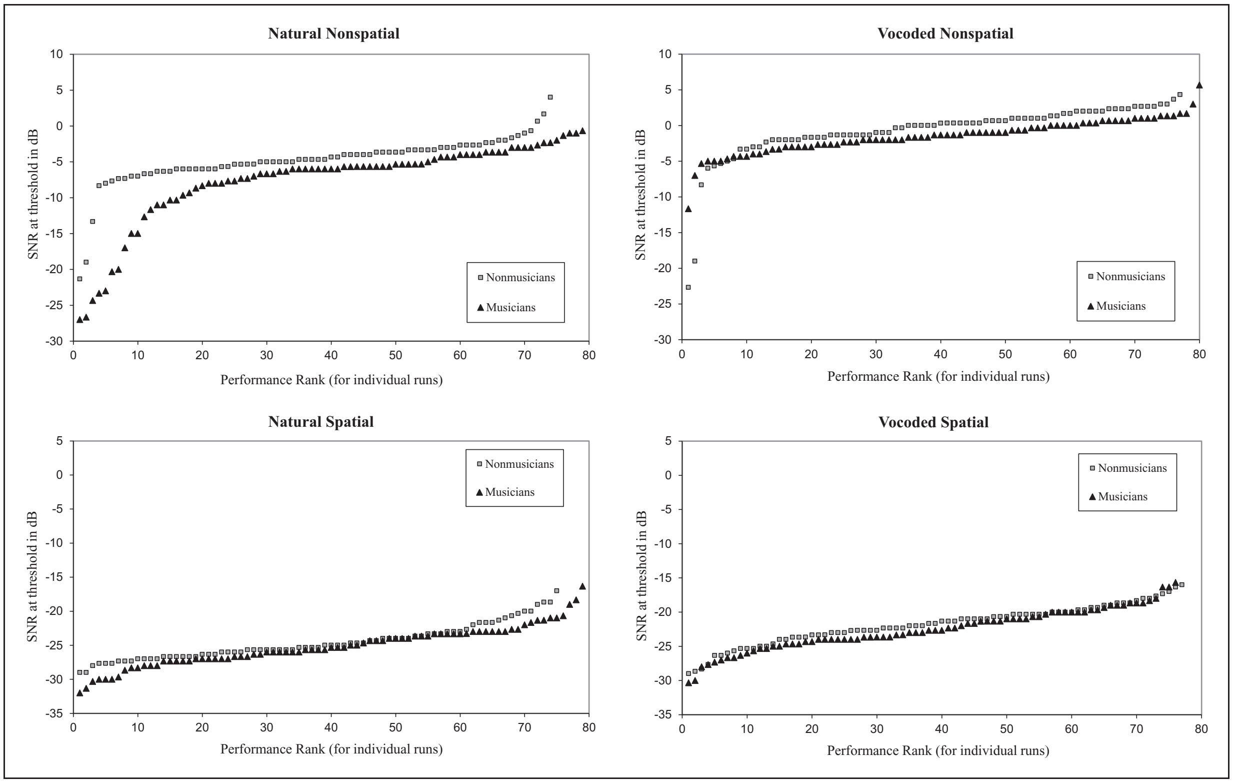

Figure 2. SNR thresholds for each individual adaptive track in each condition. Data are arranged from lowest (best) to highest thresholds in each subject group. Note the 5-dB shift in the *y* axis between upper and lower panels.
SNR = signal-to-noise ratio.

this one nonmusician revealed that the musicians had lower average thresholds than the remaining nonmusicians by a margin of 1.26 dB.

