JASA ARTICLE

Differential benefits of unmasking extended high-frequency content of target or background speech[a)]

Brian B. Monson,[1,b)] Rohit M. Ananthanarayana,[1] Allison Trine,[1] Vahid Delaram,[1] G. Christopher Stecker,[2] and Emily Buss[3]

[1]*Department of Speech and Hearing Science, University of Illinois Urbana-Champaign, Champaign, Illinois 61820, USA*

[2]*Spatial Hearing Laboratory, Boys Town National Research Hospital, Omaha, Nebraska 68131, USA*

[3]*Department of Otolaryngology/HNS, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA*

ABSTRACT:
Current evidence supports the contribution of extended high frequencies (EHFs; >8 kHz) to speech recognition, especially for speech-in-speech scenarios. However, it is unclear whether the benefit of EHFs is due to phonetic information in the EHF band, EHF cues to access phonetic information at lower frequencies, talker segregation cues, or some other mechanism. This study investigated the mechanisms of benefit derived from a mismatch in EHF content between target and masker talkers for speech-in-speech recognition. EHF mismatches were generated using full band (FB) speech and speech low-pass filtered at 8 kHz. Four filtering combinations with independently filtered target and masker speech were used to create two EHF-matched and two EHF-mismatched conditions for one- and two-talker maskers. Performance was best with the FB target and the low-pass masker in both one- and two-talker masker conditions, but the effect was larger for the two-talker masker. No benefit of an EHF mismatch was observed for the low-pass filtered target. A word-by-word analysis indicated higher recognition odds with increasing EHF energy level in the target word. These findings suggest that the audibility of target EHFs provides target phonetic information or target segregation and selective attention cues, but that the audibility of masker EHFs does not confer any segregation benefit. https://doi.org/10.1121/10.0020175

(Received 6 February 2023; revised 14 June 2023; accepted 29 June 2023; published online 25 July 2023)

[Editor: Robert A. Fox] Pages: 454–462

I. INTRODUCTION

Human speech contains audible acoustic energy at frequencies above 13 kHz (Monson and Caravello, 2019; Monson *et al.*, 2014). Recent data indicate that speech energy at frequencies above 8 kHz, or extended high-frequency (EHF) energy, provides acoustic cues useful for speech recognition in complex listening environments. EHF cues appear particularly useful for speech recognition when maskers have attenuated levels of EHF energy relative to the target. This utility has been demonstrated using natural EHF attenuation that occurs for real-world speech-in-speech (i.e., cocktail party) scenarios in which masker talkers are not facing the listener (Braza *et al.*, 2022; Flaherty *et al.*, 2021; Monson *et al.*, 2019; Trine and Monson, 2020). Masker EHF levels are attenuated in this scenario due to the greater directionality of higher-frequency speech components (i.e., the rotating of a masker talker's head away from the listener has the effect of low-pass filtering the masker talker's speech; Chu and Warnock, 2002; Kocon and Monson, 2018; Monson *et al.*, 2012), leading to reduced energetic masking of EHF cues in the target speech.

The utility of EHF cues has also been demonstrated using artificial EHF attenuation with synthetic maskers by low-pass filtering steady-state noise maskers at 8 kHz, again unmasking EHF cues in the target speech (Motlagh Zadeh *et al.*, 2019; Polspoel *et al.*, 2022). Polspoel *et al.* (2022) additionally showed that EHF cues were beneficial even in the presence of broadband noise maskers with unattenuated EHFs. In contrast, other studies have shown that EHF cues *per se* did not provide a measurable benefit for speech-in-speech when masker EHF levels were unattenuated relative to target EHF levels (i.e., target and masker talkers were all simulated to face the listener; Levy *et al.*, 2015; Moore *et al.*, 2010). However, EHF cues in those studies were restricted to the 7.5–10 or 8–10 kHz band because speech stimuli were band limited at 10 kHz. Whether the full complement of EHF cues (8–20 kHz band) is useful for speech-in-speech recognition when masker talker EHF levels are unattenuated (e.g., both masker and target talkers face the listener) has not yet been directly tested (although see Braza *et al.*, 2022), but the available data demonstrate that EHF cues are useful in the presence of EHF-attenuated speech or noise maskers and for full-band (FB) steady-state noise maskers.

The benefit of EHFs for speech-in-speech recognition raises questions regarding the nature of EHF cues. Does EHF energy provide useful phonetic information under speech-in-speech conditions, or does EHF energy provide

[a)]This paper is part of a special issue on Perception and Production of Sounds in the High-Frequency Range of Speech.

[b)]Also at: Department of Biomedical and Translational Sciences, Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA.

segregation cues that enhance a listener's ability to use low-frequency information? This question prompted the present study. It has been demonstrated that EHFs in speech convey some phonetic information, useful for vowel and consonant identification when lower-frequency energy is partially or entirely absent (Lippmann, 1996; Vitela *et al.*, 2015), or when extended bandwidth hearing aids restore EHF audibility (Seeto and Searchfield, 2018). However, whether this EHF phonetic information is useful in the presence of real-world speech maskers is unclear. On the other hand, EHFs could support talker segregation. For example, talker-specific differences in EHF level (Monson and Buss, 2022), EHF spectral tilt, or EHF spectral peaks (Maniwa *et al.*, 2009) could provide talker segregation cues. Alternatively, because speech energy at high frequencies is at least partly temporally coherent with low-frequency speech energy (Crouzet and Ainsworth, 2001) (see also Fig. 2), it may be that access to EHF energy from the target speech facilitates segregation of and selective attention to low-frequency phonetic information. This could occur because the temporal coherence of spectral features common to a given sound enables those features to be grouped together, improving stream segregation (Shamma *et al.*, 2011).

Trine and Monson (2020), demonstrated that temporal envelope cues in the EHF band (8–20 kHz) accounted for much of the observed EHF benefit. Adding a single-channel noise-vocoded EHF band to speech low-pass filtered at 8 kHz (LP8k) led to improvement in speech-in-speech recognition compared to speech LP8k alone. However, including fine spectral detail at EHFs (i.e., FB speech) provided an additional, albeit smaller, benefit. Although those data were in line with the hypothesis that both segregation (temporal) cues and phonetic (fine spectral detail) information are provided by EHFs, it is possible that the temporal envelope (e.g., EHF energy bursts) provided phonetic cues and/or that fine spectral detail (e.g., EHF spectral tilt or EHF spectral peaks) provided talker segregation cues.

