Published in final edited form as:
J Acoust Soc Am. 2006 June ; 119(6): 3940–3949.

Informational masking of speech in children: Auditory-visual integration

Frederic Wightman[a)],
Heuser Hearing Institute and Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292

Doris Kistler, and
Heuser Hearing Institute and Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292

Douglas Brungart
Human Effectiveness Directorate, Air Force Research Laboratory, 2610 Seventh Street, Wright Patterson AFB, OH 45433

Abstract

The focus of this study was the release from informational masking that could be obtained in a speech task by viewing a video of the target talker. A closed-set speech recognition paradigm was used to measure informational masking in 23 children (ages 6–16 years) and 10 adults. An audio-only condition required attention to a monaural target speech message that was presented to the same ear with a time-synchronized distracter message. In an audiovisual condition, a synchronized video of the target talker was also presented to assess the release from informational masking that could be achieved by speechreading. Children required higher target/distracter ratios than adults to reach comparable performance levels in the audio-only condition, reflecting a greater extent of informational masking in these listeners. There was a monotonic age effect, such that even the children in the oldest age group (12–16.9 years) demonstrated performance somewhat poorer than adults. Older children and adults improved significantly in the audiovisual condition, producing a release from informational masking of 15 dB or more in some adult listeners. Audiovisual presentation produced no informational masking release for the youngest children. Across all ages, the benefit of a synchronized video was strongly associated with speechreading ability.

I. Introduction

The ability to segregate and selectively attend to the components of an auditory scene is central to our appreciation of and interaction with the world around us. Often the components of an auditory scene interfere with each other and degrade our auditory source segregation ability. One type of interference is energetic masking, thought to be a consequence of temporal and spectral overlap of the target and distracting sounds. The extent of energetic masking is accurately estimated by filter-bank models of the auditory periphery (Moore and Glasberg, 1987; Slaney, 1993, 1994). Informational masking is another type of interference that inhibits source segregation and is thought to occur when distracting sounds are highly variable or are perceptually similar to target sounds (Durlach *et al.*, 2003; Lutfi, 1990). The distinction between energetic and informational masking is readily appreciated in the case of a pure-tone detection task. When a pure tone is presented in a background of wideband Gaussian noise, the detection threshold for the tone is elevated

[a)]fred.wightman@louisville.edu.

almost exclusively by energetic masking, since the noise is relatively static and has a very different quality than the tonal signal. However, when the pure tone signal is presented in a background of other tones that have random frequencies and levels, informational masking occurs because the distracting sound is variable and has a tonal quality. Informational masking of this sort produces a substantial decrement in detection performance, represented by as much as a 40-dB threshold elevation in some conditions and in some listeners (Neff and Callaghan, 1988; Oh and Lutfi, 1998).

Recent results from our laboratory suggest that informational masking in pure-tone detection tasks is much greater in young children than in adults (Lutfi *et al.*, 2003; Oh *et al.*, 2001; Wightman *et al.*, 2003). Hall *et al.* (2005) report similar findings with complex tonal stimuli. However, the degree to which these results generalize to speech stimuli has only recently been demonstrated.

With a speech signal and speech distracter that overlap each other temporally and spectrally, it is sometimes difficult to determine how much of the observed masking effects are informational and how much are energetic. The clearest identification of the informational masking effects of speech distracters is found in the work of Brungart and colleagues (Brungart, 2001b; Brungart and Simpson, 2002a, b, 2004; Brungart *et al.*, 2005, 2001). This work shows that in certain tasks, in particular the Coordinate Response Measure (CRM) task developed at the Air Force Research Laboratory (Bolia *et al.*, 2000), nearly all of the interference in speech target recognition produced by speech distracters is informational and this interference can amount to an overall threshold shift of as much as 10 dB.

Children tested in the CRM paradigm appear to demonstrate much larger amounts of informational masking than adults (Wightman and Kistler, 2005). For the youngest children (ages 4–5 years) masking is more than 15 dB greater than in adults. A clear and monotonic age effect is also observed, with children as old as 16 years still not performing at adult levels. These results are generally consistent with other studies of children's speech recognition with speech distracters (Fallon *et al.*, 2000; Hall *et al.*, 2002). Interpretation of the age effect shown by Wightman and Kistler (2005) in terms of informational masking is complicated by the many reports of higher energetic masking thresholds in children (e.g., Allen and Wightman, 1994; Allen *et al.*, 1989; Schneider *et al.*, 1989). However, it appears to be the case that only the youngest children (i.e., preschoolers) produce markedly higher masked thresholds (Allen *et al.*, 1989; Schneider *et al.*, 1989). For example, recent data from our laboratory show that children older than about 6.5 years demonstrate adultlike detection thresholds for a tone masked by wideband noise (Oh *et al.*, 2001; Wightman *et al.*, 2003). Moreover, with a male target talker in the CRM paradigm, the switch from a male to a female distracter produces an 8 -dB release from masking in all age groups tested, including the adults (Wightman and Kistler, 2005). Differences in energetic masking between male and female distracters cannot fully explain the 8-dB masking release (Brungart, 2001b). Finally, an analysis of the errors made by the children suggested that the children confused the distracter and the target; errors came primarily from the distracter message. If energetic masking were involved, one might assume that target audibility would be decreased (see also Brungart, 2001b) and that this would lead to random distributions of errors. Thus, we argue that in the CRM speech task, the interference produced by the distracter is dominated by informational masking in all age groups tested.

There are many aspects of everyday listening that may mitigate informational masking effects. For example, spatial separation of the target signal from the distracters might be expected to produce a significant release from informational masking, and a number of experiments with adults confirm this expectation (Freyman *et al.*, 2001; Freyman *et al.*, 1999; Helfer and Freyman, 2005; Kidd *et al.*, 1998, 2005b). Unfortunately, initial research with children suggests a much smaller benefit of spatial separation (Hall *et al.*, 2005; Litovsky, 2005; Wightman *et al.*, 2003).

In many real world listening situations, the listener has the advantage of being able to see the target talker. It is well known that for adults integration of the information from the auditory signal and the visual cues obtained from speechreading (lipreading) can provide a substantial improvement in speech understanding in a background of noise (Sumby and Pollack, 1954; Summerfield, 1979) or multitalker babble (Sommers *et al.*, 2005). Recent experiments on speech recognition in speech backgrounds suggest that integration of auditory and visual sources of information (A/V integration) produces as much as a 10-dB release from informational masking (Helfer and Freyman, 2005). For most listeners this would mean the difference between understanding nothing and understanding everything. Because speech recognition is such an important aspect of development in children, and since children are often forced to listen to speech in noisy classroom environments where informational masking may be especially troublesome, we view A/V integration as an important topic to address in research on children.

It is important to note here that we use the term “A/V integration” to refer to the process whereby a listener achieves some improvement in speech understanding by viewing a simultaneous visual representation of the talker. However, as Grant and colleagues argue (Grant, 2002; Grant and Seitz, 1998, 2000; Grant *et al.*, 1998; van Wassenhove *et al.*, 2005), the benefit of combining auditory and visual cues is the result of at least two independent processes, information encoding and information integration. Thus, individual differences in A/V benefit, including age effects, might be the result of differences in either encoding or integration or both. The experiments reported here do not permit us to disentangle the separate contributions of the two processes. Nevertheless, consistent with many other reports in the literature, we will continue to use the term “A/V integration” to refer to the end result of both processes, and, where appropriate, we will discuss the encoding and integration processes separately.