While Figure 2 shows the data for all runs, Figure 3 displays the performance of a subset of top performers among the musicians and nonmusicians. The figure shows the best 20 runs for all subjects in the natural nonspatial condition (including musicians and nonmusicians), wherein each subject is represented by a unique symbol. It can be observed that while there were some "star performers" who substantially overcame informational masking on three or four runs, there were also subjects who did so only on one or two runs. Note M5, a musician subject represented by the open diamond shape. This subject appeared to overcome informational masking on all four adaptive runs, although there was considerable variation in the thresholds. Subject M5's best threshold was −26.67 dB SNR, a true "breakthrough" of informational masking, but M5's other runs showed varying thresholds up to −12.67 dB SNR. NM8 (black squares), one of the two nonmusicians to have breakthrough runs, also showed considerable variation between thresholds. NM8's best performance was a SNR of −21.33 dB, with another breakthrough run of −19 dB SNR. NM8's other two runs (not shown in Figure 3) had SNR thresholds around −7 dB, close to average performance. These examples show that even those subjects who overcame informational masking in the natural nonspatial condition were often not able to do so consistently.

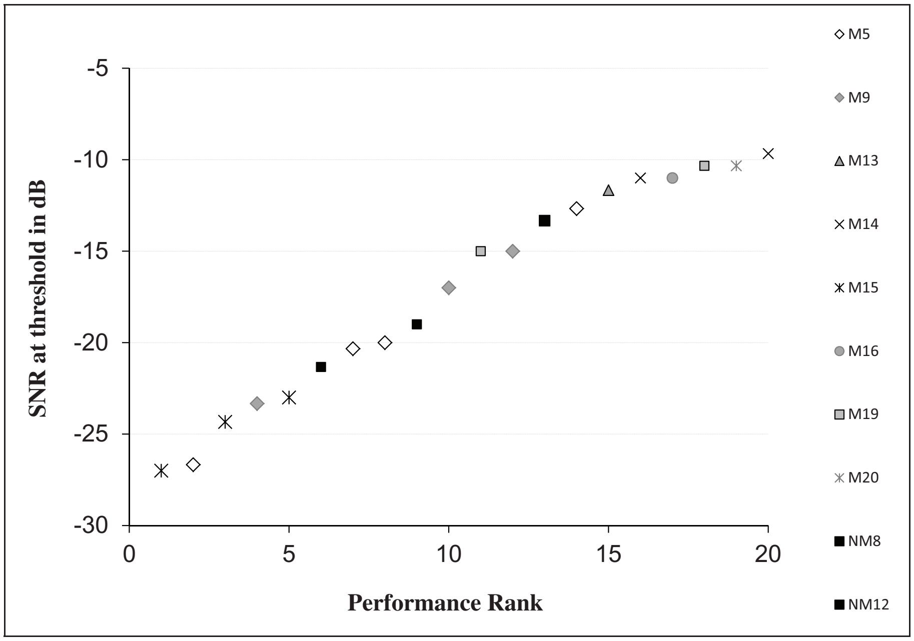

Figure 3. SNR thresholds for top performers in the natural nonspatial condition. The best 20 runs for all subjects are plotted. Data are arranged from lowest (best) threshold to highest threshold for individual runs, with each subject represented with his or her own symbol. Subject codes beginning with "M" denote musician subjects; codes beginning with "NM" denote nonmusician subjects.
SNR = signal-to-noise ratio.

The current study also investigated the relationships between musicians' speech detection performance and the duration of their musical training. Pearson correlation coefficients were conducted on the years of musical training and detection thresholds for each condition. None of these reached significance (Pearson $r < \pm 0.30$, $p > .20$), indicating that the duration of musical training did not explain a significant amount of the variance in performance on this task in any condition. All the musicians who had one or more breakthrough runs in the natural nonspatial condition had studied music for at least 10 years (Figure 4). However, 15 musicians had been playing for more than 10 years, and only 4 of them were able to substantially overcome informational masking. In addition, there did not appear to be a strong relationship between musicians' thresholds and the type of instrument played. The number of musicians who played each type of instrument was small, precluding statistical correlation calculations.

Potential effects of gender on speech detection were examined for all subjects (collapsing across musician status), and among musicians only. No significant gender effects were seen on speech detection performance.