In our previous studies on EHF cues two-talker maskers were used. Speech reception thresholds tend to be poorer for two-talker maskers than for either speech-shaped noise or one-talker maskers (Freyman *et al.*, 2004; Rosen *et al.*, 2013). There are several features of two-talker speech that could contribute to its greater effectiveness as a masker. The perceptual similarity between target speech and a speech masker creates informational masking, where the acoustic features necessary to recognize target speech are encoded at the periphery, but the listener is not able to segregate the multiple streams of speech and selectively attend to the target (reviewed by Bronkhorst, 2015). Whereas a one-talker masker typically contains sizeable pauses and spectral gaps, two-talker speech is spectro-temporally denser, providing fewer opportunities to glimpse target speech. Two-talker speech is often thought to be a stronger informational masker than one-talker speech (Freyman *et al.*, 2004), but it is also possible that energetic masking plays a role in the greater masking observed for two-talker speech. Segregating the target from two streams of maskers could also be inherently more challenging than with one masker talker. If the contributions of EHF content are most evident under conditions with greater informational and/or energetic masking, then we would predict larger effects for the two-talker masker than the one-talker masker.

Relevant to the audibility of EHF cues, several studies have reported a relationship between EHF pure-tone audiometric thresholds and speech-in-noise or speech-in-speech recognition performance for listeners with clinically normal hearing at standard audiometric frequencies (Braza *et al.*, 2022; Mishra *et al.*, 2021; Motlagh Zadeh *et al.*, 2019; Polspoel *et al.*, 2022; Trine and Monson 2020; Yeend *et al.*, 2019). Multiple mechanisms have been proposed to explain the association between EHF thresholds and speech-in-noise performance (Hunter *et al.*, 2020; Lough and Plack, 2022). One potential mechanism is that the loss of audibility of EHF cues in speech degrades performance. Because EHF cues confer benefit for speech-in-noise recognition, it stands to reason that hearing sensitivity at EHFs would be correlated with speech recognition in complex environments where low frequencies are masked but EHFs are not (Braza *et al.*, 2022; Motlagh Zadeh *et al.*, 2019; Polspoel *et al.*, 2022; Trine and Monson, 2020). Another possibility is that elevated EHF thresholds indicate subclinical hearing impairment at lower frequencies. Evidence for this possibility is that EHF thresholds are associated with speech-in-noise recognition even when EHF cues are *not* available in the speech stimuli (Ananthanarayana *et al.*, 2022; Mishra *et al.*, 2021). A third possibility is that elevated EHF thresholds may be indicative of temporal processing or other supra-threshold deficits at EHFs unrelated to audibility of EHF cues (Mishra *et al.*, 2023).

Most studies demonstrating an association between EHF thresholds and speech recognition tested listeners with large variability in thresholds at EHFs, including listeners with substantial EHF hearing loss (Braza *et al.*, 2022; Mishra *et al.*, 2021; Motlagh Zadeh *et al.*, 2019). However, Trine and Monson (2020) previously found a moderate but significant correlation ($r = 0.34$) between the 16-kHz pure-tone threshold and speech-in-speech recognition for a group of young, normal-hearing listeners who all had relatively good EHF thresholds (<25 dB hearing level, HL). This finding is of potential importance because it indicates the 16-kHz pure-tone threshold may be a good predictor of masked speech recognition, even for young normal-hearing listeners with limited variability in EHF thresholds.

With these considerations, the purpose of the present study was to further examine the mechanism of benefit from a mismatch in EHF energy between target and masker speech. If EHFs provide talker segregation cues associated with talker-specific differences in EHF content, we reasoned that creating a substantial talker mismatch in EHF energy level should be sufficient to confer the EHF benefit, regardless of whether the target or masker has greater EHF energy. This mismatch can be created by an independently low-pass filtering target or masker at 8 kHz, similar to the procedure used by Polspoel *et al.* (2022). However, if EHFs provide

either phonetic information or temporal coherence cues to access lower-frequency phonetic information, then the EHF benefit should be greater when the masker is low-pass filtered, unmasking EHF cues available in the target speech, as compared to when the target is low-pass filtered. Polspoel *et al.* (2022) tested three filtering combinations for target speech and steady-state noise maskers using FB stimuli or stimuli LP8k: FB speech with FB noise (filter matched), FB speech with LP8k noise (filter mismatched), and LP8k speech with LP8k noise (filter matched). They demonstrated that the best performance was achieved with FB speech and LP8k noise, followed by FB speech with FB noise, and then LP8k speech with LP8k noise. However, their data did not include the fourth combination (LP8k speech with FB noise), which would test whether this inverted EHF mismatch would influence performance. Furthermore, it is unclear whether the results of Polspoel *et al.* (2022) would generalize to speech maskers.

In the present study, we tested speech-in-speech recognition using four filtering combinations of FB and LP8k speech: (1) FB target, FB masker (filter matched); (2) FB target, LP8k masker (filter mismatched); (3) LP8k target, LP8k masker (filter matched); (4) LP8k target, FB masker (filter mismatched). We hypothesized that EHF energy differences provide talker segregation cues, predicting better performance in filter-mismatched conditions (2, 4) than in corresponding filter-matched conditions (1, 3). We also hypothesized that EHFs provide phonetic information directly or indirectly (through temporal coherence with low frequencies), predicting better performance in conditions with the FB target (1, 2) than in corresponding conditions with the LP8k target (3, 4). We tested whether performance in these conditions was affected by the number of masker talkers (one vs two), hypothesizing that listeners would perform better with a single masker talker overall, but that effects of EHF content would be greater for the two-talker masker. Finally, we tested whether pure-tone thresholds at 16 kHz predicted performance in our speech-in-speech task.

II. METHODS

A. Participants

A total of 39 participants (27 females, 11 males, one other), ages 18–26 years (mean 21.2 years), were tested in this experiment. Sample size was determined based on a power analysis to achieve 80% power to detect anticipated effect sizes. Participants had thresholds of ≤25 dB HL in at least one ear from 0.5 to 12.5 kHz. Four participants had better ear thresholds >25 dB HL at the extended high frequencies of 14 and/or 16 kHz. Figure 1 shows better-ear thresholds as a function of frequency. Pure tone audiometry was conducted with a GSI Audiostar Pro and RadioEar DD450 (GSI, Eden Prairie, MN) circumaural headphones for standard audiometric frequencies (0.5–8 kHz) and the extended high frequencies of 9, 10, 11.2, 12.5, 14, and 16 kHz, using the modified Hughson-Westlake procedure (Carhart and Jerger, 1959). All participants were native English speakers.