There have been several previous studies of A/V integration in children. Research on infants clearly suggests that multisensory speech information is perceived. For example, Kuhl and Meltzoff (1982) demonstrated that infants looked longer at a face, the movements of which matched a speech sound, than at a face that did not match. Also, Rosenblum *et al.* (1997) reported results suggesting that infants are influenced by the McGurk effect (McGurk and MacDonald, 1976) whereby a given speech syllable is heard differently when presented synchronously with a video of a talker speaking a different speech syllable. In another study of a McGurk-like effect in infants, Desjardins and Werker (2004) showed that the effect is much less robust than in adults. Finally, a recent report by Hollich *et al.* (2005) suggested that infants might be able to segregate a target speech stream from a speech distracter if a synchronized video display of the target talker were present.

Results from studies of somewhat older children suggest that, although A/V synchrony may be perceived, A/V integration may not always be used to assist in auditory source segregation. Research with preschool and school-aged children (Desjardins *et al.*, 1997; Kishon-Rabin and Henkin, 2000; Massaro, 1984; Massaro *et al.*, 1986) reveals less A/V integration in children than in adults. The lack of A/V integration in children is not unexpected. For example, children are known to perform poorer than adults in tasks requiring face-processing (Aylward *et al.*, 2005; de Gelder *et al.*, 1998; Doherty-Sneddon *et al.*, 2001; Mondloch *et al.*, 2004; 2003; Schwarzer, 2000; Taylor *et al.*, 2004). Additionally, there are results which suggest that most children are relatively poor at speechreading (Massaro, 1984; Massaro *et al.*, 1986), a skill that is clearly necessary for A/V integration in speech tasks. Finally, the “auditory dominance effect” is much larger in children (Monsen and Engerbretson, 1983; Napolitano and Sloutsky, 2004; Robinson and Sloutsky, 2004; Sloutsky and Napolitano, 2003). Auditory dominance refers to the fact that when individuals are presented with simultaneous auditory and visual stimuli, attention is captured by the auditory stimulus.

Previous research on A/V integration, including that with infants and young children, has focused on listening in quiet, so the extent to which the results might generalize to more realistic noisy conditions is not clear. The purpose of the experiment described here is to study A/V integration (and the resultant release from informational masking) in children of various ages using a paradigm involving speech recognition in the presence of a speech distracter. The Coordinate Response Measure task, used in previous research from our laboratory (Wightman and Kistler, 2005), will be used here. This task has several advantages. First, a large number of studies with adults (Brungart, 2001a, b; Brungart and Simpson, 2002b, 2004, ^{2005}; Brungart *et al.*, 2001, ^{2005}; Kidd *et al.*, 2003, 2005a, b, c) indicates that performance in the CRM task is dominated by informational masking. Second, the amount of informational masking produced by the time-synchronized speech distracter in the CRM task is large, so that release from informational masking is easily measured (Arbogast *et al.*, 2002, 2005; Wightman and Kistler, 2005). Finally, as shown in our own study (Wightman and Kistler, 2005), reliable data revealing large amounts of informational masking can be obtained from children as young as 4 years performing in the task.

II. Methods

A. Listeners

Ten adults and 23 school-aged children served as participants in this experiment. Four of the adults and 16 of the children had also served (9 months earlier) as listeners in our previous study using the CRM task (Wightman and Kistler, 2005). Children and adults were recruited from the University of Wisconsin and University of Louisville communities. The adults ranged in age from 18 to 31.9 years. For convenience in data interpretation, the children were divided into three age groups: six in the 6–8.9-year group, seven in the 9–11.9-year group, and ten in the 12–16.9-year group. All adults and children passed a 20 dB HL screening for hearing loss at octave frequencies from 0.25 to 8 kHz. All children passed the annual vision screening performed in their schools and no adult reported an uncorrected visual deficit. The children were tested for middle-ear problems (routine tympanometry) before the first session and again if necessary. None of the adults or children recruited was excluded due to inability to perform the task. One 11-year-old child did not complete the experiment due to scheduling difficulties.

B. Stimuli

Speech stimuli were taken from the corpus of high-quality, digitally recorded CRM stimuli made available by Bolia *et al.* (2000). The corpus includes 2048 phrases of the form, “Ready, *call sign*, go to *color number* now.” Eight talkers (four male, four female) are recorded, each speaking 256 different phrases (eight different call signs, “baron,” “ringo,” “tango,” etc.; eight numbers, 1–8; and four colors, red, white, green, blue). The target phrase was always spoken by talker 0 (male) from the corpus, using the call sign “baron.” The distracter phrases always used a different male talker (1–3), call sign, color, and number. The distracter phrase used on each trial was chosen randomly with replacement from the available phrases.

As a check on the extent of energetic masking, some of the younger children were tested with a modulated noise distracter. The modulated noise was a speech-spectrum noise (long-term spectrum derived from the CRM phrases) with a temporal envelope determined by one of the distracter phrases, randomly selected on each trial as with the speech distracters. The envelope was derived by full-wave rectification and convolution with a 7.2-ms rectangular window, exactly as described by Brungart (2001b).

The audio stimulus materials were produced digitally (CRM stimuli taken from the distribution CD), converted to analog form (22 050-Hz sample rate) by a control PC, mixed, amplified, and presented to listeners via calibrated Beyer DT990-Pro headphones. The target and distracter phrases were time aligned on the distribution CD such that the word “ready” for target and distracter phrases started synchronously. Because of small differences in speaking rate and word length the durations of the phrases were slightly different.

The videos of the target talker (same talker as used for the target in the current experiment) were recorded in the corner of a large anechoic chamber at the Air Force Research Laboratory (Brungart and Simpson, 2005). The video was captured with a digital camera (Sony Digital Handycam) located roughly 1.5 m in front of the talker, who stood in front of a black, acoustically transparent background. The audio was recorded directly onto the videotape. The talker was instructed to repeat each of the 32 possible target messages in the CRM corpus (i.e., those with the target call sign “baron”) at a monotone level while keeping his head as still as possible. Breaks were inserted between the CRM phrases to avoid any effects of coarticulation between consecutive recordings. The resulting videotapes were downloaded onto a PC where they were partitioned into individual AVI files for each of the 32 recorded phrases. Then a commercially available video editor (VirtualDub, www.virtualdub.org) was used to crop the frames of the AVI files around the locations of the talker's head, convert them from color to grayscale, and compress them into the Indeo 5.1 codec.

To produce the test stimuli for the current study, the original video tracks were individually time aligned with the corresponding audio target message from the original high-quality CRM audio corpus using custom MATLAB software. The audio track on the original videotape was not used because of its rather poor quality. Initially, the high-quality audio was inserted in place of the videotape audio such that the starting points of the two audio tracks were the same. Since the same highly practiced talker was used for both recordings this initial alignment was generally satisfactory. This was expected because this specific talker had previously been described as “extremely consistent” (Brungart and Simpson, 2005). In some cases, to compensate for slight differences in the duration of pauses between the original and the high-quality audio sentences, video frames were added (duplicates) or deleted using the VirtualDub software until alignment seemed perfect. No modifications were made to the audio files. No more than 2 frames were added to or deleted from the video tracks at any single point. At the playback rate of 29 fps, this would amount to a time shift of less than 70 ms, which is about equal to the adult threshold for detection of A/V asynchrony (Grant *et al.*, 2004; Lewkowicz, 1996). Thus, for the adults, potentially detectable misalignments were probably rendered undetectable by the manipulation. Infant asynchrony detection thresholds are much larger (Lewkowicz, 1996), so there is no reason to expect that our child or adult listeners would perceive any misalignment of the audio and video. To check for proper alignment, several lab personnel viewed the videos as they were simultaneously presented with the audio. None reported a misalignment, although no formal assessment of detectability was made. The final time-aligned video file and the corresponding audio files (target and distracter) were combined into a single file (AVI format) on a trial-by-trial basis using VirtualDub software.