Discussion

This study compared the masked speech detection thresholds of musicians and nonmusicians for natural and vocoded speech stimuli in both spatial and nonspatial conditions. Groups of 20 musicians and 20 nonmusicians detected target words spoken by a female talker within a background of a continuous masker consisting of nonsense sentences spoken by two female talkers. While musicians had lower average thresholds than nonmusicians in all four conditions tested, only in the natural nonspatial condition did the difference between the two groups of participants reach statistical significance. The musicians' average threshold was 3.19 dB lower than that of the nonmusicians in this condition. This difference was largely driven by the data from less than half of the participants in the musician group, who were able to partially overcome the informational masking produced by the two-talker speech masker for some of the four adaptive runs in that particular condition. Of the 20 adaptive runs showing the greatest resolution of informational masking, 17 were from musicians. In the natural spatial condition, it was assumed that informational masking was completely or nearly completely released, and the amount of remaining masking observed (presumably energetic masking) was not significantly different between the two groups.

The most obvious difference from the previous studies on speech-based informational masking reviewed in the Introduction section is the difference in task—detection versus recognition. Other key differences from the articles reviewed also suggest that the current study

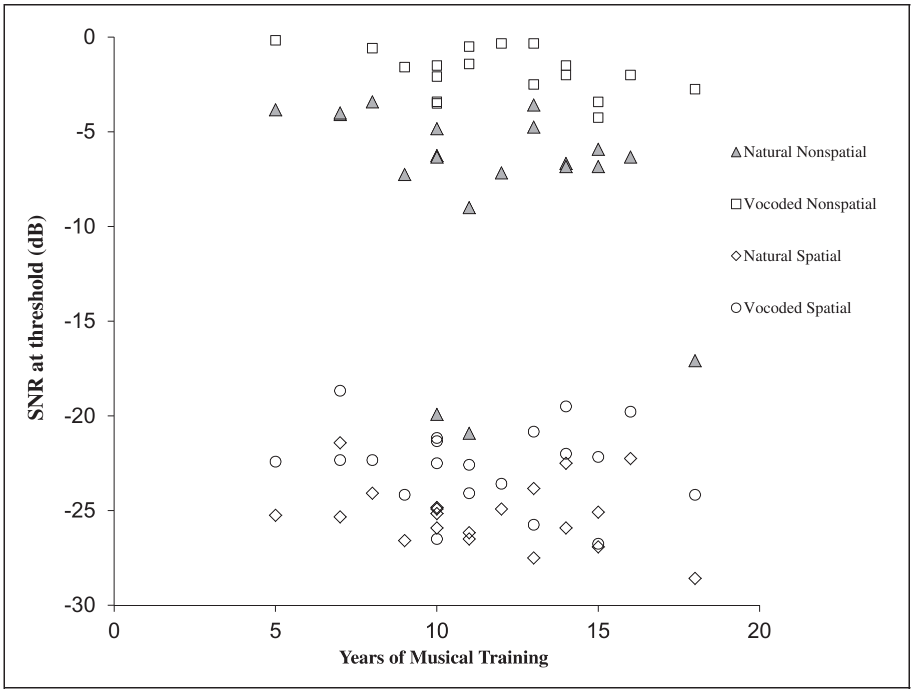

Figure 4. Mean SNR thresholds for musicians arranged by years of musical training.
SNR = signal-to-noise ratio.

explored musical training effects in the informational masking of speech from a considerably different perspective than most previous studies. First, the current conditions were shown to elicit high degrees of informational masking in a co-located target-masker configuration without the constraint of synchronous target and masker stimuli with the same sentence structure, unlike the musician studies of Swaminathan et al. (2015), Clayton et al. (2016), and many other studies on informational masking of other populations in the literature (e.g., Brungart, 2001; Kidd et al., 2016). Also, unlike the substantial reduction in masking caused by masker time reversal found by Swaminathan et al. (2015), the benefit from masker time reversal was essentially zero with the current paradigm (see Balakrishnan & Freyman, 2008). These observations suggest a fundamental difference in the nature of the masking under study.