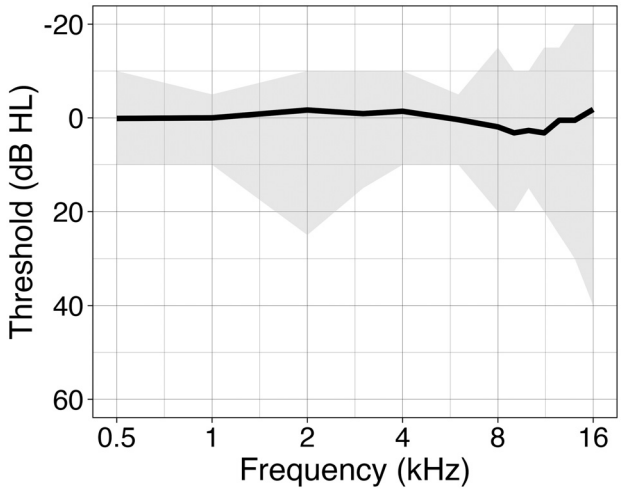

FIG. 1. Mean better ear pure tone thresholds for subjects at standard audiometric frequencies (0.5–8 kHz) and extended high frequencies (9–16 kHz). The shaded region depicts the maximum and minimum responses across participants.

B. Stimuli

The masker stimulus was either one- or two-female-talker babble. The two masker talkers had mean fundamental frequencies ($F0$s) of 229 Hz and 225 Hz (Monson *et al.*, 2012). Target speech stimuli were the Bamford-Kowal-Bench sentences (Bench *et al.*, 1979) recorded by a single female talker with a mean $F0$ of 235 Hz and a speaking rate of 3.2 words per second. Both masker and target stimuli were recorded at 44.1 kHz with 16-bit precision, with a microphone directly in front of the talker; when presented over loudspeakers, this simulates talkers directly facing the listener. A comparison of stimulus spectra is shown in Fig. 2. Stimuli were low-pass filtered with a 32-pole Butterworth filter with a cutoff frequency at either 20 or 8 kHz (Fig. 3). These conditions are described FB and LP8k, respectively. A total of four filtering conditions were tested: two matched conditions where both target and masker were either low-pass filtered at 20 kHz (FB Target, FB Masker) or 8 kHz (LP8k Target, LP8k Masker), and two mismatched

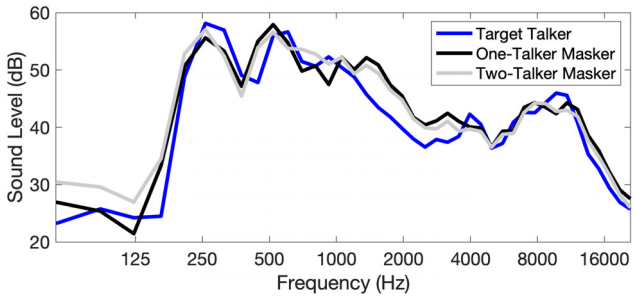

FIG. 2. (Color online) Long-term average spectra of the target talker, one-talker masker, and two-talker masker. The overall level for each stimulus was set to 65 dB SPL.

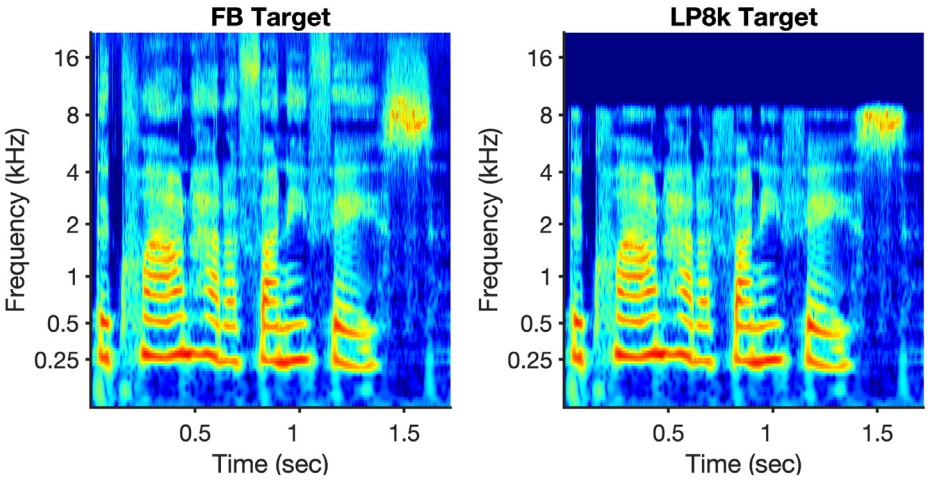

FIG. 3. (Color online) Cochleogram of the sentence "The clown had a funny face," showing the spectral content of the FB target signal (left) and the LP8k target signal (right).

conditions where either the target or masker was FB and the other was not (FB Target, LP8k Masker or LP8k Target, FB Masker).

C. Procedure

Stimuli were presented in a double-walled sound booth using a single KRK Rokit 8 G3 loudspeaker (KRK, Nashville, TN) placed 1 m in front of the listener, thus target and masker were co-located. The level of the masker was set to 65 dB sound pressure level (SPL) at 1 m. The level of the target stimulus was varied using two interleaved adaptive tracks, each using a one-down, one-up adaptive rule, following methods described by Sobon *et al.* (2019). For the first track, when one or more words (out of three to five target words) were correctly repeated, the signal-to-noise ratio (SNR) was decreased; otherwise, the SNR increased. For the second track, when all target words or all but one target word was repeated correctly, the SNR decreased; otherwise, the SNR increased. Both tracks had the same initial starting level of 4 dB SNR; level adjustments were made in steps of 4 dB prior to the first reversal, and in steps of 2 dB thereafter. Each of the two tracks comprised 16 sentences, for a total of 32 sentences in each condition. Percent correct word recognition as a function of SNR was fitted using a logit function with asymptotes at 0% and 100% correct,

$$y = \frac{1}{1 + \exp\left(-\frac{(x-\alpha)}{\beta}\right)}, \tag{1}$$

where y is the percent correct, x is the SNR, and α and β are the fit parameters. Fits were made by minimizing the sum of squared error, weighted by the number of words at each SNR. The resulting functions for each listener in each condition were used to estimate speech reception thresholds (SRTs), defined as the SNR required to produce 50% correct performance. Data fits accounting for $< 50\%$ of variance were removed from further analysis ($n = 14$, accounting for 4% of runs). Final data fits were associated with r^2 values ranging from 0.54 to 0.98, with a median value of 0.82.