All conditions involved trials in which a single target and single distracter were presented to the listener's right ear. The overall level of the distracter in the target ear was held constant for all conditions at approximately 65 dB SPL. The level of the target was varied randomly from trial to trial in order to obtain complete psychometric functions, percent correct versus target/distracter ratio (T/D), from each listener. Depending on listener and condition the target/distracter ratio (T/D) ranged from −35 to +15 dB (in 5-dB steps). Thus the highest level of the target was 80 dB SPL.

C. Conditions

All listeners were tested in three conditions. In the video-only condition, no audio was presented and listeners responded to the video alone. This condition allowed assessment of a listener's untrained speechreading ability. Two other conditions were evaluated: the audio-only condition in which only the audio target and distracter messages were presented, and the audiovisual condition in which the audio was presented along with a simultaneous video of the target talker saying the target message. Data from the video-only condition were obtained in the first and last sessions. Data from the audio-only and audiovisual conditions were obtained in all sessions with the two conditions alternated.

Approximately 6–8 months after the completion of the experiment, seven of the younger listeners (four in the 6–8.9-year-old group and three in the 9–11.9-years-old group) were tested in a condition identical to the audio-only condition but using a modulated noise distracter. This condition, presumably involving mostly energetic masking, was included as a test of our assumption that masking in the audio-only condition and the audiovisual conditions was dominated by informational masking.

D. Procedure

Listeners sat in a sound-isolated room in front of a computer display. The display showed a start button and 32 response buttons arranged in four colored matrices of eight buttons each, numbered 1–8. For the conditions in which the video of the target talker was presented, a box approximately 6 in. wide and 8 in. tall appeared in the center of the screen in which the video could be seen. Individual trials were initiated by the listener by a mouse-click on the start button. After hearing the phrases, the listener moved the mouse cursor to the matrix of the heard color and clicked on the number corresponding to the heard number. No feedback was given regarding the correctness of the response. The response method was identical in the video-only condition. However, the listeners were told that they would see, but not hear, the “baron” talker. They were instructed to “watch the man's face” and respond according to what they thought the man said.

All participants completed a practice run of 30 trials listening to the target talker with no distracter to assure perfect performance at five levels ranging from 45 to 65 dB SPL. Next each listener completed a practice run of 60 trials in each of the audio-only and audiovisual conditions followed by a practice run of 30 trials in the video-only condition. The practice runs were used to determine the target-distracter (T/D) levels to be used for the test runs. In the audio-only condition, the level of the target was varied randomly (five or six levels) from trial to trial so that an entire psychometric function, from near perfect performance to chance, could be estimated during each session. The levels were 5 dB apart so that a 20-dB (five levels) or a 25-dB (six levels) range was covered. In the audiovisual condition all listeners completed runs in which the levels were 5 dB apart. Listeners who performed well above 50% correct at the lowest level were tested in one or more runs in which the levels were 10 dB apart (covering a 50-dB range) to determine if performance had reached plateau.

Both children and adults were tested in 60 or 120 trial blocks in the audio-only and audiovisual conditions and in 30 or 60 trial blocks in the video-only condition. Children completed 240–300 trials in each of the audio-only and audiovisual conditions and 30–90 trials in the video-only condition. Adults completed 120 trials in the video-only condition and 480 trials in the other two conditions.

Seven children were tested in the modulated noise condition. These children completed 210–240 trials in this condition, and, in the same session, an additional 180 trials in the audio-only condition.

Testing was conducted over the course of several sessions. Sessions were approximately 1–2 h long for both children and adults. Frequent breaks during each session were encouraged. Sessions were scheduled at the participant's convenience, usually once or twice per week. Most listeners completed the experiment in two to four sessions. Listeners were paid $8/h for their participation.

III. Results and Discussion

A. Data analysis

Although complete psychometric functions were obtained from the listeners in all conditions, the irregular form of many of the functions led us not to fit them with a smooth curve (e.g., logistic) and extract parameters of the fitted functions. The irregularity was characterized in most cases by a plateau in performance at T/D ratios of between 0 and −10 dB. This plateau has been observed in several previous studies using both the CRM and other paradigms (Brungart, 2001b; Brungart and Simpson, 2002b; Dirks and Bower, 1969; Egan *et al.*, 1954; Wightman and Kistler, 2005). Here we will present complete psychometric functions.

Individual differences are large in informational masking studies, especially those involving children (Lutfi *et al.*, 2003; Oh *et al.*, 2001; Wightman *et al.*, 2003; Wightman and Kistler, 2005), so averaging must be done with caution. In a previous study involving similar conditions (Wightman and Kistler, 2005), we argued that averaging produced psychometric functions that were reasonable representations of the individual psychometric functions of the members of the age group. The results described here show larger intersubject differences so some individual psychometric functions will be discussed in addition to the averages.

To quantify the release from informational masking provided by the simultaneous video display, an "A/V$_{\text{benefit}}$" score (Grant and Seitz, 1998) was computed for each individual. The A/V$_{\text{benefit}}$ score is defined as

$$A/V_{\text{benefit}}=\frac{(A/V-A)}{1-A}.$$

A/V and A are proportion correct recognition scores from the audiovisual and audio-only conditions, respectively, averaged over T/D ratios of −15 to 0 dB where informational masking is expected to be maximal. Thus the A/V$_{\text{benefit}}$ score reflects the difference between the recognition scores in the audiovisual and audio-only conditions relative to the amount of improvement possible given the audio-only score.

B. Audio-only condition

Figure 1 shows the data from individual listeners in each age group in the audio-only condition. In the data from each age group, the dashed line represents mean performance. As was the case in the previous study (Wightman and Kistler, 2005) individual differences are large. However, as before, we conclude that, in general, the mean psychometric functions appear to represent fairly the general shape of the psychometric functions in each group. The non-monotonicities (dips) in many of the individual functions (especially in the data from the 12– 16-years olds) are not well captured by the mean. However, those non-monotonicities are not statistically significant. Because of the limited amount of data at each T/D ratio (less than 50 trials for the children) the confidence limits around each percent correct value near the middle of the function are quite large, in some cases more than ±15%.

The audio-only condition in this experiment is nearly identical to the "monaural" condition of our previous experiment (Wightman and Kistler, 2005). The only differences are procedural: the stimulus levels in this experiment were presented randomly rather than in an up-down staircase, and no feedback was given. Not surprisingly, since many of the same listeners participated in both experiments, the mean data from this condition match well with the mean data from the comparable condition of the previous experiment (Wightman and Kistler, 2005). Figure 2 shows the two sets of mean psychometric functions (means weighted according to the number of trials included for each listener at each T/D) with the previous data age-grouped according to the current scheme. The only obvious difference is the apparent lack of the performance plateau in the data from the adults and older children at T/Ds from 0 to −10 dB. Although some individual listeners showed the plateau (Fig. 1), it was not as common an observation in this as compared to the earlier study. It seems reasonable to suggest that the use of the up-down staircase in the previous study may have facilitated the use of the level-difference segregation strategy whereby even though the target may be less intense than the distracter, it is still intelligible and can be recognized as the "softer talker." Random level presentation may not draw a listener's attention to this strategy, since the target levels do not systematically increase and decrease. It is also possible that the lack of feedback contributed to the apparent ineffectiveness of the level-difference segregation strategy. However, it should be noted that the overall level of performance was apparently not influenced by either the lack of feedback or the random level presentation. Other than the lack of a performance plateau in the current data, the two data sets are nearly identical.

