Speech recognition in one- and two-talker maskers in school-age children and adults: Development of perceptual masking and glimpsing

Emily Buss,[1,a)] Lori J. Leibold,[2] Heather L. Porter,[3] and John H. Grose[1]
[1]*Department of Otolaryngology/Head and Neck Surgery, University of North Carolina, Chapel Hill, North Carolina 27599, USA*
[2]*Center for Hearing Research, Boys Town National Research Hospital, Omaha, Nebraska 68131, USA*
[3]*Hearing and Speech Department, Children's Hospital Los Angeles, Los Angeles, California 90027, USA*

(Received 31 October 2016; revised 24 March 2017; accepted 27 March 2017; published online 13 April 2017)

Children perform more poorly than adults on a wide range of masked speech perception paradigms, but this effect is particularly pronounced when the masker itself is also composed of speech. The present study evaluated two factors that might contribute to this effect: the ability to perceptually isolate the target from masker speech, and the ability to recognize target speech based on sparse cues (glimpsing). Speech reception thresholds (SRTs) were estimated for closed-set, disyllabic word recognition in children (5–16 years) and adults in a one- or two-talker masker. Speech maskers were 60 dB sound pressure level (SPL), and they were either presented alone or in combination with a 50-dB-SPL speech-shaped noise masker. There was an age effect overall, but performance was adult-like at a younger age for the one-talker than the two-talker masker. Noise tended to elevate SRTs, particularly for older children and adults, and when summed with the one-talker masker. Removing time-frequency epochs associated with a poor target-to-masker ratio markedly improved SRTs, with larger effects for younger listeners; the age effect was not eliminated, however. Results were interpreted as indicating that development of speech-in-speech recognition is likely impacted by development of both perceptual masking and the ability recognize speech based on sparse cues.

© *2017 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4979936]

[VB] Pages: 2650–2660

I. INTRODUCTION

Children are worse than adults at recognizing masked speech, particularly when the masker is composed of other talkers (Corbin *et al.*, 2016; Hall *et al.*, 2002; Wightman and Kistler, 2005). This age effect cannot be attributed to maturation of the auditory periphery, which is functionally mature by 6 months of age (Werner, 2007). Prolonged development of speech recognition in a speech masker is more consistent with observations that maturation of the auditory cortex extends into late childhood (Moore and Linthicum, 2007). Speech-in-speech recognition is often described as entailing two stages of processing: (1) perceptual isolation of the target speech from the masker, through segregation and selective attention (Bregman, 1990; Leibold, 2012), and (2) combination of cues associated with the target across time and frequency (Cooke, 2006; Miller and Licklider, 1950). A failure to perceptually isolate target cues is sometimes described as perceptual or informational masking (Carhart *et al.*, 1969a), in contrast to energetic masking. In adults, speech-in-speech masking is thought to reflect perceptual masking rather than an inability to understand the target speech based on spectrotemporally sparse cues (Brungart *et al.*, 2006). Likewise, immature speech-in-speech recognition in children is typically attributed to immature segregation and/or selective attention (Leibold *et al.*, 2016; Sussman and Steinschneider, 2009). This interpretation in terms of perceptual masking is consistent with psychoacoustic data showing maturation of both auditory stream segregation (Sussman *et al.*, 2007) and perceptual masking for non-speech stimuli (Hall *et al.*, 2005). However, other data indicate that children are poorer than adults at recognizing speech based on spectrally and/or temporally sparse cues, even in cases where perceptual isolation of target cues is not thought to play a role in performance (Buss *et al.*, 2016; Elliott *et al.*, 1987; Mlot *et al.*, 2010). These data raise the possibility that the ability to understand speech based on sparse cues could impact speech-in-speech recognition differently at different points in development. The present study was designed to evaluate the contributions of two factors—perceptual masking and the ability to utilize sparse speech cues—to speech-in-speech recognition in children and adults with normal hearing.

A growing number of studies demonstrate a pronounced developmental effect for speech-in-speech masking, larger than that observed with noise maskers. Large child/adult differences have been observed for closed-set VCV recognition (Leibold and Buss, 2013), open- and closed-set word recognition (Buss *et al.*, 2016; Corbin *et al.*, 2016; Hall *et al.*, 2002), and sentence recognition (Wightman and Kistler, 2005). Development of speech-in-speech recognition appears to continue into adolescence (Corbin *et al.*, 2016; Wightman and Kistler, 2005).

While children tend to perform more poorly than adults, they benefit from many of the same cues that improve

[a)] Electronic mail: ebuss@med.unc.edu

performance in adult listeners, such as target/masker differences in location on the horizontal plane (Litovsky, 2005; Yuen and Yuan, 2014), sex (Wightman and Kistler, 2005), and language (Calandruccio *et al.*, 2016). The fact that children benefit from perceptual differences between target and masker speech implies that they have some ability to perceptually isolate the target from concurrent streams of masking speech, at least under some conditions. The smaller age effect for noise maskers than speech maskers has been interpreted as indicating that children experience little or no perceptual masking with noise maskers. This conclusion receives support from the finding that spatial segregation provides marked benefits for speech-in-speech, but only modest benefit for speech-in-noise recognition in children or adults (Corbin *et al.*, 2017; Freyman *et al.*, 2001). Interestingly, infants may experience perceptual masking in both noise and speech-based maskers (Leibold *et al.*, 2016), suggesting that the ability to perceptually isolate target speech from a noise masker may be learned.

Children's immature speech-in-speech recognition is often attributed to development of the ability to perceptually isolate target speech from the masker, and there are several findings in the literature to support this view. Sussman *et al.* (2007) demonstrated immature auditory stream segregation based on frequency differences in 5- to 8-year-olds and 9- to 11-year-olds relative to adults. The stimulus used in that study was a repeating sequence of tones (ABBABB…). When A and B tones are close in frequency this stimulus is perceived as a single stream, but as the frequency separation is increased the stimulus is increasingly likely to be heard as comprising two auditory streams. Using this stimulus, Sussman *et al.* (2007) evaluated performance in two tasks: a subjective task, in which listeners were asked to report whether they heard one or two streams, and a psychophysical task, in which listeners were asked to detect a level increment in the A stream in the face of level variability in the B stream, a task that requires segregation. In both tasks, children required a larger frequency separation between A and B tones to segregate streams compared to adults. These results were interpreted as showing that stream segregation is immature in children as old as 9 to 11 years of age. Other evidence of development in the ability to perceptually isolate target cues comes from error patterns obtained in speech-in-speech experiments. One hallmark of perceptual masking is when listeners repeat back words or phrases from the masker speech (e.g., Brungart *et al.*, 2006; Lee and Humes, 2012). This type of error is sometimes described as an intrusion from the masker stream. Several studies showing poor speech-in-speech recognition in children have also reported evidence of more intrusions in children's responses (Leibold and Buss, 2013; Wightman and Kistler, 2005), suggesting that their poor performance may be due to a failure of stream segregation and/or a failure to selectively attend to the target stream.