Custom scripts written in MATLAB (The MathWorks Inc., Natick, MA) were used for signal processing and experimental control. Following a training block (FB target, FB two-talker masker) consisting of 16 sentences, the eight conditions (four filtering conditions × two masker conditions) were tested in separate blocks, with block order randomized across participants.

D. Analysis

Data were analyzed in R (R Core Team, 2022). Linear mixed-effects models, with a random intercept for a subject, were created using the function *nlme* (Pinheiro *et al.*, 2022). These models were used to examine the influence of low-pass filtered target, filtering mismatch, and number of masker talkers on the SRT, as well as the influence of condition on the slope of the psychometric function fits. Exploratory analyses examined the potential contribution of average or better-ear pure-tone extended high-frequency thresholds on performance in the two-talker-masker FB Target conditions. Thresholds at 16 kHz were selected for this analysis due to their use in previous studies (Braza *et al.*, 2022; Trine and Monson, 2020).

A *post hoc* analysis was carried out to determine the relationship between energy level in the EHF band and word recognition in the FB Target conditions with the two-talker masker. A mixed-effects logistic regression model was built using *lme4* (Bates *et al.*, 2015). Each target keyword was a data point, and the outcome was a dichotomous variable indicating whether the word was correctly identified. To estimate word-by-word EHF levels, onset and offset times for each keyword were first obtained using IBM Watson® Speech-to-Text (IBM, 2022) and the speech2text function (MathWorks Audio Toolbox Team, 2022). This analysis evaluates speech recordings and returns text transcription, including approximate onset and offset times for each word that is recognized. Manual editing was required for 3% of keywords, due to misidentification of keywords. Next, the target sentence was bandpass filtered between 8 and 20 kHz to isolate the EHF band. The EHF level was computed in dB for each keyword.

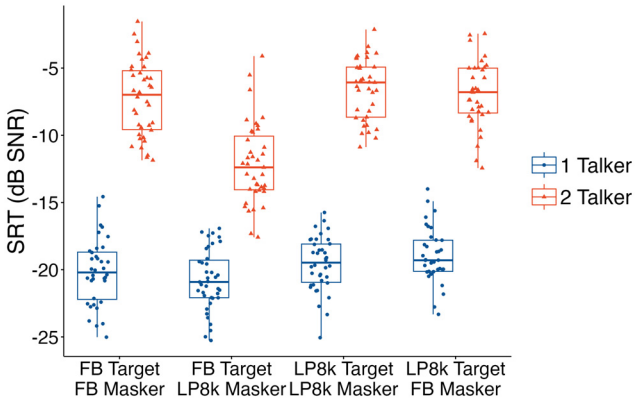

FIG. 4. (Color online) Distributions of SRT values in dB SNR for each of the four filtering conditions, indicated on the abscissa. Color reflects the number of masker talkers, as defined in the legend. SRT, speech reception threshold.

III. RESULTS

The distributions of SRT values are plotted for each condition in Fig. 4, and mean SRTs are reported in Table I. The mean SRT is consistently lower for the one-talker masker than the two-talker masker, with masker effects ranging from 8.7 dB (FB Target, LP8k Masker) to 12.9 dB (FB Target, FB Masker). With the one-talker masker, mean SRTs were similar for the four conditions, with values of –19.0 (LP8k Target, FB Masker) to –20.8 dB SNR (FB Target, LP8k Masker). In contrast, SRTs differed by up to 5.5 dB across conditions with the two-talker masker. With the two-talker masker, the mean SRT for the FB Target, LP8k Masker condition was –12.1 dB SNR; SRTs in the remaining three conditions were similar, with values of –7.3 (FB Target, FB Masker) to –6.6 dB SNR (LP8k Target, LP8k Masker; see Table I).

These observations were evaluated using a linear mixed model with fixed effects of mismatch between target and masker filters, number of masker talkers, and EHF content in target (LP8k or FB), as well as all two- and three-way interactions (Table II). This model indicates a significant effect of the number of masker talkers, a two-way interaction of mismatch and LP8k target content, and a three-way interaction between mismatch, number of masker talkers, and LP8k target content. Compared to the baseline condition with FB target and FB one-talker masker, there was no significant improvement by either introducing a target-masker mismatch or introducing LP8k target content. The interaction between mismatch and LP8k target content reflects a beneficial effect of mismatch in EHF content for the FB target and detrimental effects for the LP8k target in the data for the one-talker masker. The three-way interaction supports the observation that SRTs with the two-talker masker for the FB Target, LP8k Masker condition were lower (better) than expected based on the lower order effects. These results support the observation that a mismatch in EHF content only improved performance for the FB Target conditions, with larger effects for the two-talker masker than the one-talker masker.

TABLE I. Mean SRT values in dB SNR observed in each of the eight conditions.

	SRT (dB SNR)	
Target Condition Masker Condition	***One-talker***	***Two-talker***
FB Target FB Masker	−20.2 (±2.5)	−7.3 (±2.8)
FB Target LP8k Masker	−20.8 (±2.2)	−12.1 (±3.0)
LP8k Target LP8k Masker	−19.6 (±2.0)	−6.6 (±2.2)
LP8k Target FB Masker	−19.0 (±2.0)	−7.0 (±2.4)

A second linear mixed model with the same fixed and random effects was applied to the slopes of the psychometric function fits for data from each participant in each of the eight conditions. This analysis did not indicate any significant differences in slopes across conditions. None of the fixed effects or interactions reached significance. The mean slope was 3.63, corresponding to a change in performance of approximately 6.3 percentage points for every 1-dB change in SNR between 25% to 75% correct performance.