In both data sets, and in our previous studies of informational masking with tonal stimuli (Oh *et al.*, 2001; Wightman *et al.*, 2003), individual variability was greater in the intermediate age groups than in either the adult group or the youngest group of children. This is inconsistent with data from other detection and discrimination studies from our laboratory in which variability is highest in preschoolers (Allen and Wightman, 1994, 1995; Allen *et al.*, 1989). However, it seems plausible that the non-monotonic change in variability could result from the developmental course of selective attention strategies. Attentional strategies for dealing with informational masking may not develop until the early school years, and then develop at different rates in different children. Supporting this view are the results of several studies (e.g., Geffen and Sexton, 1978; Geffen and Wale, 1979; Sexton and Geffen, 1979) which suggest that strategies for focusing attention do not begin to develop until about age 7 and continue to develop until the teenage years. Gibson (1969) argued that selective attention develops as a consequence of perceptual learning. If so it is reasonable to expect different rates of development in different children since each child's individual environment would play an important role in perceptual learning.

Figure 3 shows the data from the seven children who were tested with the modulated noise distracter and retested in the audio-only condition. Note that in each case, performance with the noise distracter was substantially better than with the speech distracter, amounting to a shift in the psychometric function of about 10 dB for all but one listener. In other words, the noise (presumed to be an energetic masker) was 10 dB less effective as a masker than the speech, supporting the view that masking in this experiment was dominated by informational masking. Figure 3 also shows that, for all but one listener, the retest performance in the audio-only condition was the same as in the original test. One listener (LBV) improved considerably at T/D ratios less than 0 dB; the performance plateau at T/D ratios between 0 and −10 dB improved by about 35%, reflecting much more adultlike performance (see Fig. 2). Also, note that the mean adult data from Brungart's modulated noise condition (Brungart, 2001b) are nearly identical to the modulated noise data from the seven children tested here (lower right panel of Fig. 3). This suggests that in the CRM task, children do not show substantially more energetic masking than adults.

C. Video-only condition

Figure 4 shows a scatterplot of the scores in the video-only condition as a function of age. The individual differences in listeners' abilities to speechread in this experiment are striking. Among the adults, some scored as high as 85% but most were a bit lower. The lowest score in the adult group was about 48%. This result is consistent with that reported recently by Brungart and Simpson (2005) who also used the CRM task. Although there is an obvious age effect, with the younger children showing relatively poorer speechreading abilities, even among the youngest children (group median score of 10.8%) there is one who scored 80%.

Large individual differences in speechreading ability are common. Watson *et al.* (1996) reported a range of scores from 50 subjects on a CID sentence test of speechreading from 4% to 87% total words correct (mean 37%; sd 17%). In a study of the speechreading ability of 60 undergraduates on a sentence test, Yakel *et al.* (2000) reported a mean score of 51% correct keywords reported with a standard deviation of 12 keywords. This suggests that more than 30% of the subjects had scores below about 39% or above 63%. These are but two examples of the results of the many experiments on speechreading in normal-hearing adults that show large individual differences [see Campbell *et al.* (1998) for a review of the classical work].

Individual differences in speechreading ability are also large in children. In one relatively recent study, Lyxell and Holmberg (2000) obtained a range of scores from 0% correct to 41% correct on a sentence test administered to 23 normal-hearing children. These results are consistent with those from the classic studies of children's speechreading reported by Massaro (1984; Massaro *et al.*, 1986) and demonstrate not only large individual differences but a level of speechreading proficiency that is generally lower in children than in adults. Both of these features of previous data can be seen in the data shown in Fig. 4.

D. Audiovisual condition

Figure 5 shows the individual performance of all listeners in the audiovisual condition. In the data from each age group, the dashed line represents mean performance. The data from the listener in the youngest group who performed much better than all the others were not included in the computation of the mean. This is the same listener who produced a video-only score of 80%. As was the case in the audio-only condition, large individual differences are evident in the data shown in Fig. 5, especially in the age 12–16.9-year group.

Mean psychometric functions from the audiovisual condition are plotted in Fig. 6. This figure also shows the mean functions from the audio-only condition and the mean speechreading scores. This figure clarifies the general developmental course of audiovisual integration and release from masking. The results suggest that on average there is no A/V release in children up to age 9, very little in children from 9 to 11.9, and much more in older children and adults. However, it seems clear that except for a few children in the age 12–16.9-year group, adultlike A/V integration and release from masking are not seen until the later teenage years. One should not expect much A/V integration when video encoding is lacking, as in those listeners who scored poorly in the video-only condition. Unfortunately, the current experiment does not allow us to identify whether a low A/$V_{benefit}$ score is a result of poor encoding, poor integration, or both.

Figure 6 suggests a strong relationship between A/V release from masking and speechreading ability, both of which increase with age. Recall that Fig. 4 shows the speechreading score for each participant as a function of age. Figure 7 shows the A/$V_{benefit}$ score for each participant as a function of age. The correlation between A/$V_{benefit}$ and age is 0.74. However, when speechreading score is factored out, the correlation of A/$V_{benefit}$ and age is only 0.03. The lack of correlation suggests that the age dependence we see in A/$V_{benefit}$ is almost entirely a result of the age-related changes in speechreading ability, which probably reflects age-related changes in the encoding of the visual information in speech. However, this does not imply that A/$V_{benefit}$ is determined only by speechreading ability. When the variance due to age is controlled, the correlation between A/$V_{benefit}$ and speechreading is 0.65, implying that other factors contribute to A/$V_{benefit}$. Thus, these data alone do not provide strong support for an age-related change in the ability of a listener to integrate auditory and visual information. This conclusion is identical to that reached by Sommers *et al.* (2005) in their study of younger and older adults.

Although interpreting data from individual participants is almost always problematic, there is one individual in this study who produced data that have very suggestive features. The data from this individual, a 7-year-old child, are shown in Fig. 8. The speechreading (video-only condition) score for this child was 80%, thus considerably better than any of the other children and on a par with scores from the adults. If speechreading were the sole determinant of A/V release from informational masking, we would expect a large A/V release from this individual. However, as the data show, the impact of the added video in the audiovisual condition was modest, producing about a 4–5-dB shift in the psychometric function. The A/$V_{benefit}$ score for this individual was only 0.35. Note that in the case of adults, for whom the mean speechreading score was almost 80%, the mean A/$V_{benefit}$ score was 0.65, and the lower end of the psychometric function in the audiovisual condition asymptotes at the speechreading score (Fig. 6). However, for the 7-year-old with an 80% speechreading score, the lower end of the psychometric function did not asymptote, suggesting a different strategy of combining the audio and video information. For this individual one might hypothesize that encoding of the video information was quite good, but integration of auditory and video information was less than optimal.

Given the importance of speechreading revealed by our data, we might speculate that speechreading training of young children may increase the release from informational masking they can achieve in everyday listening situations. The results from the study reported by Massaro *et al.* (1993) support this suggestion. In that study, speechreading training of adults on syllables, words, and sentences significantly improved A/V speech perception.

The results reported here do not permit a determination of the extent to which A/V integration reduced only informational masking. Many previous studies have reported an A/V benefit for adults recognizing speech targets in noise backgrounds (see, for example, Grant, 2002; Grant and Seitz, 1998; Grant *et al.*, 1998; Sommers *et al.*, 2005; Sumby and Pollack, 1954). Our results are generally consistent with the previous findings. However, since noise was used as the masker in the previous studies, the contribution of informational masking in these studies was probably minimal. In our experiment, informational masking was dominant. Given the conditions studied here, we cannot provide independent measures of A/V release from energetic and informational masking.

A recent study of informational masking reported by Helfer and Freyman (2005) shows an A/V release of almost 10 dB with a two-talker distracter. This is quite similar to the extent of A/V release we report here (Fig. 6) with adult listeners. There are no previous studies of which we are aware on A/V release from informational masking in children.

IV. Conclusions

A/V release from informational masking was measured in 23 children and 10 adults using a task requiring recognition of a target speech message in the presence of a single speech competitor. As observed in previous studies, the adults tested here achieved a large release from informational masking in the task as a result of watching a video of the target talker during the task. The effect was largest at the lowest T/D ratios; at T/D=−20 dB, the video produced a mean increase of about 40% in the recognition score.