Even when perceptual masking is not thought to pose a challenge, children generally do not perform as well as adults when presented with spectrally and/or temporally sparse speech information. One example is speech recognition in the presence of modulated noise. Modulation of a noise masker allows adults to make use of speech cues available during time/frequency epochs associated with advantageous target-to-masker ratios (TMRs) and ignore epochs associated with poor TMRs (Brungart *et al.*, 2006; Howard-Jones and Rosen, 1993; Miller and Licklider, 1950), a process sometimes described as glimpsing (Cooke, 2006). Some data indicate that children benefit less from noise masker modulation than adults (Buss *et al.*, 2016; Hall *et al.*, 2012), although other studies report no child/adult difference (Stuart, 2008; Stuart *et al.*, 2006). Another demonstration of more stringent cue requirements in children than adults used the forward gating procedure, where performance is evaluated for stimuli that are gated off before the end of the target word; young children require a longer-duration segment of the target word in order to recognize it, compared to older children and adults (Elliott *et al.*, 1987; Metsala, 1997). Children also require a wider bandwidth to recognize bandpass-filtered speech (Eisenberg *et al.*, 2000; Mlot *et al.*, 2010) compared to adults. These results indicate development in the ability to recognize speech based on temporally and/or spectrally sparse speech cues, such that younger children require more cues or higher quality cues than older children and adults. Greater cue requirements could be due to reduced linguistic experience of younger listeners, reduced cognitive resources of younger listeners, or a combination of factors. One question posed in the present research is whether an age effect in the number or quality of cues required to understand speech contributes to development of speech-in-speech recognition.

A number of studies have shown that one-talker maskers are less effective than maskers composed of two or more talkers (Miller, 1947). Interestingly, the number of talkers needed to maximize masking differs across studies. At the low end, Freyman *et al.* (2004) assessed nonsense sentence recognition and found maximal masking for a two-talker masker. At the high end, Simpson and Cooke (2005) evaluated consonant recognition and found maximal masking for an eight-talker masker, with similar performance for 8–128 talkers. Regardless of the number of masker talkers associated with worst performance, the largest decrements in performance are typically observed when increasing from one to two masker talkers (reviewed by Iyer *et al.*, 2010). The decrement in performance going from a one-talker to a two-talker masker is sometimes described as the multi-masker penalty. One factor that may contribute to the multi-masker penalty is the availability of envelope modulation minima. The envelope of a one-talker masker provides the listener with a larger number of high-quality glimpses than a multi-talker masker (e.g., Festen and Plomp, 1990). Another factor that may contribute to the multi-masker penalty is increased difficulty associated with segregating three or more concurrent steams of speech (a target and two or more maskers). These possible factors are not mutually exclusive, in that expending greater cognitive resources to perceptually isolate the target from the masker could reduce a listener's ability to benefit fully from the available glimpses. Considering both of these factors, one prediction tested in the present study was that the child/adult difference is larger in a two- than a one-talker masker, due to reduced cognitive resources available for perceptually isolating target cues and subsequent utilization of those cues in younger children.

Two experiments were conducted to evaluate developmental effects for speech-in-speech recognition in a one-talker and a two-talker masker. The first experiment evaluated performance in a speech masker with and without a steady speech-shaped noise that was 10-dB down from the level of that speech masker. Because noise is not associated with perceptual masking in school-age children, detrimental effects of steady noise would be interpreted as reflecting reduced access to low-level speech cues which would otherwise be audible in the speech-masker modulation minima. The hypothesis is that young children are unable to make use of these low-level speech cues, due to difficulties perceptually isolating the target from the masker and/or due to greater cue requirements for target speech recognition. If young children are not relying on low-level speech cues, then their performance should be less detrimentally affected by noise that masks those cues as compared to adults. The second experiment assessed the relative contributions of perceptual masking and the ability to recognize speech based on sparse cues by using a signal processing procedure designed to mimic stream segregation and thereby facilitate selective attention to the target. This procedure separates the stimulus into time-frequency epochs, evaluates the TMR in each, and then replaces any epoch below some criterion TMR with silence (Brungart *et al.,* 2006; Wang, 2005). The resulting stimulus provides the listener with sparse cues that are relatively free of masking, greatly reducing or eliminating challenges associated with perceptually isolating target cues.

II. GENERAL METHODS

Listeners were children (5–16 years) and adults (18–35 years) with normal hearing. Exclusion criteria included a history of hearing problems, known cognitive delays, and an abnormal tympanogram on the day of test. None of the child listeners had delayed speech and language development, as evaluated by parental report. All listeners were native speakers of American English. All listeners had pure-tone detection thresholds of 20 dB hearing level or lower at octave frequencies 250–8000 Hz (ANSI, 2010).

Stimuli were a subset of those previously used by Calandruccio *et al.* (2014). Targets were 30 disyllabic words (e.g., tiger), spoken by a female talker, with durations of 425–680 ms (mean 550 ms). Each target was associated with a custom illustration. Speech maskers were generated based on recordings of two additional females reading different passages from Jack and the Beanstalk, each 2 min 48 s in duration. Speech-shaped noise (SSN) maskers matched the average long-term power spectrum of the two-talker speech masker. Speech recordings were originally made using a sampling rate of 44 100 Hz. This rate was reduced to 24 414 Hz for the present experiment.

The listener's task was a four-alternative forced choice. Four randomly-selected illustrations appeared on the screen at the beginning of the trial. Once the target finished playing, those pictures changed from black and white to color, prompting the listener to respond using the computer mouse or touchscreen. After a response was entered, correct-answer feedback was provided by removing non-target illustrations from the screen. The target level was adaptively varied to estimate the speech reception threshold (SRT) associated with 71% correct using a 2-down, 1-up stepping rule. The step size was 4 dB at the beginning of the track and 2 dB after the second track reversal. Each track continued for eight reversals, and the SRT was calculated as the mean target level at the last six reversals. Two estimates were obtained from each listener in each condition, with a third estimate obtained in cases where the first two differed by 3 dB or more. The final SRTs reported below are means of all SRTs obtained for each listener in each condition. Within an experiment, SRTs in different conditions were obtained in random interleaved order. Procedures were implemented using custom MATLAB scripts which controlled dedicated experimental hardware (RZ6, TDT). Stimuli were presented diotically over headphones (Sennheiser, HD-25 II), and all data were collected in a double-walled soundproof booth.

The SRT as a function of age was evaluated using linear and non-linear regression, with parameter estimates obtained by minimizing sum of squared error. A log(base 10) transformation was applied to age in years, to accommodate decelerating effects of development with increasing age. The age associated with mature performance was estimated as the intersection between the 95% confidence interval around adult performance and the best-fit regression model.

III. EXPERIMENT 1: ONE- AND TWO-TALKER MASKERS WITH AND WITHOUT SPEECH-SHAPED NOISE

The first experiment compared performance in a one-talker and a two-talker masker, with and without the addition of speech-shaped noise that was 10-dB below the level of the speech masker. Several published studies have evaluated performance in a combined speech-plus-noise masker as a means of estimating the relative contributions of energetic and perceptual masking (Agus *et al.,* 2009; Carhart *et al.,* 1968). In those studies, SRTs in a noise masker were compared to SRTs in a speech-plus-noise masker. Any worsening in performance with the addition of a speech masker, after accounting for increases in overall masker energy, was attributed to perceptual masking. In contrast, the present study compared speech-in-speech recognition with and without an additional noise masker. In this paradigm, worsening in performance with the addition of noise was attributed to the masking of low-level cues coincident with speech-masker modulation minima.