The finding of better thresholds in musicians on average in the high informational masking condition is consistent with the detection advantage for musicians observed by Oxenham et al. (2003) for nonspeech tonal stimuli. The musicians in the present study may have been using similar types of analytic listening skills to those that are required for detecting tones in the presence of other tones to segregate the target from the masker in the current investigation. There are obvious differences in task and task difficulty between the studies, which could explain why fewer participants in the present study resolved informational masking. Nearly all 12 of the musician subjects tested by Oxenham et al. (2003) were able to overcome informational masking, when compared with a much smaller percentage in the current study.

Task difficulty likely also explains qualitative differences between the current results and those of both Swaminathan et al. (2015) and Clayton et al. (2016). In the current natural speech conditions, the advantage for musicians was much greater in the nonspatial condition, whereas in these two earlier studies cited, the musician advantage for unprocessed masking sentences was much greater in the spatial condition. In the two earlier studies it is apparent that the co-located masking condition was so difficult that neither musicians nor nonmusicians were able to resolve speech-on-speech masking until the SNR was positive (and then all subjects succeeded with a few dB). In that condition, detection appeared to require that the target stand out as being louder than the masker, and musicians were not better at this task than nonmusicians. However, in the conditions where the maskers were slightly and symmetrically displaced, musicians outperformed nonmusicians, possibly indicating superior spatial attention abilities on average.

In the spatial conditions in the current study, the masker was perceived well off to the right and was apparently easy to ignore for all subjects while attending to the front target. Thresholds for spatially separated speech were 17 to 20 dB better than for the nonspatial speech condition, consistent with the differences found by

Balakrishnan and Freyman (2008) and Freyman et al. (2008). This significant effect of spatial configuration has only been found for speech stimuli delivered in the presence of speech maskers, not for speech in the presence of noise maskers. The absence of energetic masking release results from the disruption of binaural and head shadow cues produced by the delayed copy of the masker presented from the front loudspeaker (Freyman et al., 1999). This suggests that the spatial release from masking observed in the current study is due to a release from informational masking, and thus estimates the informational component in the nonspatial configuration to be about 17 to 20 dB.

In the current spatial condition, presumably dominated by energetic masking, thresholds from musicians were less than 1 dB better than those from nonmusicians. This finding is consistent with the small (~1 dB) musician advantage noted on the HINT and QuickSIN tests by Parbery-Clark et al. (2009b). The maskers used in the HINT (speech-shaped noise, test developed by Nilsson, Soli, & Sullivan, 1994) and the QuickSIN (four-talker babble, test developed by Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004) may present less informational masking than the current two-talker babble masker. The effect size observed by Parbery-Clark et al. (2009b) is similar to the small musician benefit shown in the current study for the spatial conditions where energetic masking was dominant, although the current result did not reach statistical significance.

Electrophysiological studies have revealed that musicians have more robust encoding of fundamental frequency and harmonics in the brainstem, more rapid and more accurate responses to timing cues in speech signals, increased activation in Heschl's gyrus, and more robust encoding of speech harmonics in the presence of noise than individuals without musical training (Krizman, Marian, Shook, Skoe, & Kraus, 2012; Parbery-Clark et al., 2009a; Parbery-Clark, Strait, & Kraus, 2011; Schneider et al., 2002; Strait & Kraus, 2014; Zendel & Alain, 2009). Strait and Kraus (2014) also suggested that these subcortical enhancements signal increased cortical control of sensory processing. There could be still other top-down processes at work, such as disparities in verbal and working memory between musicians and nonmusicians (Bialystok & Depape, 2009; Brandler & Rammsayer, 2003; Carey et al., 2015; Chan, Ho, & Cheung, 1998; Clayton et al., 2016). Such subcortical or top-down differences may help explain the small musician advantages seen in the conditions dominated by energetic masking, although they were not addressed in the current study design.