The youngest children obtained very little benefit from the video. This was at least in part a result of the fact that their speechreading scores were low. Obviously, to achieve a benefit, visual information must first be encoded, and with such low speechreading scores, there is little evidence that the youngest children were encoding information from the video. As mentioned earlier, several previous studies have also shown that children produce low speechreading scores (Lyxell and Holmberg, 2000; Massaro, 1984; Massaro *et al.*, 1986). The results of one particularly intriguing study (Doherty-Sneddon *et al.*, 2001) suggested that in some tasks looking at a face would actually interfere with a child's ability to attend to an auditory message.

Children in intermediate age ranges obtained a smaller but potentially useful release from informational masking with the video. In the 12–16-years-old group, the improvement at a T/D of −20 dB was about 30%, although individual differences were large. The dependence of the extent of release from informational masking on age was a result of the strong association of age and speechreading ability.

Acknowledgments

The authors would like to thank Jen Junion Dienger, Rebekah Drska, the teachers in the University of Wisconsin Waisman Early Childhood Program, and the staff of the Heuser Hearing Institute for their contributions to the research. The research was supported financially by a grant to the first author from the National Institutes of Health (Grant No. R01-HD023333-14).

References

Allen P, Wightman F. Psychometric functions for children's detection of tones in noise. J Speech Hear Res 1994;37:205–215. [PubMed: 8170124]

Allen P, Wightman F. Effects of signal and masker uncertainty on children's detection. J Speech Hear Res 1995;38:503–511. [PubMed: 7596115]

Allen P, Wightman F, Kistler D, Dolan T. Frequency resolution in children. J Speech Hear Res 1989;32:317–322. [PubMed: 2739383]

Arbogast TL, Mason CR, Kidd G Jr. The effect of spatial separation on informational and energetic masking of speech. J Acoust Soc Am 2002;112:2086–2098. [PubMed: 12430820]

Arbogast TL, Mason CR, Kidd G Jr. The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. J Acoust Soc Am 2005;117:2169–2180. [PubMed: 15898658]

Aylward EH, Park JE, Field KM, Parsons AC, Richards TL, Cramer SC, Meltzoff AN. Brain activation during face perception: Evidence of a developmental change. J Cogn Neurosci 2005;17:308–319. [PubMed: 15811242]

Bolia RS, Nelson WT, Ericson MA, Simpson BD. A speech corpus for multitalker communications research. J Acoust Soc Am 2000;107:1065–1066. [PubMed: 10687719]

Brungart DS. Evaluation of speech intelligibility with the coordinate response measure. J Acoust Soc Am 2001a;109:2276–2279. [PubMed: 11386582]

Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am 2001b;109:1101–1109. [PubMed: 11303924]

Brungart DS, Simpson BD. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J Acoust Soc Am 2002a;112:664–676. [PubMed: 12186046]

Brungart DS, Simpson BD. Within-ear and across-ear interference in a cocktail-party listening task. J Acoust Soc Am 2002b;112:2985–2995. [PubMed: 12509020]

Brungart DS, Simpson BD. Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. J Acoust Soc Am 2004;115:301–310. [PubMed: 14759023]

Brungart DS, Simpson BD. Interference from audio distracters during speechreading. J Acoust Soc Am 2005;118:3889–3902. [PubMed: 16419831]

Brungart DS, Simpson BD, Ericson MA, Scott KR. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am 2001;110:2527–2538. [PubMed: 11757942]

Brungart DS, Simpson BD, Darwin CJ, Arbogast TL, Kidd G. Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task. J Acoust Soc Am 2005;117:292–304. [PubMed: 15704422]

Campbell, R.; Dodd, B.; Burnham, DK., editors. Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. Psychology Press; Hove, East Sussex, UK: 1998.

de Gelder, B.; Vroomen, J.; Laeng, B. Impaired speechreading related to arrested development of face processing. In: Burnham, D.; Robert-Ribes, J.; Vatikiotis-Bateson, E., editors. Auditory-Visual Speech Processing (AVSP'98). Terrigal-Sydney, Australia: 1998. p. 157-161.http://www.isca-speech.org/archive/avsp98

Desjardins RN, Werker JF. Is the integration of heard and seen speech mandatory for infants? Dev Psychobiol 2004;45:187–203. [PubMed: 15549681]

Desjardins RN, Rogers J, Werker JF. An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. J Exp Child Psychol 1997;66:85–110. [PubMed: 9226935]

Dirks DD, Bower DR. Masking effects of speech competing messages. J Speech Hear Res 1969;12:229–245. [PubMed: 5808851]

Doherty-Sneddon G, Bonner L, Bruce V. Cognitive demands of face monitoring: Evidence for visuospatial overload. Mem Cognit 2001;29:909–919.

Durlach NI, Mason CR, Kidd G Jr, Arbogast TL, Colburn HS, Shinn-Cunningham BG. Note on informational masking. J Acoust Soc Am 2003;113:2984–2987. [PubMed: 12822768]

Egan JP, Carterette EC, Thwing EJ. Some factors affecting multi-channel listening. J Acoust Soc Am 1954;26:774–782.

Fallon M, Trehub SE, Schneider BA. Children's perception of speech in multitalker babble. J Acoust Soc Am 2000;108:3023–3029. [PubMed: 11144594]

Freyman RL, Balakrishnan U, Helfer KS. Spatial release from informational masking in speech recognition. J Acoust Soc Am 2001;109:2112–2122. [PubMed: 11386563]

Freyman RL, Helfer KS, McCall DD, Clifton RK. The role of perceived spatial separation in the unmasking of speech. J Acoust Soc Am 1999;106:3578–3588. [PubMed: 10615698]

Geffen G, Sexton MA. The development of auditory strategies of attention. Dev Psychol 1978;14:11–17.

Geffen G, Wale J. Development of selective listening and hemispheric asymmetry. Dev Psychol 1979;15:138–146.

Gibson, EJ. Principles of Perceptual Learning and Development. Appleton-Century-Crofts; New York: 1969.

Grant KW. Measures of auditory-visual integration for speech understanding: A theoretical perspective. J Acoust Soc Am 2002;112:30–33. [PubMed: 12141356]

Grant KW, Seitz PF. Measures of auditory-visual integration in nonsense syllables and sentences. J Acoust Soc Am 1998;104:2438–2450. [PubMed: 10491705]

Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 2000;108:1197–1208. [PubMed: 11008820]

Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am 1998;103:2677–2690. [PubMed: 9604361]

Grant KW, Wassenhove Vv, Poeppel D. Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. Speech Commun 2004;44:43–53.

Hall JW III, Buss E, Grose JH. Informational masking release in children and adults. J Acoust Soc Am 2005;118:1605–1613. [PubMed: 16247871]

Hall JW III, Grose JH, Buss E, Dev MB. Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children. Ear Hear 2002;23:159–165. [PubMed: 11951851]

Helfer KS, Freyman RL. The role of visual speech cues in reducing energetic and informational masking. J Acoust Soc Am 2005;117:842–849. [PubMed: 15759704]

Hollich G, Newman RS, Jusczyk PW. Infants' use of synchronized visual information to separate streams of speech. Child Dev 2005;76:598–613. [PubMed: 15892781]

Kidd G Jr, Mason CR, Gallun FJ. Combining energetic and informational masking for speech identification. J Acoust Soc Am 2005a;118:982–992. [PubMed: 16158654]

Kidd G Jr, Mason CR, Brughera A, Hartmann WM. The role of reverberation in release from masking due to spatial separation of sources for speech identification. Acust Acta Acust 2005b;91:526–536.