Noise was expected to elevate SRTs of adult listeners by masking target speech that would otherwise be audible in envelope minima of the speech masker. This effect was expected to be larger in the one- than the two-talker masker, due to the longer-duration envelope minima associated with the one-talker masker. The added noise masker was expected to have a less detrimental effect on younger children. This somewhat unusual prediction—*less* susceptibility to noise in young children—is based on the hypothesis that young children have a reduced ability to recognize the target based on sparse glimpses coincident with masker envelope minima in the speech-alone conditions. If they are not able to use these cues in the speech-alone conditions, then masking those cues with noise would have little or no effect on performance. A reduced effect of steady noise is also expected due to differences in SNR at threshold in the absence of SSN; if children have higher SRTs than adults, then a low-level SSN would mask a smaller proportion of the dynamic range of the target speech for children than adults. Nonetheless, age effects in susceptibility to steady noise are expected to reflect the extent to which low-level glimpses contribute to speech recognition. With respect to the multi-masker penalty, the child/adult difference was expected to be larger for the two-talker masker than the one-talker masker, due to the additional challenges posed by the presence of an additional talker.

A. Methods

Listeners were children (5.3–16.6 years, n = 33) and adults (18–33 years, n = 10) with normal hearing. Performance was evaluated in the one- and two-talker maskers, with noise (+SSN) or without noise (alone). The four masker conditions were: one-talker alone, one-talker + SSN, two-talker alone, and two-talker + SSN. The masker speech was played continuously at 60 dB SPL, and speech-shaped noise, when present, was played continuously at 50 dB SPL. This noise level was selected based on two considerations: (1) the noise masker should be high enough in level to effectively mask target speech coincident with a speech-masker envelope minimum and (2) it should not be so high as to reduce intelligibility of the speech masker. The decision to present the noise 10-dB down from the level of the speech masker was based on the observation that normal-hearing 5-year-olds perform at ceiling when tasked with recognizing noise-masked sentences presented at 5 to 10 dB SNR (Holder *et al.*, 2016). The stimuli and procedures otherwise followed those described in the general methods.

Supplemental data were also collected on a second set of adult listeners (18–31 years, n = 9). In this cohort, SRTs were evaluated in a speech-shaped noise that was 30, 40, 50, 60, or 70 dB SPL. This noise was either presented alone or in combination with a 60-dB-SPL one-talker masker. The rationale for collecting these additional data was to better understand the effect of speech-shaped noise on mature performance and to provide a point of comparison with previous data (e.g., Agus *et al.*, 2009; Carhart *et al.*, 1968). The one-talker speech masker was chosen for these conditions rather than the two-talker based on the expectation of more pronounced effects of SSN on speech recognition in the one-talker masker. These data also allow an estimate of the noise level for adults in the one-talker + SSN condition required to match children's performance in the one-talker alone condition.

B. Results

The SRTs for individual child listeners are plotted in Fig. 1 as a function of age. The left panel shows performance in the one-talker masker, and the right panel shows performance with the two-talker masker. The mean of adult SRTs is shown at the right of each panel, with error bars indicating the 95% confidence intervals. Symbol fill indicates whether or not the speech-shaped noise was present in addition to the speech masker. Lines show fits to the data, described in detail below. For both maskers, performance improved with listener age, but the trajectory of this improvement depended on the number of masker talkers.

Improvement in SRTs with child age for the one-talker masker was fitted using a power function of the form $y = b + m \times x^k$, where y is SRT in dB SPL and x is child age in log(base 10) of years. A pair of power functions provided a significantly better fit to the data than a pair of straight lines ($F_{2,60} = 13.96$, $p = 0.001$). Fitting power functions with a common value of k for both conditions did not significantly reduce the quality of the fit, so this simpler model was adopted.[1] No further model reduction was indicated. The quality of power-function fits was high for both the one-talker alone ($r^2 = 0.80$) and the one-talker + SSN ($r^2 = 0.79$) conditions. Over the age range tested here, children's SRTs

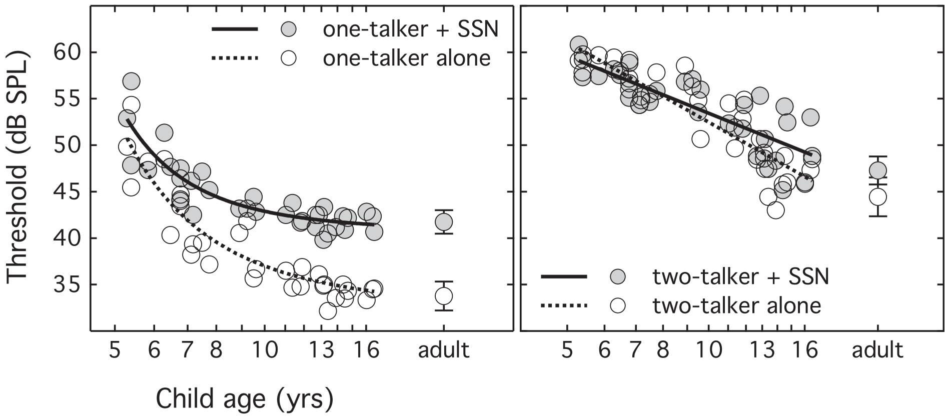

FIG. 1. SRTs plotted for individual child listeners as a function of age for the one-talker masker (left panel) and the two-talker masker (right panel). Mean SRTs for adult listeners appear at the right of each panel, with error bars indicating the 95% confidence interval. Symbol fill reflects masker condition, and lines indicate data fits, as defined in the legend. Power functions were fitted to data in the one-talker + SSN and one-talker alone conditions, and lines were fitted to data in the two-talker + SSN and two-talker alone conditions.

improved by 16.5 dB for the one-talker alone and by 11.1 dB in the one-talker +SSN condition. The age associated with mature performance was 12.9 years for the one-talker alone and 10.0 years for the one-talker + SSN.

In contrast to the one-talker data, straight lines provided good fits to SRTs as a function of child age for both the two-talker alone ($r^2 = 0.80$) and the two-talker + SSN ($r^2 = 0.67$) data sets.[2] The quality of data fits was significantly reduced when a common value of slope was fitted simultaneously to data in the two conditions ($F_{1,62} = 5.02$, $p = 0.029$), but the effect of fitting a common value of intercept just missed significance. Over the age range tested here, children's SRTs improved by 14.2 dB for the two-talker alone and by 10.2 dB in the two-talker + SSN condition. Estimates of mature performance were 16.1 years for the two-talker alone and 16.8 years for the two-talker +SSN.

1. Effect of speech-shaped noise for children and adults

One question of interest is how noise affected performance for children and adults in each of the speech maskers. To evaluate that question, two difference scores were computed for each listener, one for the one-talker masker and one for the two-talker masker; SRTs for the speech-masker alone were subtracted from those for the speech-masker + SSN. These scores are plotted in Fig. 2, following the conventions of Fig. 1. Lines show the difference between data fits to the SRTs obtained with and without noise, described above.