The current study also included vocoded conditions in which both the target words and the masking sentences were transformed into minimally intelligible fluctuating noises with little pitch, pitch variation, or voice quality. The two-talker vocoded masker clearly produced a great deal of masking for the target words in the co-located condition for almost every subject and adaptive run. While it is unknown which acoustic cues were used to detect the presence of the target when pitch cues were unavailable, musicians were slightly better on average than the nonmusicians in using them—although the difference was not statistically significant. Results and interpretations for the spatial vocoded condition (involving largely energetic masking) are essentially the same as for the spatial natural speech condition described earlier. The lack of evidence for a musician advantage for stimuli with limited pitch cues is consistent with similar results for whispered speech by Ruggles et al. (2014). Data from Fuller, Galvin, Maat, Free, and Başkent (2014) showed a small statistically significant musician advantage in only one vocoded condition of eight speech recognition conditions tested, but the overall conclusions were similar to those of Ruggles et al. (2014) and the current study. It also should be noted that neither of these two previous studies found robust musician advantages even in the natural speech conditions tested.

Previous studies (Ho, Cheung, & Chan, 2003; Jackobson, Cuddy, & Kilgour, 2003) found a direct correlation between performance on verbal memory and the duration for which musician subjects played an instrument. Parbery-Clark et al. (2009b) found positive relationships between working memory, QuickSIN performance, and years of musicianship. For the musicians in their study, Clayton et al. (2016) found no relationship between SNR thresholds on their spatial speech-in-noise task with the length or age of onset of musical training. The current study also found no significant effect of duration of musicianship for any of the conditions. In addition, this study found no clear link between type of instrument played and performance among musicians (although a full statistical analysis was not possible due to small sample size).

In summary, the current study investigated speech detection in four conditions, only one of which, the natural nonspatial speech condition, could have been predicted from previous research to show a large benefit for musicians (Başkent & Gaudrain, 2016; Clayton et al., 2016; Oxenham et al., 2003; Swaminathan et al., 2015). Indeed, this condition was the only one of the four to show a sizable and statistically significant advantage for musicians. In the current subject sample, musicians were more likely to have the sharper listening abilities needed to overcome informational masking when pitch cues are available, even when spatial cues are not.

When viewed in the context of previous research on this topic, the current results contribute to what appears to be a common finding in the recent literature: that musicians have an advantage on at least some informational speech masking tasks. It is clear from a comparison of the current and previous findings that the nature of the target and masker stimuli, the degree of spatial separation, and task difficulty all have significant effects on the advantages attributed to musicianship. Future research will hopefully clarify the nature of these effects and advantages more precisely.

Acknowledgments

The authors would like to thank the associate editor for this article, Maria Chait, as well as two anonymous reviewers for their very helpful comments on an earlier draft of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The National Institute on Deafness and Other Communication Disorders R01 DC-01625.

References

Balakrishnan, U., & Freyman, R. L. (2008). Speech detection in spatial and nonspatial speech maskers. *Journal of the Acoustical Society of America*, *123*, 2680–2691.

Başkent, D., & Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *Journal of the Acoustical Society of America*, *139*, EL51–EL56.

Bialystok, E., & Depape, A. M. (2009). Musical expertise, bilingualism, and executive functioning. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 565–574.

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *Journal of the Acoustical Society of America, 137*, 378–387.

Brandler, S., & Rammsayer, T. H. (2003). Differences in mental abilities between musicians and non-musicians. *Psychology of Music*, *31*, 123–138.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*, 1101–1109.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America*, *120*, 4007–4018.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition, 137*, 81–105.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *Journal of the Acoustical Society of America*, *45*, 694–703.

Chan, A. S., Ho, Y. C., & Cheung, M. C. (1998). Music training improves verbal memory. *Nature*, *396*, 128.

Clayton, K. K., Swaminathan, J., Yazdanbakhsh, A., Zuk, J., Patel, A. D., & Kidd, G. Jr. (2016). Executive function, visual attention and the cocktail party problem in musicians and non-musicians. *PloS One, 11*, e0157638.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research*, *352*, 49–69.

Duquesnoy, A. J. (1983). Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. *Journal of the Acoustical Society of America*, *74*, 739–743.