Given the very low SRTs for the one-talker masker, it is possible that target EHFs were inaudible at threshold for the FB target conditions. SRTs of –20 dB result in a target speech level of 45 dB SPL, resulting in target EHF levels that are approximately 25 dB SPL (see Fig. 2). This might have caused a floor effect, reducing the effect of low-pass filtering for the one-talker conditions. To test this possibility, we ran a secondary analysis defining SRT at the 80%-correct SNR for the one-talker masker conditions and at the 20%-correct SNR for the two-talker masker conditions. These points of the psychometric function were selected because the target speech level for the 80%-correct SNR for the one-talker FB Target, FB Masker condition was similar (within 2 dB) to the 20%-correct SNR for the two-talker FB Target, FB Masker condition, and it was ∼6 dB higher than the target speech level for the 50%-correct SNR for the one-talker FB Target, FB Masker condition. These new values of SRT were estimated for each listener using the psychometric function generated for each listener in each condition. A linear mixed-effects model evaluating the same fixed and random effects was conducted (Table III). Significant factors were similar to those of the original model, with the addition of a significant main effect of mismatch compared to the baseline condition of FB target and FB one-talker masker. These results support the idea that low-pass filtering the one-talker masker improves performance, but that this benefit is greater for the two-talker masker.

Lower SRTs for the FB Target, LP8k Masker condition compared to the FB Target, FB Masker condition could reflect greater access to EHF speech cues when the masker is LP8k. Two exploratory analyses were carried out to confirm this interpretation. The first evaluated individual differences in EHF sensitivity and SRTs with the two-talker masker for the FB Target, LP8k Masker condition, across participants. EHF sensitivity was characterized as the average or better ear 16 kHz pure-tone thresholds. The expectation was that participants with better EHF sensitivity might experience greater benefit when the masker was low-pass

https://doi.org/10.1121/10.0020175

TABLE II. Results of a linear mixed model evaluating effects of mismatch between target and masker filters, number of masker talkers, and EHF content in target speech, on SRTs. The reference condition was the one-talker masker, with the FB target and FB masker.

	Estimate	*Standard error (SE)*	*t-value*	*p*
(Intercept)	−20.09	0.40	−50.23	<0.001
LP8k Target	0.55	0.45	1.24	0.217
Mismatch	−0.69	0.44	−1.53	0.126
Two-talker Masker	12.87	0.45	28.89	**<0.001**
*LP8k Target * Mismatch*	1.30	0.63	2.08	**0.038**
*LP8k Target * Two-talker Masker*	0.02	0.63	0.03	0.975
*Mismatch * Two-talker Masker*	−4.23	0.62	−6.81	**<0.001**
*LP8k Target * Mismatch * Two-talker Masker*	3.41	0.88	3.86	**<0.001**

filtered. Contrary to this prediction, SRTs did not correlate with either average ($r = -0.21$, $p = 0.199$) or better ear ($r = 0.02$, $p = 0.917$) 16-kHz thresholds. Similarly, the EHF benefit, defined as the difference between the FB Target, LP8k Masker condition and the LP8k Target, LP8k Masker condition, showed no significant correlation with 16-kHz thresholds ($r = 0.01$, $p = 0.97$). One limitation of this analysis is the fact that only five of the 39 participants (13%) had average 16-kHz thresholds greater than 20 dB HL.

The second exploratory analysis evaluated the association between EHF content of each target word and the probability of getting that word correct for the FB Target, LP8k Masker and FB Target, FB Masker conditions with the two-talker masker. The prediction was that target words with greater EHF content would benefit more from LP8k filtering the masker as compared to targets with less EHF content. EHF levels across target words spanned a wide range from roughly 7 to 66 dB SPL, with a standard deviation of 10 dB around mean values of 32.3 and 36.0 dB SPL, respectively, in the LP8k Masker and FB Masker conditions.

This analysis used logistic regression with fixed effects that included the global SNR for each sentence, masker condition, and target-word EHF level along with its interaction with masker condition. Subject and sentence were used as random effects to account for correlations within subjects and within sentences. The values of SNR and EHF levels were both centered around their means to improve model convergence. We note that SNR captures the target-to-masker ratio for the overall sentence and not individual words; word-by-word masker level was unavailable since the randomly chosen masker sample for each trial was not saved during run-time.

TABLE III. Results of a linear mixed model evaluating effects of mismatch between target and masker filters, number of masker talkers, and EHF content in target speech, on SRTs defined at 80% correct for the one-talker masker and at 20% correct for the two-talker masker. The reference condition was the one-talker masker, with the FB target and FB masker.

	Estimate	*SE*	*t-value*	*p*
(Intercept)	−14.43	0.67	75.49	**<0.001**
LP8k Target	−0.02	0.79	−0.02	0.981
Mismatch	−1.79	0.79	−2.27	**0.024**
Two-talker Masker	1.79	0.79	2.27	**0.024**
*LP8k Target * Mismatch*	2.30	1.12	2.06	**0.041**
*LP8k Target * Two-talker Masker*	0.31	1.12	0.28	0.779
*Mismatch * Two-talker Masker*	−2.35	1.12	−2.10	**0.037**
*LP8k Target * Mismatch * Two-talker Masker*	0.62	1.58	0.39	0.695

TABLE IV. Logistic regression model for word-level analysis with the estimated odds ratios.

	Odds Ratios	*CI*	*p*
(Intercept)	2.808	2.250–3.060	**<0.001**
SNR	1.276	1.256–1.297	**<0.001**
FB Masker	0.277	0.244–0.315	**<0.001**
Target EHF Level	1.031	1.021–1.041	**<0.001**
*FB Masker * Target EHF Level*	0.985	0.973–0.996	**0.010**

All fixed effects were significant (Table IV). In the baseline condition with low-pass filtered masker (FB Target, LP8k Masker), a 1-dB increase in SNR at a fixed EHF level increased the odds of word recognition by about 28% (95% CI [1.26, 1.30], $p < 0.001$), while a 1-dB increase in EHF level at a fixed SNR improved the odds by about 3% (95% CI [1.02, 1.04], $p < 0.001$). Increasing the masker bandwidth (FB Target, FB Masker) led to a large drop of 72% (95% CI [0.24, 0.32], $p < 0.001$) in the odds, consistent with the results of the linear mixed model. The effect of target EHF level was reduced in this condition (95% CI [0.973, 0.996], $p = 0.01$), with a 1-dB increase in EHF level giving an improvement of only about 1% in the odds. These results demonstrate that the EHF content of each word has a greater effect on recognition when the masker is low-pass filtered than when it is full bandwidth, as would be expected if the benefit of filtering the masker was related to a reduction in energetic masking.

IV. DISCUSSION

The primary question motivating the present study was whether the EHF benefit for speech-in-speech recognition can be attributed to talker segregation cues associated with talker-specific differences in EHF content. One possibility considered at the outset of this study was that a mismatch in EHF energy might support improved segregation, leading to better SRTs in both mismatch conditions. This was not the case. We observed that a large mismatch in EHF level between target and masker speech substantially improved speech recognition performance only when the target contained EHFs. This effect was somewhat greater for the two-talker masker.