For adults, the noise/no-noise difference was larger for the one-talker masker than the two-talker masker, with mean effects of 8.0 and 2.9 dB, respectively. In both cases the noise/no-noise difference was significantly greater than zero ($t_9 = 12.07$, $p < 0.001$; $t_9 = 2.69$, $p = 0.025$), and values for the two speech maskers were significantly different from each other ($t_9 = 7.10$, $p < 0.001$). As in the adult data, noise had a larger effect on children's SRTs in the one-talker than the two-talker masker, with the caveat that the effect of noise increased with listener age. For the one-talker masker, the noise/no-noise difference rose from approximately 1.5 to 6.9 dB between 5.3 and 16.6 years of age. In contrast, the effect of noise was more modest for the two-talker masker, rising from −1.3 to 2.8 dB over the same age range. The noise/no-noise difference was significantly correlated with the log of child age for both the one-talker ($r = 0.54$, $p = 0.001$) and the two-talker masker ($r = 0.51$, $p = 0.002$).

2. Effect of number of talkers for children and adults

Another question of interest is how performance differed as a function of the number of masker talkers. Figure 3 shows pairwise differences between SRTs in the two-talker masker and the one-talker masker, following plotting conventions of Fig. 1. Results for the speech masker alone are shown in the left panel, and those for the speech masker with speech-shaped noise are shown in the right. Solid lines show the difference between fits to the data collected with two-talker and one-talker maskers, described above. Based on these fits, the additional masking associated with the second masker talker fell above the 95% confidence interval around adult means for children 5.8–15.8 years of age for the speech-masker alone, and 5.2–19.7 years of age for the speech-masker with speech-shaped noise. The largest difference between one- and two-talker SRTs occurred at 8.5 years for the speech-alone masker and 8.3 years for the speech + SSN masker. The difference between SRTs with one- and two-talker maskers was larger for the speech-masker alone than the speech-masker + SSN in both adult data ($t_9 = 7.10$, $p < 0.001$) and child data ($t_{32} = 8.27$, $p < 0.001$).

3. Supplemental data: Effect of speech-shaped noise level for adults

Mean SRTs in the supplemental conditions with adult listeners are plotted as a function of speech-shaped noise level in Fig. 4, with error bars indicating one standard deviation. SRTs in the one-talker + SSN condition were expected to asymptote at low noise levels, as thresholds approach those in the one-talker alone condition. Thresholds in the SSN masker were likewise expected to asymptote as levels approached absolute threshold. Data were therefore fitted with a power function, of the form $y = b + m \times x^k$, where y is SRT and x is the masker level, both in units of dB SPL. Those functions are shown with lines in Fig. 4.[3] Thresholds for the one-talker-alone condition, from the primary dataset, are included at the left of the figure for reference.

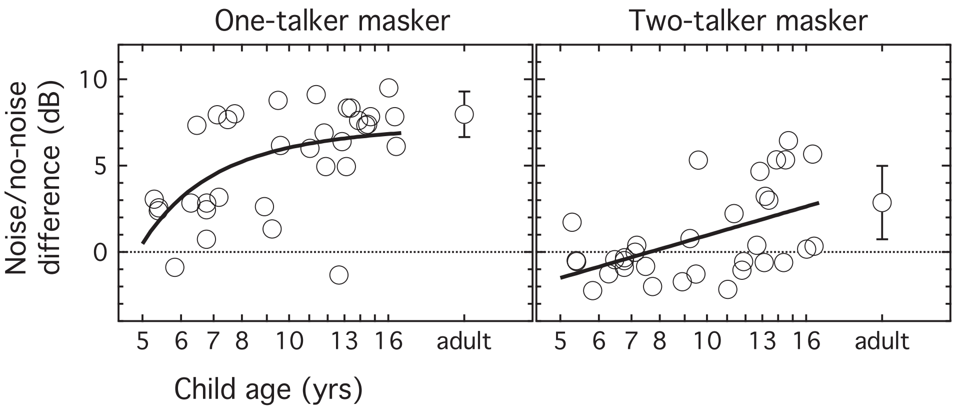

FIG. 2. The difference between SRTs with and without speech-shaped noise, plotted for individual child listeners as a function of age. Mean values for adults appear at the right of each panel, with error bars indicating the 95% confidence interval. Results for the one-talker masker are shown in the left panel, and those for the two-talker masker are shown in the right panel. Lines show differences between functions fitted to child listeners' SRTs.

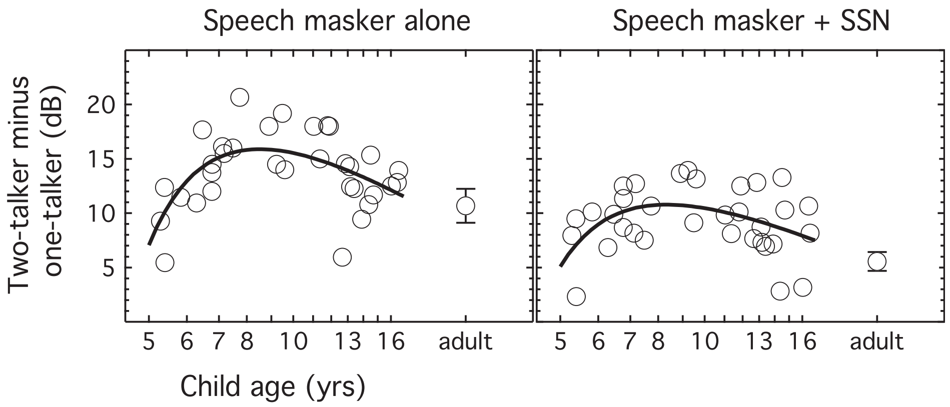

FIG. 3. The difference between SRTs in two-talker and one-talker maskers, plotted for individual child listeners as a function of age. Mean values for adults appear at the right of each panel, with error bars indicating the 95% confidence interval. Results for speech maskers alone are shown in the left panel, and those for speech maskers with speech-shaped noise are shown in the right panel. Lines show differences between functions fitted to child listeners' SRTs.

Not surprisingly, SRTs rose approximately linearly with increasing level of the speech-shaped noise alone, with approximately 10 dB of SRT elevation for every 10-dB increase in masker level. Growth of masking in the one-talker + SSN condition was relatively shallow at the low end of the range, with SRTs rising only 3–5 dB per 10-dB increment in noise level between 30 and 50 dB SPL. Growth of masking was approximately linear between 60 and 70 dB SPL. For speech-shaped noise at 30 dB SPL, SRTs for the one-talker + SSN condition (33.9 dB) were not significantly different from the adult mean SRT in the one-talker alone condition from experiment 1 (33.8 dB; $t_8 = 0.21$, $p = 0.842$). For speech-shaped noise at 60 dB SPL, SRTs were not significantly different for the noise-alone and one-talker + SSN after accounting for the 3-dB increase in overall level ($t_8 = 0.42$, $p = 0.687$). At 70-dB-SPL, SRTs *were* significantly higher for one-talker + SSN than the noise alone masker even after correcting for the 0.4-dB increase in overall level ($t_8 = 5.13$, $p = 0.001$).