Kidd G Jr, Arbogast TL, Mason CR, Gallun FJ. The advantage of knowing where to listen. J Acoust Soc Am 2005c;118:3804–3815. [PubMed: 16419825]

Kidd G Jr, Mason CR, Arbogast TL, Brungart DS, Simpson BD. Informational masking caused by contralateral stimulation. J Acoust Soc Am 2003;113:1594–1603. [PubMed: 12656394]

Kidd G Jr, Mason CR, Rohtla TL, Deliwala PS. Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. J Acoust Soc Am 1998;104:422–431. [PubMed: 9670534]

Kishon-Rabin L, Henkin Y. Age-related changes in the visual perception of phonologically significant contrasts. Br J Audiol 2000;34:363–374. [PubMed: 11201323]

Kuhl PK, Meltzoff AN. The bimodal perception of speech in infancy. Science 1982;218:1138–1141. [PubMed: 7146899]

Lewkowicz DJ. Perception of auditory-visual temporal synchrony in human infants. J Exp Psychol Hum Percept Perform 1996;22:1094–1106. [PubMed: 8865617]

Litovsky RY. Speech intelligibility and spatial release from masking in young children. J Acoust Soc Am 2005;117:3091–3099. [PubMed: 15957777]

Lutfi RA. How much masking is informational masking? J Acoust Soc Am 1990;88:2607–2610. [PubMed: 2283433]

Lutfi RA, Kistler DJ, Oh EL, Wightman FL, Callahan MR. One factor underlies individual differences in auditory informational masking within and across age groups. Percept Psychophys 2003;65:396–406. [PubMed: 12785070]

Lyxell B, Holmberg I. Visual speech reading and cognitive performance in hearing-impaired and normal hearing children (11–14 years). Br J Educ Psychol 2000;70:505–518. [PubMed: 11191184]

Massaro DW. Children's perception of visual and auditory speech. Child Dev 1984;55:1777–1788. [PubMed: 6510054]

Massaro DW, Cohen MM, Gesi AT. Long-term training, transfer, and retention in learning to lipread. Percept Psychophys 1993;53:549–562. [PubMed: 8332424]

Massaro DW, Thompson LA, Barron B, Laren E. Developmental changes in visual and auditory contributions to speech perception. J Exp Child Psychol 1986;41:93–113. [PubMed: 3950540]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature (London) 1976;264:746–748. [PubMed: 1012311]

Mondloch CJ, Dobson KS, Parsons J, Maurer D. Why 8-years-olds cannot tell the difference between Steve Martin and Paul Newman: Factors contributing to the slow development of sensitivity to the spacing of facial features. J Exp Child Psychol 2004;89:159–181. [PubMed: 15388304]

Mondloch CJ, Geldart S, Maurer D, Le Grand R. Developmental changes in face processing skills. J Exp Child Psychol 2003;86:67–84. [PubMed: 12943617]

Monsen RB, Engerbretson AM. The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. J Speech Hear Res 1983;26:91–97.

Moore BCJ, Glasberg BR. Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. Hear Res 1987;28:209–225. [PubMed: 3654390]

Napolitano AC, Sloutsky VM. Is a picture worth a thousand words? The flexible nature of modality dominance in young children. Child Dev 2004;75:1850–1870. [PubMed: 15566384]

Neff DL, Callaghan BP. Effective properties of multicomponent simultaneous maskers under conditions of uncertainty. J Acoust Soc Am 1988;83:1833–1838. [PubMed: 3403798]

Oh EL, Lutfi RA. Nonmonotonicity of informational masking. J Acoust Soc Am 1998;104:3489–3499. [PubMed: 9857508]

Oh EL, Wightman F, Lutfi RA. Children's detection of pure-tone signals with random multitone maskers. J Acoust Soc Am 2001;109:2888–2895. [PubMed: 11425131]

Robinson CW, Sloutsky VM. Auditory dominance and its change in the course of development. Child Dev 2004;75:1387–1401. [PubMed: 15369521]

Rosenblum LD, Schmuckler MA, Johnson JA. The McGurk effect in infants. Percept Psychophys 1997;59:347–357. [PubMed: 9136265]

Schneider BA, Trehub SE, Morrongiello BA, Thorpe LA. Developmental changes in masked thresholds. J Acoust Soc Am 1989;86:1733–1742. [PubMed: 2808922]

Schwarzer G. Development of face processing: The effect of face inversion. Child Dev 2000;71:391–401. [PubMed: 10834472]

Sexton MA, Geffen G. Development of three strategies of attention in dichotic monitoring. Dev Psychol 1979;15:299–310.

Slaney, M. An efficient implementation of the Patterson-Holdsworth auditory filter bank (35). Apple Computer, Inc.; Cupertino, CA: 1993.

Slaney, M. Auditory Toolbox. Apple Computer, Inc.; Cupertino, CA: 1994.

Sloutsky VM, Napolitano AC. Is a picture worth a thousand words? Preference for auditory modality in young children. Child Dev 2003;74:822–833. [PubMed: 12795392]

Sommers MS, Tye-Murray N, Spehar B. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. Ear Hear 2005;26:263–275. [PubMed: 15937408]

Sumby W, Pollack I. Visual contribution to speech intelligibility in noise. J Acoust Soc Am 1954;26:212–215.

Summerfield Q. Use of visual information for phonetic perception. Phonetica 1979;36:314–331. [PubMed: 523520]

Taylor MJ, Batty M, Itier RJ. The faces of development: A review of early face processing over childhood. J Cogn Neurosci 2004;16:1426–1442. [PubMed: 15509388]

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 2005;102:1181–1186. [PubMed: 15647358]

Watson CS, Qiu WW, Chamberlain MM, Li X. Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. J Acoust Soc Am 1996;100:1153–1162. [PubMed: 8759968]

Wightman FL, Kistler DJ. Informational masking of speech in children: Effects of ipsilateral and contralateral distracters. J Acoust Soc Am 2005;118:3164–3176. [PubMed: 16334898]

Wightman FL, Callahan MR, Lutfi RA, Kistler DJ, Oh E. Children's detection of pure-tone signals: Informational masking with contralateral maskers. J Acoust Soc Am 2003;113:3297–3305. [PubMed: 12822802]

Yakel DA, Rosenblum LD, Fortier MA. Effects of talker variability on speechreading. Percept Psychophys 2000;62:1405–1412. [PubMed: 11143452]

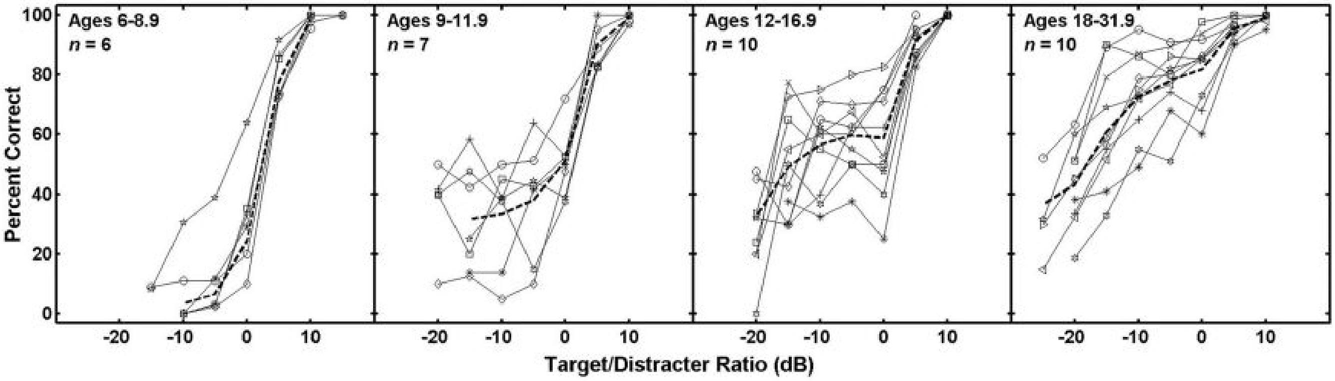

FIG. 1.