Mismatches in EHF content being beneficial only for the FB target could indicate that listeners take advantage of reduced energetic masking in the EHF band to make use of phonetic cues in that frequency region. Indeed, words that had higher levels of EHFs were more likely to be recognized in both FB and LP8k two-talker speech maskers. These

findings are in line with previous data that indicate EHFs convey phonetic information for both consonants and vowels (Lippmann, 1996; Monson *et al.*, 2014; Vitela *et al.*, 2015). This possibility is notable because, apart from studies on voiceless fricatives (Jongman *et al.*, 2000; Maniwa *et al.*, 2009; Monson *et al.*, 2012; Shadle and Mair, 1996), relatively little is understood about phoneme-specific differences in acoustic structure at EHFs. Phoneme-specific EHF characteristics may distinguish other phonemes besides voiceless fricatives.

This possibility leads to an alternative interpretation of the data from Trine and Monson (2020). In that study, the benefit of including a vocoded band of speech in the EHF region was interpreted as providing primarily segregation cues. However, it is also likely that a single-channel noise-vocoded EHF band provides phonetic information, albeit somewhat degraded. For example, bursts of EHF energy occur for voiceless fricatives such as /s/ or /f/ (see Fig. 3). Although these phonemes have characteristic spectral slopes at EHFs (Monson *et al.*, 2012) that would be modified by noise-vocoding the EHF band, it may be that listeners still associate such vocoded noise bursts with fricatives because other phonemes are not associated with sustained EHF noise bursts at high sound levels. However, EHF spectral detail afforded listeners in that study additional benefit for speech-in-speech recognition beyond that provided by EHF level alone, suggesting that EHF acoustic characteristics for speech may be worth further investigation.

An alternative interpretation of the present data is that EHF target information helps listeners segregate and selectively attend to phonetic information *across* the speech spectrum by, for example, taking advantage of temporal coherence of target EHF information and low-frequency information. This interpretation is also corroborated by Trine and Monson (2020). With this interpretation, the association we observed here between higher target EHF level and improved word recognition could be explained by greater audibility of these temporally coherent EHF cues. Thus, although the present data appear to rule out the possibility that talker-specific differences at EHFs facilitate stream segregation, they do not definitively favor either of the other interpretations considered here.

An additional possibility not previously considered is that the presence (or absence) of EHF content in a speech signal could affect a listener's ability to selectively attend to that signal. In real-world scenarios, speech signals with the highest levels of EHFs are produced by talkers who are directly facing the listener (Chu and Warnock, 2002; Kocon and Monson, 2018; Monson *et al.*, 2012). Through experience with natural signals, the auditory brain may learn that EHFs provide cues to determine the head orientation of a talker (Monson *et al.*, 2019). A talker facing the listener presumably indicates the intent to address the listener (Neuhoff, 2003). Thus, it may be that high levels of EHFs flag a speech signal as ecologically important, thereby facilitating selective attention to that signal and only that signal. However, if this were the case, one would expect worse performance in the LP8k Target, FB Masker condition, relative to the filter-matched LP8k condition. Although we did observe a worsening in this condition for the one-talker masker, this effect was small (0.6 dB) and the opposite trend was observed with the two-talker masker, providing little evidence to support this hypothesis.

Another observation from the present data is that EHF cues conferred little benefit in the FB masker condition (compare filter-matched FB conditions to filter-matched LP8k conditions). Thus, unlike non-facing masker scenes, EHF cues appear to be of little utility when FB masker talkers are simulated to face the listener. This finding is important because, apart from only a few studies of which we are aware (Braza *et al.*, 2022; Flaherty *et al.*, 2021; Monson *et al.*, 2019; Strelcyk *et al.*, 2014; Trine and Monson, 2020), nearly all speech-in-speech studies use masker talker recordings that simulate maskers facing the listener. Our data indicate that the utility of EHF cues is substantially reduced in this unnatural listening scenario, which may be why EHFs have been thought previously to provide little benefit for speech recognition (Levy *et al.*, 2015; Moore *et al.*, 2010).

While there are few data in the literature to compare to the present dataset, Polspoel *et al.* (2022) used a similar research design to evaluate the effect of EHF content on speech-in-noise recognition. Notable differences from the present study were the use of speech-shaped noise maskers, Dutch speech materials, and monaural presentation over headphones. That study measured sentence recognition at –5 dB SNR and reported significantly better sentence recognition with an FB target compared to an LP8k target, irrespective of the speech-shaped noise masker bandwidth. Scoring responses by words correct, mean performance was 27.5% when both the target and masker were LP8k, 40.3% when the target and masker were both FB, and 51.3% when the target was FB and the masker was LP8k. One obstacle to comparing their results with the present data is the fact that the current protocol estimated SRT rather than percent correct at a fixed SNR. However, the psychometric function fits to the present data can be used to estimate associated changes in percent correct. For this analysis, logit functions were fitted to the proportion of words correct by SNR, weighted by the number of observations, for each participant and each condition. For the LP8k Target, LP8k Masker condition, 27.5% correct was associated with a mean of –23.1 dB SNR for the one-talker masker and –10.1 dB SNR for the two-talker masker. Both of these values are well below the –5 dB SNR level used by Polspoel *et al.* (2022). Discrepancies in SRTs could be due to the use of different speech materials and talker recordings, both factors known to affect speech-in-speech recognition (Brungart *et al.*, 2001; Buss *et al.*, 2019; Calandruccio *et al.*, 2017), or due to other differences noted above.

Changes in percent correct with the introduction of EHF target content at −23.1 dB SNR (one-talker) and –10.1 dB SNR (two-talker) can be compared to those reported by Polspoel *et al.* (2022). For the one-talker masker, scores were 30.9% for the FB Target, FB Masker

and 34.6% for the FB Target, LP8k masker. For the two-talker masker, those values were 31.5% for the FB Target, FB Masker and 63.3% for the FB Target, LP8k masker. Thus, whereas Polspoel *et al.* (2022) observed a 12.8-point increase in performance when increasing the bandwidth of both the target and masker, this benefit was only 3.4 percentage points for the one-talker masker and 4.0 percentage points for the two-talker masker in the present dataset. This could indicate that extending the stimulus bandwidth is more beneficial for speech-in-noise than speech-in-speech recognition. Speech-shaped noise with the same long-term average power spectrum as the target will tend to exert more energetic masking for low- and mid-frequency speech cues than high-frequency cues. The reason for this is related to the larger variance in target energy at EHFs than lower frequencies; whereas low- and mid-frequency energy tends to be associated with long-duration phonemes like vowels, high-frequency energy tends to be briefer and therefore contributes less to the long-term average (Phatak and Allen, 2007). A larger role of EHF bandwidth for speech-in-noise than speech-in-speech is not consistent with all prior research, however. For example, Best *et al.* (2019) argued that bandwidth requirements for speech recognition are wider when the listener has access to spectro-temporally sparse cues, such as those that are audible in the presence of a two-talker masker, compared to intact speech cues.