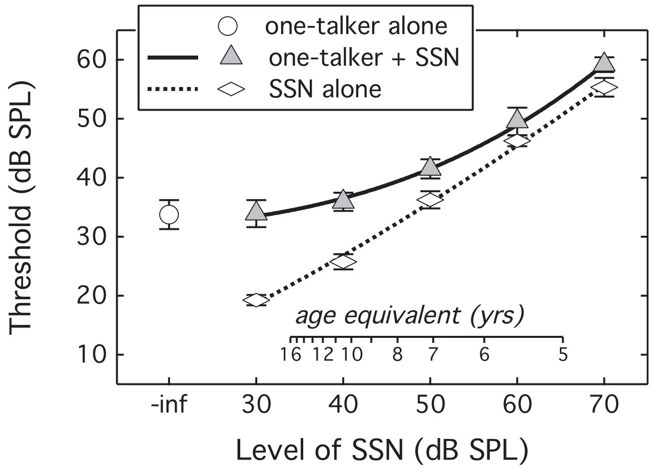

FIG. 4. Mean SRTs of adult listeners plotted as a function of speech-shaped noise level, with error bars indicating one standard deviation around the mean. Filled triangles show SRTs with the 60-dB-SPL one-talker plus speech-shaped noise, and diamonds show SRTs with the speech-shaped noise alone. Lines show fits to the data. Mean thresholds in the one-talker alone condition from the primary dataset are shown with the open circle. The noise level in adults associated with children's performance in quiet as a function of child age is indicated at the bottom of the panel.

Supplemental data from the one-talker + SSN condition were compared to child data from the one-talker alone condition in the primary dataset to estimate the magnitude of added speech-shaped noise necessary to achieve the mean SRTs obtained by children of different ages. These calculations were based on the function fits reported in footnotes 1 and 3. The results are indicated in Fig. 4, at the bottom of the panel. To match performance of a 5-year-old tested in the one-talker alone condition, an adult would require a speech-shaped noise level of 65.3 dB SPL. That level drops to 50.3 dB SPL to match performance of a 7-year-old, and 40.9 dB SPL to match performance of a 10-year-old.

C. Discussion

There is substantial improvement in SRTs between 5.3 and 16.6 years of age in all four masker conditions. In the one- and two-talker alone maskers, SRTs improved by approximately 16.5 and 14.2 dB, respectively (Fig. 1). However, the time-course of development differed for the two maskers. In the one-talker masker performance was estimated to be adult-like by 10–12.9 years of age, depending on the presence of speech-shaped noise. In contrast, performance in the two-talker masker was not projected to be adult-like until 16.1–16.8 years of age. Significant differences in regression fits to SRTs as a function of age provided additional support for these observations. One possible explanation for the finding of different developmental effects with the one- and two-talker maskers is that 10- to 12.9-year-olds are relatively adept at perceptually isolating the target from the one-talker masker, but were unsuccessful applying the same strategies with a two-talker masker, due to the greater cognitive demands associated with three simultaneous speech streams. Another possibility is that the cues available in the two-talker masker are poorer than those in the one-talker masker, and this reduction in cue quality has a larger detrimental effect on performance of younger listeners.

Another way of looking at these results is in terms of the multi-masker penalty. The difference between SRTs in the one- and two-talker maskers changed non-monotonically as a function of child age (Fig. 3). In the absence of speech-shaped noise, the multi-masker penalty was approximately

7 dB in 5-year-olds and 12 dB in 16-year-olds, with a peak value of approximately 16 dB at 8.5 years of age. This result is consistent with the idea that the ability to recognize speech masked by a single talker develops more rapidly in early childhood than the ability to recognize speech masked by a two-talker masker.

The present results can be compared with those of Litovsky (2005). That study measured speech-in-speech recognition in one- and two-talker maskers in 4.5- to 7.5-year-olds and adults with normal hearing. Targets were spondee words spoken by a male talker, and maskers were sentences spoken by a female talker. In one set of conditions the target and masker stimuli were presented from a loudspeaker directly in front of the listener, and the task was a four-alternative forced-choice. After accounting for the higher level of the two-talker masker, the multi-masker penalty was 3 dB for children and 3.6 dB for adults. In contrast, the multi-masker penalty in the present study ranged from 7 dB for 5-year-olds to 15 dB for 7-year-olds. One factor that may have reduced the multi-masker penalty in the data of Litovsky (2005) is the perceptual difference between the male target and the female masker talkers. Differences in target and masker talker sex have been shown to improve performance in both children and adults, often dramatically, presumably by facilitating perceptual isolation of the target (Wightman and Kistler, 2005). It is possible that the multi-masker penalty observed by Litovsky (2005) was lower than that observed here due to the greater perceptual masking associated with perceptually similar, matched-sex speech stimuli.

The presence of speech-shaped noise had a larger detrimental effect on performance in the one- than the two-talker masker for both children and adults (Fig. 2). This result is consistent with previous data from adults reported by Carhart *et al.* (1969b). That study measured spondee word recognition thresholds for a wide variety of masker conditions. Of most interest here were one-talker and two-talker maskers, comprised of sentences, presented with and without white noise at the same level as the speech masker. SRTs reported by Carhart *et al.* were 11-dB poorer in the two-talker than the one-talker masker, with the caveat that including the second talker increased overall masker power by 3 dB. Including white noise raised SRTs by 7.5 dB for one-talker masker and by 0.9 dB in the two-talker masker; SRT elevation was greater than expected based on increases in overall masker level for the one-talker masker, but not for the two-talker masker. In the main conditions of the present study the noise masker was presented 10-dB below the level of the speech-masker, so including speech-shaped noise had a negligible (0.4-dB) effect on overall masker level. Despite its lower level, noise significantly elevated adults' SRTs in the present study in both the one- and two-talker maskers, with mean effects of 8.0 and 2.9 dB, respectively. The pronounced detrimental effects of speech-shaped noise in the one-talker masker are consistent with the idea that the noise interfered with glimpsing of target speech during envelope minima in the speech masker. The relatively modest effect of speech-shaped noise in the two-talker masker is consistent with the idea that glimpsing plays a smaller role when more than one speech stream is present in the masker; this could be due to greater difficulty segregating the target and masker, poorer cue availability due to faster and/or shallower envelope fluctuation, or a combination of these two factors.

Noise also tended to elevate SRTs for children, but the magnitude of this effect depended on child age: younger children were less susceptible to the additional masking associated with the speech-shaped noise than older children. This general result is consistent with the idea that younger children are not as effective as older children and adults at perceptually isolating low-level target cues. However, it is also consistent with the idea that children require more cues to understand speech. If a listener relies primarily on cues that can be masked by a 50-dB-SPL noise, then a large effect of noise would be expected. However, reliance on more cues—including higher-level cues that are not effectively masked by a 50-dB-SPL noise—would result in a smaller effect of noise. Further consideration of the relative contribution of these factors motivated experiment 2.

Supplemental data on adults for a range of speech-shaped noise levels indicated that the detrimental effect of adding noise to the 60-dB-SPL one-talker masker reached asymptote at around 30 dB SPL—30-dB down from the speech masker level—at which point SRTs resembled those in the one-talker alone. This pattern of results is consistent with previous data on the dynamic range of speech (Studebaker and Sherbecoe, 2002). For a 60-dB-SPL noise level, SRTs were not significantly different in the noise-alone and one-talker with noise condition, after accounting for overall level effects. This finding might be interpreted as indicating no perceptual masking (Agus *et al.*, 2009; Carhart *et al.*, 1968). However, threshold elevation associated with the one-talker masker added to the 70-dB-SPL noise is consistent with perceptual masking. These results suggest that additivity of masking may not reliably indicate an absence of perceptual masking. Perhaps the most pertinent aspect of the supplemental data was the estimate of noise level required to equate adult performance in the one-talker + SSN condition to child performance in the one-talker alone condition. Adults required a speech-shaped noise that was up to 3-dB *higher* than the one-talker masker in order to obtain SRTs comparable to those of the youngest children tested in a one-talker alone condition.