Edmonds, B. A., & Culling, J. F. (2005). The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay. *Journal of the Acoustical Society of America*, *117*, 3069–3078.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, *115*, 2246–2256.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *Journal of the Acoustical Society of America*, *124*, 1627–1637.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, *106*, 3578–3588.

Fuller, C. D., Galvin, J. J. III, Maat, B., Free, R. H., & Başkent, D. (2014). The musician effect: Does it persist under degraded pitch conditions of cochlear implant simulations? *Frontiers in Neuroscience*, *8*, 1–16.

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

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *40*, 432–443.

Ho, Y. C., Cheung, M. C., & Chan, A. S. (2003). Music training improves verbal but not visual memory: Cross-sectional and longitudinal explorations in children. *Neuropsychology*, *17*, 439–450.

Jackobson, L. S., Cuddy, L. L., & Kilgour, A. R. (2003). Time tagging: A key to musicians' superior memory. *Music Perception*, *20*, 307–313.

Kidd, G. Jr., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *Journal of the Acoustical Society of America*, *124*, 3793–3802.

Kidd, G. Jr., Mason, C. R., & Arbogast, T. L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *Journal of the Acoustical Society of America*, *111*, 1367–1376.

Kidd, G. Jr., Mason, C. R., Brughera, A., & Hartmann, W. M. (2005). The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica United with Acustica*, *91*, 526–536.

Kidd, G. Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *Journal of the Acoustical Society of America*, *95*(6), 3475–3480.

Kidd, G. Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *Journal of the Acoustical Society of America, 140*, 132–144.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *116*, 2395–2405.

Kishon-Rabin, L., Amir, O., Vexler, Y., & Zaltz, Y. (2001). Pitch discrimination: Are professional musicians better than non-musicians? *Journal of Basic and Clinical Physiology and Pharmacology*, *12*, 125–143.

Kraus, N., & Nicol, T. (2010). The musician's auditory world. *Acoustics Today*, *6*, 15–27.

Krizman, J., Marian, V., Shook, A., Skoe, E., & Kraus, N. (2012). Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. *Proceedings of the National Academy of Sciences*, *109*, 7877–7881.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*(Suppl 2), 467–477.

Madsen, S. M., Whiteford, K. L., & Oxenham, A. J. (2017). Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. *Scientific Reports*, *7*, 12624.

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*, 36–47.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, *95*, 1085–1099.

Oxenham, A. J., Fligor, B. J., Mason, C. R., & Kidd, G. Jr. (2003). Informational masking and musical training. *Journal of the Acoustical Society of America*, *114*, 1543–1549.

Parbery-Clark, A., Skoe, E., & Kraus, N. (2009a). Musical experience limits the degradative effects of background noise on the neural processing of sound. *Journal of Neuroscience*, *29*, 14100–14107.

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009b). Musician enhancement for speech-in-noise. *Ear and Hearing*, *30*, 653–661.

Parbery-Clark, A., Strait, D. L., & Kraus, N. (2011). Context-dependent encoding in the auditory brainstem subserves enhanced speech-in-noise perception in musicians. *Neuropsychologia*, *49*, 3338–3345.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *Journal of the Acoustical Society of America*, *114*, 446–454.

Ruggles, D. R., Freyman, R. L., & Oxenham, A. J. (2014). Influence of musical training on understanding voiced and whispered speech in noise. *PloS One*, *9*, e86980.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl s gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience, 5*, 688–694.

Strait, D. L., & Kraus, N. (2014). Biological impact of auditory expertise across the life span: Musicians as a model of auditory learning. *Hearing Research*, *308*, 109–121.

Strait, D. L., Kraus, N., Parbery-Clark, A., & Ashley, R. (2010). Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance. *Hearing Research*, *261*, 22–29.

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V., Kidd, G. Jr., & Patel, A. D. (2015). Musical training, individual differences and the cocktail party problem. *Scientific Reports, 5*, 11628.

Zendel, B. R., & Alain, C. (2009). Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience*, *21*, 1488–1498.