For the LP8k Masker, introducing EHF target content improved speech-in-noise performance by 23.8 percentage points in the data of Polspoel *et al.* (2022); improvements were 3.6 percentage points for the one-talker masker and 35.8 percentage points for the two-talker masker in the current dataset. Less benefit of unmasked EHF target content for the one-talker masker than for the other two conditions could be due to inaudibility of target EHF cues, given the low overall levels for target speech at the threshold for the one-talker masker, as highlighted earlier. However, it could also be due to the quality and quantity of available speech cues at lower frequencies. One-talker maskers exert less informational masking than two-talker maskers (Buss *et al.*, 2017; Rosen *et al.*, 2013), presumably due to easier stream segregation and/or greater opportunities for glimpsing in the one-talker masker. Intermediate benefit of EHF content for the noise masker could be related to the near-absence of spectro-temporal modulation in the masker, and therefore no opportunities for glimpsing, coupled with a good ability to make use of audible speech cues. The marked informational masking obtained with a two-talker masker (Rosen *et al.*, 2013) could make any cues that improve glimpsing very valuable.

We did not replicate the previous finding of an association between EHF pure-tone thresholds and speech-in-speech recognition for young, normal-hearing listeners with good EHF hearing (Trine and Monson, 2020). We and others have reported a relationship between EHF audiometric thresholds and speech-in-noise for populations that exhibited a wider range of EHF thresholds (Braza *et al.*, 2022; Mishra *et al.*, 2021; Motlagh Zadeh *et al.*, 2019; Polspoel *et al.*, 2022; Yeend *et al.*, 2019), but the present data provide no evidence that this relationship is present in young, normal-hearing listeners with generally good EHF thresholds, suggesting effects of reduced audibility at EHFs for speech recognition may not be detectable until substantial EHF hearing loss is present. The present data do demonstrate that loss of audibility of EHFs in speech (*via* low-pass filtering) reduces speech recognition, and this finding has implications for individuals who may have reduced audibility of EHFs in speech due to EHF hearing loss.

One limitation of the present study is that the choice of stimuli may have impacted the results. Speech-in-speech recognition is known to vary markedly across speech materials and across recordings of the same corpora (Freyman *et al.*, 2007). The pattern and magnitude of results reported here could therefore depend on the stimuli used in the present experiment. For example, using a one-talker masker stimulus eliciting more informational masking could result in a higher SRT and a larger effect of EHF target content in the mismatch condition. Similarly, target sentences associated with less semantic context could affect reliance on EHF target audibility. Furthermore, the stimuli used here were all female talkers, which tend to exhibit higher EHF levels than male talkers (Monson *et al.*, 2012). Whether gender influences the utility of EHF cues is currently unknown but is a question worthy of further investigation.

In summary, EHFs in target speech make an appreciable contribution to speech-in-speech recognition when they are not masked by competing talkers. This condition is met in real-world auditory scenes in which background talkers have different head orientations but is *not* met in traditional speech-in-speech testing paradigms where masker talkers are simulated to face the listener. Notably, words with higher EHF levels are more likely to be recognized for speech-in-speech. These findings contribute to the growing evidence that EHF hearing is useful for speech recognition. The present data demonstrate that the EHF benefit is influenced somewhat by the number of masker talkers. It will be important to assess whether and how EHF cues are affected by other factors such as spatial separation of talkers, talker gender, or EHF hearing loss.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Deafness and Other Communication Disorders grant number R01 DC019745 (B.B.M.).

Ananthanarayana, R., Trine, A., and Monson, B. B. (**2022**). "Extended high-frequency pure-tone thresholds predict speech-in-speech recognition even when extended high-frequency speech cues are absent," J. Acoust. Soc. Am **151**(4), A224.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially hearing children," Br. J. Audiol. **13**(3), 108–112.

Best, V., Roverud, E., Baltzell, L., Rennies, J., and Lavandier, M. (**2019**). "The importance of a broad bandwidth for understanding 'glimpsed' speech," J. Acoust. Soc. Am **146**(5), 3215–3221.

Braza, M. D., Corbin, N. E., Buss, E., and Monson, B. B. (**2022**). "Effect of masker head orientation, listener age, and extended high-frequency sensitivity on speech recognition in spatially separated speech," Ear Hear **43**(1), 90–100.

Bronkhorst, A. W. (**2015**). "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," Atten. Percept. Psychophys. **77**(5), 1465–1487.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Buss, E., Hodge, S. E., Calandruccio, L., Leibold, L. J., and Grose, J. H. (**2019**). "Masked sentence recognition in children, young adults, and older adults: Age-dependent effects of semantic context and masker type," Ear Hear **40**, 1117–1126.

Buss, E., Leibold, L. J., Porter, H. L., and Grose, J. H. (**2017**). "Speech recognition in one- and two-talker maskers in school-age children and adults: Development of perceptual masking and glimpsing," J. Acoust. Soc. Am. **141**(4), 2650–2660.

Calandruccio, L., Buss, E., and Bowdrie, K. (**2017**). "Effectiveness of two-talker maskers that differ in talker congruity and perceptual similarity to the target speech," Trends Hear. **21**, 1–14.

Carhart, R., and Jerger, J. F. (**1959**). "Preferred method for clinical determination of pure-tone thresholds," J. Speech Hear. Disord. **24**(4), 330–345.

Chu, W. T., and Warnock, A. C. C. (**2002**). "Detailed directivity of sound fields around human talkers," Technical Report, Institute for Research in Construction (National Research Council of Canada, Ottawa, Canada), pp. 1–47.

Crouzet, O., and Ainsworth, W. A. (**2001**). "On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation," in *Proceedings of the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, September 2, Aalborg, Denmark.

Flaherty, M., Kelsey, L., and Monson, B. B. (**2021**). "Extended high-frequency hearing and head orientation cues benefit children during speech-in-speech recognition," Hear. Res. **406**, 108230.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

Freyman, R. L., Helfer, K. S., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**(2), 1040–1046.