IV. EXPERIMENT 2: EFFECTS OF DIGITAL TARGET/MASKER SEGREGATION

Experiment 1 showed that speech-shaped noise had a larger detrimental effect on speech-in-speech recognition for adults and older children than younger children. This result could be due to development of the ability to perceptually isolate target cues in the context of masker speech, but it could also reflect development in the ability to recognize speech based on sparse cues. The present experiment was designed to evaluate these two factors by comparing performance with and without the application of a digital technique designed to isolate the auditory stream associated with the target. The one-talker masker was chosen for experiment 2 because the noise/no-noise difference varied more as a function of age in the one- than the two-talker masker. This data pattern was interpreted as reflecting large age effects in the ability to make use of low-level target speech cues that are available in the one-talker alone masker condition.

The procedure for digitally segregating the target from the masker entails estimating the TMR as a function of time in narrow frequency regions of the target-plus-masker stimulus, and then eliminating energy in those epochs dominated by the masker. This technique is sometimes referred to as ideal time-frequency separation, reflecting the fact that it approximates the cues available to the listener after optimal segregation of the target and masker. This general approach has been used to enhance masked speech perception for normal-hearing and hearing-impaired listeners (Healy *et al.,* 2015; Wang, 2005), and to estimate the contributions of energetic masking to speech-in-speech recognition (Brungart, 2001; Kidd *et al.,* 2016). The implementation in the present study is described as digital time-frequency separation (DTFS), to avoid the implication that it represents optimal processing. The rationale for isolating time-frequency epochs dominated by the target in the present study is to evaluate two factors that could limit speech-in-speech recognition for children: the ability to perceptually isolate target cues and the ability to utilize those sparse cues to recognize the target speech. If perceptual masking were the critical factor responsible for child/adult differences, then performance should be comparable across age for DTFS-processed stimuli.

In previous studies of speech-in-speech recognition using ideal time-frequency separation, the stimulus consists of epochs of target plus masker for which the TMR is greater than some criterion (e.g., −6 dB). The present study included an additional condition in which the masker was omitted altogether, leaving just the target present in epochs coincident with a favorable TMR *had the masker been present*. The rationale for this manipulation was to evaluate susceptibility to masking after the most informational bits of the stimulus had been isolated. While there are data indicating that performance is not acutely sensitive to the criterion value (Brungart *et al.,* 2006), there are no data verifying that the optimal value is the same for children and adults. Evaluating performance for DTFS-processed stimuli with and without the masker included supports an evaluation of possible age effects in susceptibility to masking after digital segregation.

A. Methods

Listeners were children (5.0–15.2 years, n = 32) and adults (18–34 years, n = 13) with normal hearing. Only one of these listeners (5.8 years) had previously participated in experiment 1; data collection for the two experiments was separated by 3 months for this listener.

As in experiment 1, the target was a disyllabic word in a four-alternative forced-choice context. The masker was a one-talker masker presented alone (no noise). In contrast to experiment 1, the masker was presented for 1.4 s in each listening interval, gated on and off with 10-ms raised cosine ramps. The target was presented 10-ms after masker onset. Gated presentation was used in the present experiment to accommodate the additional signal processing associated with the DTFS technique. Both the target and masker stimuli were filtered into 30 bands between 100 and 8060 Hz. Each band was approximately one equivalent rectangular bandwidth (ERB) (Glasberg and Moore, 1990). Filtering was implemented with finite impulse response (FIR) filters; filter slopes spanned 36 Hz, and skirts of neighboring bands crossed at the 6-dB-down point.[4] In the unprocessed condition the bands were summed without further processing. In the two DTFS conditions, each band was temporally windowed using a series of 20-ms Hann functions, which overlapped at the 6-dB-down points. The output of each window was evaluated to determine whether the TMR met the local criterion of −6 dB. Windows meeting this criterion were retained, and the rest were scaled to an amplitude of zero. The next step was to reconstitute the target and masker arrays and sum across frequency bands. In the DTFS/T + M condition the processed target and masker arrays were added together. In the DTFS/T condition the stimulus presented to the listener included just the processed target array.

B. Results

SRTs are shown in Fig. 5. Individual data for child listeners are plotted as a function of age, and means of adult data are shown at the far right of the panel, with error bars indicating the 95% confidence interval. As in the one-talker alone condition of experiment 1, power functions with a common exponent (k) provided a significantly better fit to the data than straight lines ($F_{1,89} = 30.16$, $p < 0.001$). Fitting the two DTFS conditions with common parameter values did not significantly reduce the quality of the fit, so this simpler model was adopted.[5] Those fits are shown in Fig. 5. Based on these observations, individual listeners' mean SRTs in

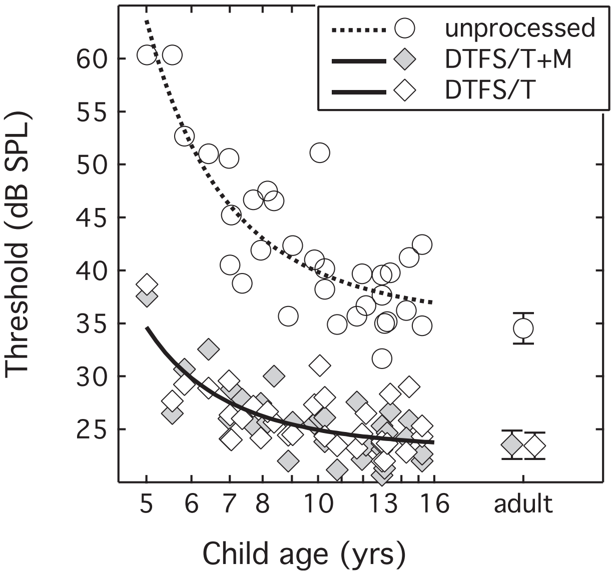

FIG. 5. SRTs for individual child listeners are plotted as a function of age. Mean SRTs for adults are shown at the far right of the panel, with error bars indicating the 95% confidence interval. Symbol shape and fill reflects the listening condition, and lines show power function fits to the child data, as indicated in the legend.

the two DTFS conditions were used in further analyses. For adults, performance was approximately 11-dB worse in the unprocessed condition than in either of the DTFS conditions; performance in the two DTFS conditions was nearly identical (0.10 dB; $t_{12} = 0.15$, $p = 0.885$). Based on line fits and 95% confidence intervals around adult data, child SRTs were estimated to be adult-like by 10.5 years of age for the two DTFS conditions. In contrast, performance in the unprocessed condition was not expected to reach adult levels until 19 years of age; this estimate should be interpreted with caution given that it is well beyond the age range of the study population.

Figure 6 shows the difference between SRTs in the unprocessed condition and the mean SRT in the two DTFS conditions. Values for individual children are plotted as a function of age, and adult means are shown at the right of the panel, with error bars indicating the 95% confidence interval. The solid line shows the difference between the power functions fitted to SRTs. Based on this fit to the data, the benefit of DTFS fell from 26.0 to 12.4 dB between 5.0 and 15.2 years of age. The nonlinear trend notwithstanding, there is a strong correlation between the benefit of DTFS and child age ($r = -0.65$, $p < 0.001$), with younger children experiencing greater benefit.