Psychometric functions for individual listeners in the audio-only condition of the CRM speech recognition task. The four panels show data from listeners in the four different age groups. The different symbols in each panel represent different listeners. The mean psychometric function in each group is shown by the dashed line. The mean excludes data from the one “outlier” in the 6–9.9 years age group whose data are represented by open stars.

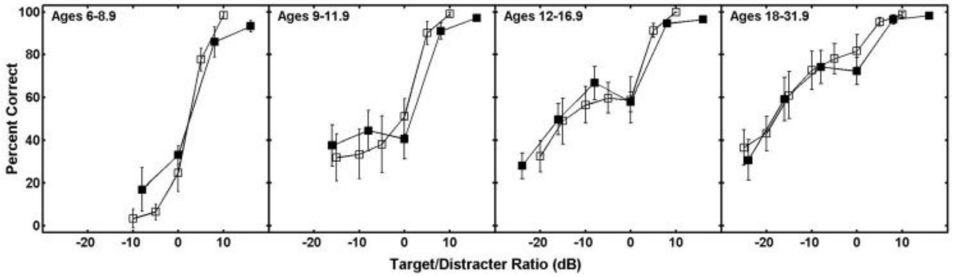

FIG. 2.

Mean psychometric functions in the audio-only condition of the CRM task for listeners in each of the four age groups. Data from the current study (excluding that from the “outlier” shown in Fig. 1) are represented by open symbols and data from the previous study (Wightman and Kistler, 2005) by filled symbols. The error bars represent 95% confidence intervals for the mean.

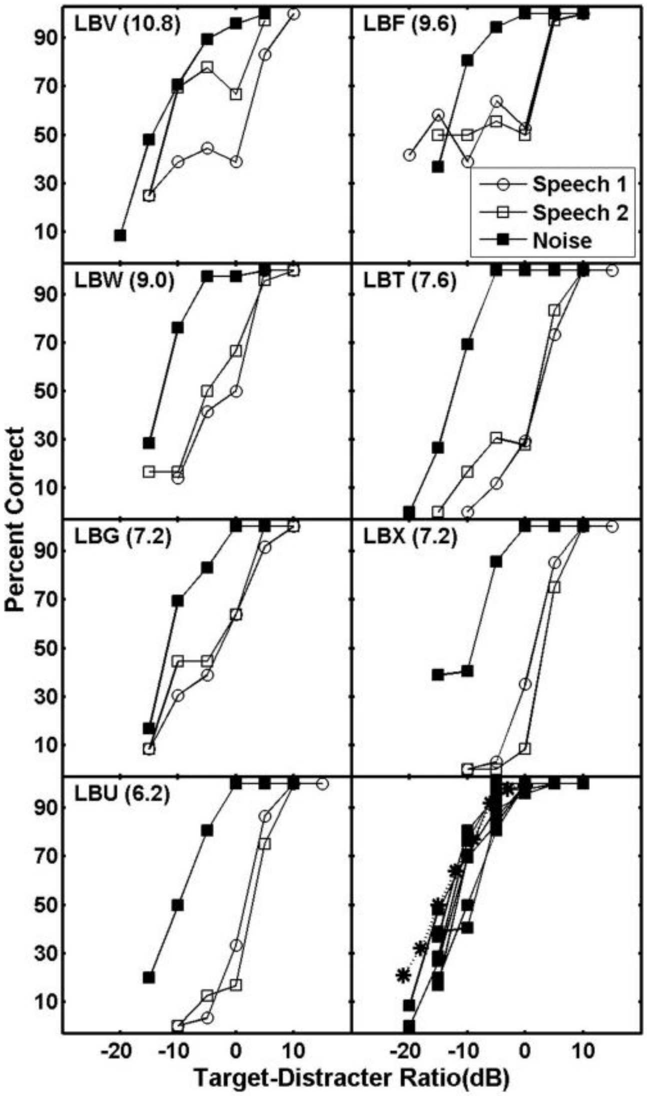

FIG. 3.

Individual psychometric functions from seven children tested with modulated noise as a distracter (audio-only) and retested with the speech distracter. The original data obtained with a speech distracter are labeled “Speech 1” and the retest data are labeled “Speech 2.” The bottom right panel shows all the individual modulated noise functions along with the data obtained in a similar condition by (Brungart, 2001b). Brungart's data are plotted as stars.

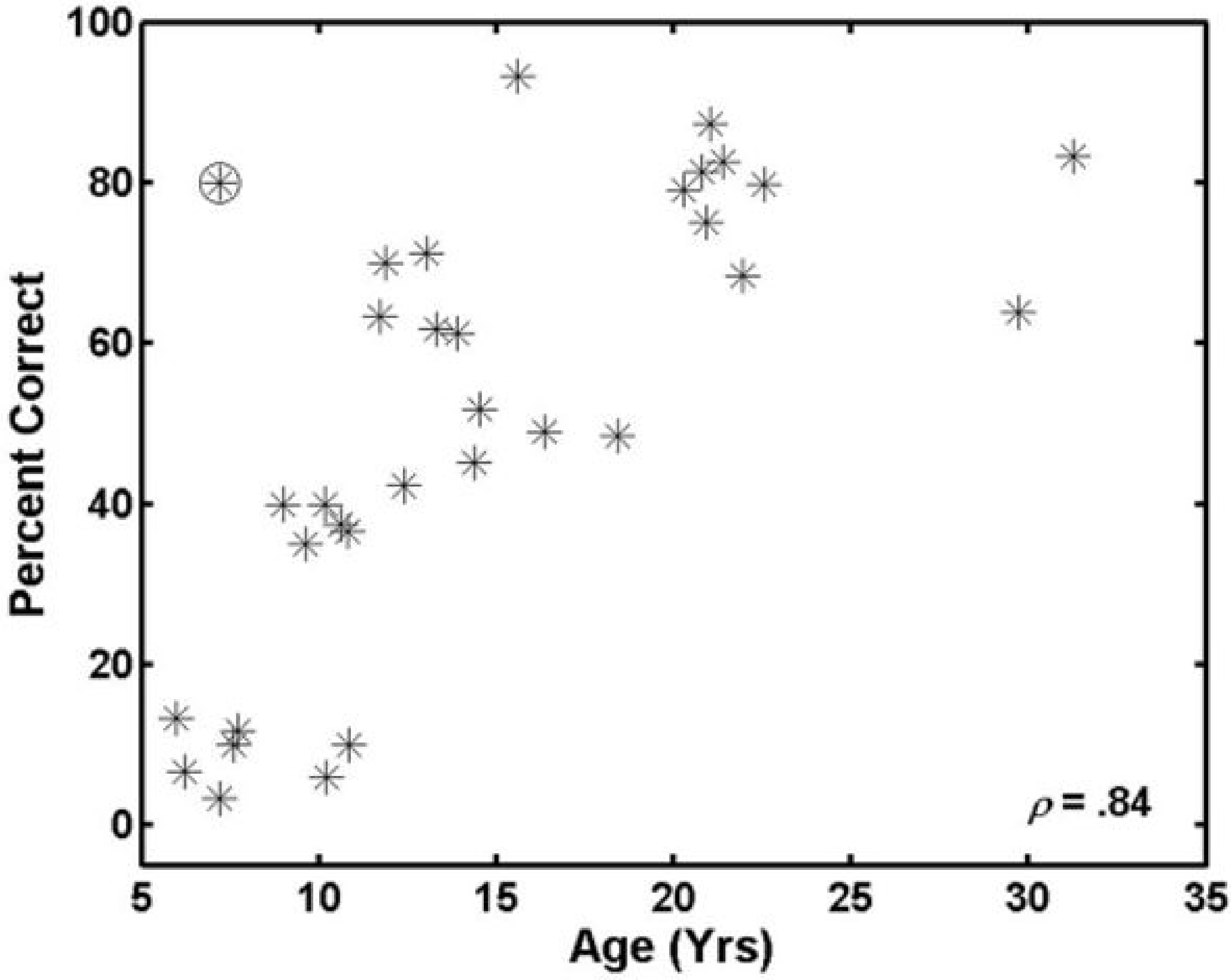

FIG. 4.