Hunter, L. L., Monson, B. B., Moore, D. R., Dhar, S., Wright, B. A., Munro, K. J., Zadeh, L. M., Blankenship, C. M., Stiepan, S. M., and Siegel, J. H. (**2020**). "Extended high frequency hearing and speech perception implications in adults and children," Hear. Res. **397**, 107922.

IBM (**2022**). "IBM Watson—Speech to text," https://www.ibm.com/cloud/watson-speech-to-text (Last viewed June 25, 2022).

Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**(3), 1252–1263.

Kocon, P., and Monson, B. B. (**2018**). "Horizontal directivity patterns differ between vowels extracted from running speech," J. Acoust. Soc. Am. **144**(1), EL7–EL12.

Levy, S. C., Freed, D. J., Nilsson, M., Moore, B. C., and Puria, S. (**2015**). "Extended high-frequency bandwidth improves speech reception in the presence of spatially separated masking speech," Ear Hear. **36**(5), e214–e224.

Lippmann, R. P. (**1996**). "Accurate consonant perception without mid-frequency speech energy," IEEE Trans. Speech Audio Process. **4**(1), 66–69.

Lough, M., and Plack, C. J. (**2022**). "Extended high-frequency audiometry in research and clinical practice," J. Acoust. Soc. Am. **151**(3), 1944–1955.

Maniwa, K., Jongman, A., and Wade, T. (**2009**). "Acoustic characteristics of clearly spoken English fricatives," J. Acoust. Soc. Am. **125**(6), 3962–3973.

MathWorks Audio Toolbox Team (**2022**). "speech2text," https://www.mathworks.com/matlabcentral/fileexchange/65266-speech2text (Last viewed July 4, 2022).

Mishra, S. K., Fu, Q., Galvin, J. J., III., and Galindo, A. (**2023**). "Suprathreshold auditory processes in listeners with normal audiograms but extended high-frequency hearing loss," J. Acoust. Soc. Am. **153**(5), 2745–2745.

Mishra, S. K., Saxena, U., and Rodrigo, H. (**2021**). "Extended high-frequency hearing impairment despite a normal audiogram: Relation to early aging, speech-in-noise perception, cochlear function, and routine earphone use," Ear Hear. **43**(3), 822–835.

Monson, B. B., and Buss, E. (**2022**). "On the use of the TIMIT, QuickSIN, NU-6, and other widely used bandlimited speech materials for speech perception experiments," J. Acoust. Soc. Am. **152**(3), 1639–1645.

Monson, B. B., and Caravello, J. (**2019**). "The maximum audible low-pass cutoff frequency for speech," J. Acoust. Soc. Am **146**(6), EL496–EL501.

Monson, B. B., Hunter, E. J., and Story, B. H. (**2012**). "Horizontal directivity of low- and high-frequency energy in speech and singing," J. Acoust. Soc. Am. **132**(1), 433–441.

Monson, B. B., Lotto, A. J., and Story, B. H. (**2012**). "Analysis of high-frequency energy in long-term average spectra (LTAS) of singing, speech, and voiceless fricatives," J. Acoust. Soc. Am. **132**(3), 1754–1764.

Monson, B. B., Lotto, A. J., and Story, B. H. (**2014**). "Detection of high-frequency energy level changes in speech and singing," J. Acoust. Soc. Am. **135**(1), 400–406.

Monson, B. B., Rock, J., Schulz, A., Hoffman, E., and Buss, E. (**2019**). "Ecological cocktail party listening reveals the utility of extended high-frequency hearing," Hear. Res. **381**, 107773.

Moore, B. C., Füllgrabe, C., and Stone, M. A. (**2010**). "Effect of spatial separation, extended bandwidth, and compression speed on intelligibility in a competing-speech task," J. Acoust. Soc. Am. **128**(1), 360–371.

Motlagh Zadeh, L., Silbert, N. H., Sternasty, K., Swanepoel, W., Hunter, L. L., and Moore, D. R. (**2019**). "Extended high-frequency hearing enhances speech perception in noise," Proc. Natl. Acad. Sci. U.S.A. **116**(47), 23753–23759.

Neuhoff, J. G. (**2003**). "Twist and shout: Audible facing angles and dynamic rotation," Ecol. Psychol. **15**(4), 335–351.

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

Pinheiro, J., and Bates, D., and R Core Team (**2022**). "nlme: Linear and nonlinear mixed effects models," R package version 3.1-158, https://CRAN.R-project.org/package=nlme (Last viewed July 18, 2023).

Polspoel, S., Kramer, S. E., van Dijk, B., and Smits, C. (**2022**). "The importance of extended high-frequency speech information in the recognition of digits, words, and sentences in quiet and noise," Ear Hear **43**(3), 913–920.

R Core Team (**2022**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

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

Seeto, A., and Searchfield, G. D. (**2018**). "Investigation of extended bandwidth hearing aid amplification on speech intelligibility and sound quality in adults with mild-to-moderate hearing loss," J. Am. Acad. Audiol. **29**(3), 243–254.

Shadle, C. H., and Mair, S. J. (**1996**). "Quantifying spectral characteristics of fricatives," in *Proceedings of ICSLP 96*, October 3–6, Philadelphia, PA, pp. 1521–1524.

Shamma, S. A., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**(3), 114–123.

Sobon, K. A., Taleb, N. M., Buss, E., Grose, J. H., and Calandruccio, L. (**2019**). "Psychometric function slope for speech-in-noise and speech-in-speech: Effects of development and aging," J. Acoust. Soc. Am. **145**(4), EL284–EL290.

Strelcyk, O., Pentony, S., Kalluri, S., and Edwards, B. (**2014**). "Effects of interferer facing orientation on speech perception by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **135**, 1419–1432.

Trine, A., and Monson, B. B. (**2020**). "Extended high frequencies provide both spectral and temporal information to improve speech-in-speech recognition," Trends Hear. **24**, 233121652098029.

Vitela, A. D., Monson, B. B., and Lotto, A. J. (**2015**). "Phoneme categorization relying solely on high-frequency energy," J. Acoust. Soc. Am. **137**(1), EL65–EL70.

Yeend, I., Beach, E. F., and Sharma, M. (**2019**). "Working memory and extended high-frequency hearing in adults: Diagnostic predictors of speech-in-noise perception," Ear Hear. **40**(3), 458–467.