1. Comparison with data from experiment 1

The unprocessed one-talker masker condition of the present experiment was similar to the one-talker alone condition of experiment 1, with two exceptions: (1) the masker was gated in experiment 2 but it played continuously in experiment 1 and (2) stimuli were bandpass filtered (100–8060 Hz) in experiment 2 but not in experiment 1. Overall, the pattern of results was very similar between the two experiments. Adults' SRTs were slightly lower in experiment 1 than experiment 2, with mean values of 33.8 and 34.5 dB. However, this difference was not significant ($t_{21} = 0.72$, $p = 0.481$). A similar trend for poorer performance in experiment 2 was also observed in data from children, with a larger effect in younger listeners. Based on power-function fits to the data in each experiment, the mean differences in SRTs across experiments fell from 9.5 dB for 5-year-olds to 2.4 dB for 16-year-olds. This finding suggests that there may be an age effect associated with gating the masker or with bandpass filtering the target.

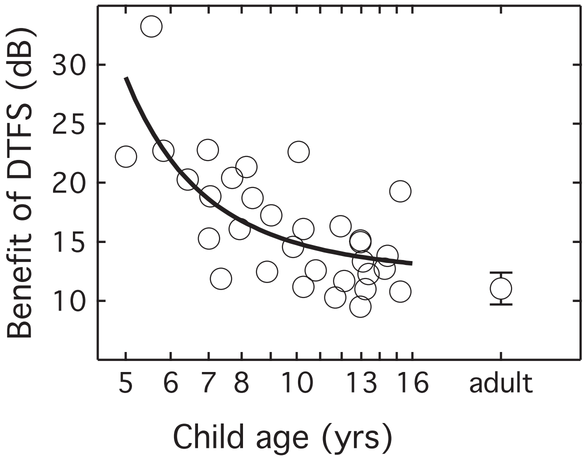

FIG. 6. The difference between SRTs in the unprocessed condition and mean SRTs in the two DTFS conditions (DTFS/T + M and DTFS/T), plotted for individual child listeners as a function of age. Mean values for adults appear at the right of the panel, with error bars indicating the 95% confidence interval. The line shows the difference between functions fitted to child listeners' SRTs.

C. Discussion

The most important result of experiment 2 is the observation that DTFS reduced the child/adult difference, but did not eliminate it. Comparable performance in the DTFS/T and DTFS/T+M conditions indicates that children and adults experienced little or no masking for these digitally segregated stimuli, such that performance was limited by the sparsity of cues available.

The difference between SRTs in the unprocessed and DTFS conditions was 11 dB for adults and 13 dB for the oldest children, but it was larger for younger children; the benefit associated with DTFS was 14-dB larger in 5- to 6-year-olds than adults. The increased benefit of DTFS in younger children supports the idea that development of segregation and/or selective attention plays an important role in immature speech-in-speech recognition, consistent with results obtained using other paradigms (Sussman and Steinschneider, 2009; Sussman *et al.,* 2007). Despite the benefit associated with DTFS, this processing did not eliminate the child/adult difference. Children under 10.5 years of age performed worse than adults in the DTFS stimulus conditions, with a child/adult difference of 8.3 dB observed for 5-year-olds. This finding is consistent with previous data indicating maturation in the ability to recognize speech based on spectrally and/or temporally sparse cues (Buss *et al.,* 2016; Hall *et al.,* 2012; Mlot *et al.,* 2010).

Similarity between the magnitudes of the DTFS benefit (14 dB for 5-year-olds) and the effect of age on performance in the DTFS maskers (8 dB for 5-year-olds) could be interpreted as indicating a greater contribution of perceptual masking to development. One consideration, however, is the interconnection between these factors under natural listening conditions. In the unprocessed stimulus conditions, the listener's ability to perceptually isolate target cues will affect the number and/or quality of cues available to listeners. For example, increased susceptibility to perceptual masking could result in the child listener attempting to identify the target based on an unfavorable mixture of target and masker cues, further increasing the target cue requirements. Another consideration has to do with interpretation of the DTFS results. The DTFS technique used here provides a rough approximation of cues that might be available to the listener after target/masker segregation, but this approximation is not precise. A conservative interpretation of the present results is that development of both perceptual masking and the ability to recognize speech based on sparse glimpses of the target speech likely contribute to speech-in-speech recognition for a one-talker masker.

Comparison of SRTs in the one-talker masker as a function of age in experiments 1 and 2 revealed a larger age effect in experiment 2. Theoretically, this could be due to differences in stimulus gating or to the use of filtering in experiment 2. An interpretation with respect to filtering seems unlikely, however. With A weighting, the target energy outside the passband of stimuli in experiment 2 was 32-dB down. This difference was perceptually discriminable for a subset of words (e.g., "pencil"), but only when original and filtered stimuli were played in close temporal proximity. There is some precedent in the literature for differential effects of gating for speech-in-speech recognition in children and adults. Hall *et al.* (2002) measured spondee recognition SRTs in 5- to 10-year-olds and adults, with a two-talker masker that either played continuously or gated on only during the listening interval. While gating had no effect on adults' SRTs, it did impact performance of child listeners: SRTs were 4.1-dB better in the gated two-talker masker. In contrast, the present data are consistent with better SRTs in the continuous one-talker masker. It is unclear whether the number of masker talkers plays a role in this discrepancy.

V. CONCLUSIONS

The present study evaluated speech-in-speech recognition for school-age children and adults, with special consideration of two factors that may contribute to performance: perceptual isolation of the target speech from the masker and the ability to recognize target speech based on spectro-temporally sparse cues. Results support the following conclusions.

(1) The ability to recognize speech in a one-talker masker develops earlier in childhood than the ability to recognize speech in a two-talker masker. This effect may be related to the cognitive resources needed to perceptually isolate the target in the context one vs two concurrent streams of masker speech.

(2) Adding speech-shaped noise to a speech masker tended to elevate SRTs for listeners of all ages, an effect that was larger for the one-talker than the two-talker masker. The effect of noise on SRT was correlated with listener age: young children were *less* affected by noise than older children and adults. This result is consistent with the idea that children are not as adept as adults at recognizing speech based on low-level glimpses.

(3) Digitally segregating the target and masker reduced SRTs for all listeners, but this effect was larger in younger children than older children and adults. This finding is consistent with a reduction in susceptibility to perceptual masking with increasing child age. While child-adult differences were reduced by digital segregation of the target and masker, a significant improvement in SRT with increasing age remained, indicating that the ability to recognize speech based on sparse cues could also play a role.

ACKNOWLEDGMENTS

This work was funded by a grant from NIH NIDCD (R01 DC011038, L.J.L.). Dr. Joseph Hall III and Dr. Jacob Oleson provided helpful feedback on this manuscript.