Speechreading scores (video-only condition) for all listeners in the current study plotted as a function of age. The circled symbol represents the score for the young listener whose data were excluded from the mean computations shown in Figs. 1, 2, 5, and 6. The correlation listed in the inset is a Spearman *r*, computed with the data indicated by the circled symbol removed.

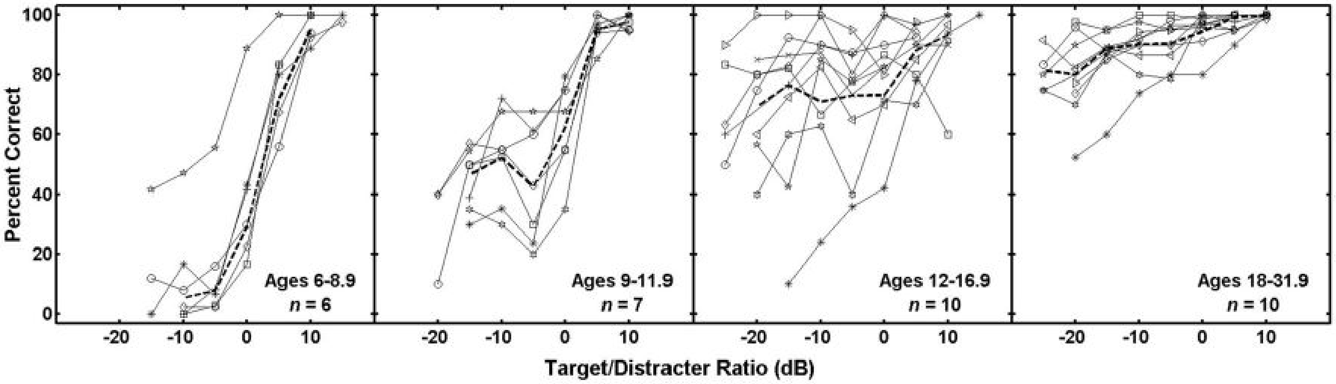

FIG. 5.

Same as Fig. 1, except here the individual psychometric functions are from the audiovisual condition.

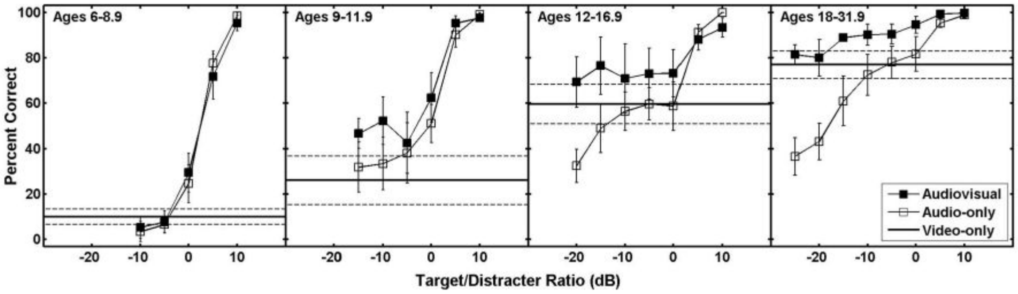

FIG. 6.

Mean psychometric functions in the audio-only condition (open symbols) and the audiovisual condition (filled symbols) of the CRM task for listeners in each of the four age groups. Solid horizontal lines show mean performance in the video-only condition. The error bars and dashed lines represent 95% confidence intervals for the mean.

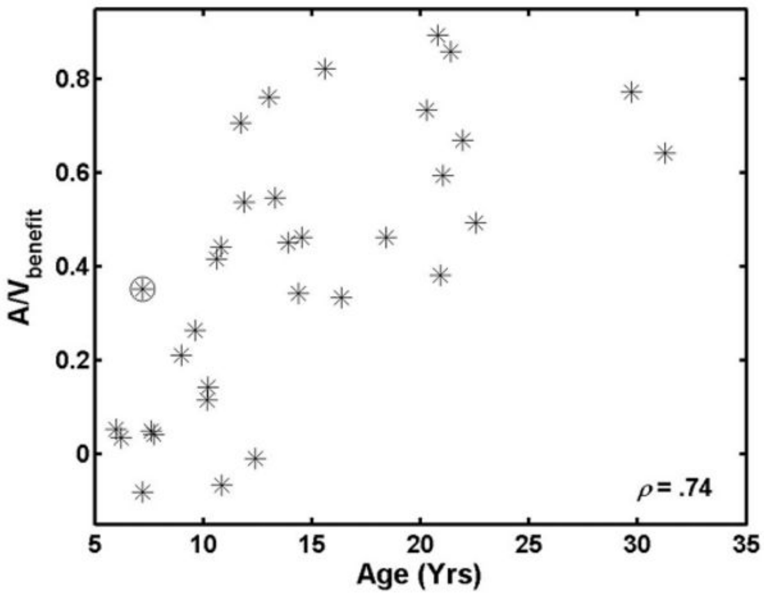

FIG. 7.

A/V$_{\text{Benefit}}$ scores for all listeners in the current study plotted as a function of age. The circled symbol represents the score for the young listener whose data were excluded from the mean computation shown in Figs. 1, 2, 5, and 6. The correlation listed in the inset is a Spearman *r*, computed with the data indicated by the circled symbol removed.

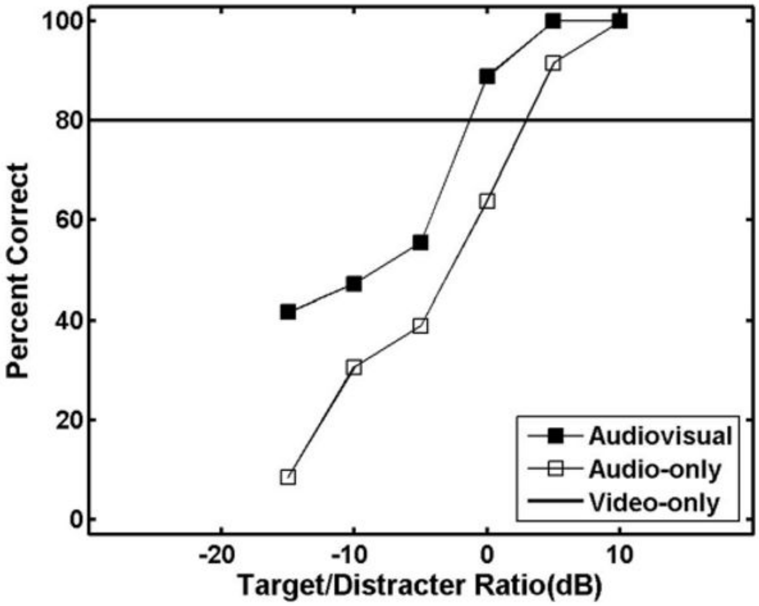

FIG. 8.

Same as Fig. 5 except this figure includes only the data from the 7-years-old who performed better than the others in the same age group. This is the listener whose data were excluded from the mean computation shown in Figs. 1, 2, 5, and 6.