[1] Power functions fitted to child SRTs (y, dB SPL) in the one-talker as a function of age [x, log(base 10) transform of yrs] from experiment 1 were

$$y = 32.6 + 4.31(x^{-4.48})\ (\text{one-talker alone}),$$

$$y = 40.1 + 2.91(x^{-4.48})\ (\text{one-talker} + \text{SSN}).$$

[2] Lines fitted to child SRTs (y, dB SPL) in the two-talker as a function of child age [x, log(base 10) transform of yrs] from experiment 1 were

$$y = 81.2 - 28.7(x)\ (\text{two-talker alone}),$$

$$y = 73.9 - 20.5(x)\ (\text{two-talker} + \text{SSN}).$$

[3] Power functions fitted to SRTs (y, dB SPL) as a function of level (x, dB SPL) from supplemental conditions of experiment 1 were

$$y = 31.2 + 9.28\text{E} - 5(x^{2.97})\ (\text{one-talker} + \text{SSN}),$$

$$y = 1.76 + 0.16(x^{1.37})\ (\text{SSN alone}).$$

[4] Previous implementations of ideal time-frequency processing in the literature have typically used Gammatone filters. Procedures in the current study were adopted for speed.

[5] Power functions fitted to child SRTs (y, dB SPL) in the one-talker as a function of age [x, log(base 10) transform of yrs] from experiment 2 were

$$y = 35.0 + 4.83(x^{-4.97})\ (\text{unprocessed}\Big),$$

$$y = 23.0 + 1.97(x^{-4.97})\ (\text{DTFS/T} + \text{M and DTFS/T}\Big).$$

Agus, T. R., Akeroyd, M. A., Gatehouse, S., and Warden, D. (**2009**). "Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise," J. Acoust. Soc. Am. **126**, 1926–1940.

ANSI (**2010**). ANSI S3.6-2010, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Bregman, A. S. (**1990**). *Auditory Organization in Speech Perception, Auditory Scene Analysis* (MIT Press, Cambridge, MA), pp. 529–594.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

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

Buss, E., Leibold, L. J., and Hall, J. W. I. (**2016**). "Effect of response context and masker type on word recognition in school-age children and adults," J. Acoust. Soc. Am. **140**, 968–977.

Calandruccio, L., Gomez, B., Buss, E., and Leibold, L. J. (**2014**). "Development and preliminary evaluation of a pediatric Spanish-English speech perception task," Am. J. Audiol. **23**, 158–172.

Calandruccio, L., Leibold, L. J., and Buss, E. (**2016**). "Linguistic masking release in school-age children and adults," Am. J. Audiol. **25**, 34–40.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969a**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969b**). "Release from multiple maskers: Effects of interaural time disparities," J. Acoust. Soc. Am. **45**, 411–418.

Carhart, R., Tillman, T. W., and Johnson, K. R. (**1968**). "Effects of interaural time delays on masking by two competing signals," J. Acoust. Soc. Am. **43**, 1223–1230.

Cooke, M. A. (**2006**). "Glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Corbin, N., Bonino, A. Y., Buss, E., and Leibold, L. J. (**2016**). "Development of open-set word recognition in children: Speech-shaped noise and two-talker speech maskers," Ear Hear. **37**, 55–63.

Corbin, N. E., Buss, E., and Leibold, L. J. (**2017**). "Spatial release from masking in children: Effects of simulated unilateral hearing loss," Ear Hear. **38**, 223–235.

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**, 2704–2710.

Elliott, L. L., Hammer, M. A., and Evan, K. E. (**1987**). "Perception of gated, highly familiar spoken monosyllabic nouns by children, teenagers, and older adults," Percept. Psychophys. **42**, 150–157.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

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.

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

Hall, J. W., III, Buss, E., and Grose, J. H. (**2005**). "Informational masking release in children and adults," J. Acoust. Soc. Am. **118**, 1605–1613.

Hall, J. W., Buss, E., Grose, J. H., and Roush, P. A. (**2012**). "Effects of age and hearing impairment on the ability to benefit from temporal and spectral modulation," Ear Hear. **33**, 340–348.

Hall, J. W., Grose, J. H., Buss, E., and Dev, M. B. (**2002**). "Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children," Ear Hear. **23**, 159–165.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660–1669.

Holder, J. T., Sheffield, S. W., and Gifford, R. H. (**2016**). "Speech understanding in children with normal hearing: Sound field normative data for BabyBio, BKB-SIN, and QuickSIN," Otol. Neurotol. **37**, e50–e55.

Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.

Iyer, N., Brungart, D. S., and Simpson, B. D. (**2010**). "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," J. Acoust. Soc. Am. **128**, 2998–2910.

Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**, 132–144.

Lee, J. H., and Humes, L. E. (**2012**). "Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background," J. Acoust. Soc. Am. **132**, 1700–1717.

Leibold, L. J. (**2012**). *Development of Auditory Scene Analysis and Auditory Attention, Human Auditory Development* (Springer, New York), pp. 137–161.

Leibold, L. J., and Buss, E. (**2013**). "Children's identification of consonants in a speech-shaped noise or a two-talker masker," J. Speech Lang. Hear. Res. **56**, 1144–1155.

Leibold, L. J., Yarnell Bonino, A., and Buss, E. (**2016**). "Masked speech perception thresholds in infants, children, and adults," Ear Hear. **37**, 345–353.

Litovsky, R. Y. (**2005**). "Speech intelligibility and spatial release from masking in young children," J. Acoust. Soc. Am. **117**, 3091–3099.

Metsala, J. L. (**1997**). "An examination of word frequency and neighborhood density in the development of spoken-word recognition," Mem. Cogn. **25**, 47–56.

Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibiligy of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.

Mlot, S., Buss, E., and Hall, J. W. (**2010**). "Spectral integration and bandwidth effects on speech recognition in school-aged children and adults," Ear Hear. **31**, 56–62.

Moore, J. K., and Linthicum, F. H., Jr. (**2007**). "The human auditory system: A timeline of development," Int. J. Audiol. **46**, 460–478.

Simpson, S. A., and Cooke, M. (**2005**). "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**, 2775–2778.

Stuart, A. (**2008**). "Reception thresholds for sentences in quiet, continuous noise, and interrupted noise in school-age children," J. Am. Acad. Audiol. **19**, 135–146.

Stuart, A., Givens, G. D., Walker, L. J., and Elangovan, S. (**2006**). "Auditory temporal resolution in normal-hearing preschool children revealed by word recognition in continuous and interrupted noise," J. Acoust. Soc. Am. **119**, 1946–1949.

Studebaker, G. A., and Sherbecoe, R. L. (**2002**). "Intensity-importance functions for bandlimited monosyllabic words," J. Acoust. Soc. Am. **111**, 1422–1436.

Sussman, E., and Steinschneider, M. (**2009**). "Attention effects on auditory scene analysis in children," Neuropsychol. **47**, 771–785.

Sussman, E., Wong, R., Horvath, J., Winkler, I., and Wang, W. (**2007**). "The development of the perceptual organization of sound by frequency separation in 5-11-year-old children," Hear. Res. **225**, 117–127.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines* (Kluwer Academic Publishers, New York), pp. 181–197.

Werner, L. A. (**2007**). "Issues in human auditory development," J. Commun. Disord. **40**, 275–283.

Wightman, F. L., and Kistler, D. J. (**2005**). "Informational masking of speech in children: Effects of ipsilateral and contralateral distracters," J. Acoust. Soc. Am. **118**, 3164–3176.

Yuen, K. C. P., and Yuan, M. (**2014**). "Development of spatial release from masking in Mandarin-speaking children with normal hearing," J. Speech Lang. Hear. Res. **57**, 2005–2023.