Detection and Attention for Auditory, Visual, and Audiovisual Speech in Children with Hearing Loss

Susan Jerger,[1,2] Markus F. Damian,[3] Cassandra Karl,[1,2] and Hervé Abdi[1]

[1]School of Behavioral Brain Sciences, University of Texas at Dallas, Richardson, TX

[2]Callier Center for Communication Disorders, University of Texas at Dallas, Richardson, TX

[3]School of Psychological Science, University of Bristol, Bristol, United Kingdom

Conflicts of Interest and Source of Funding: No author has any conflicts of interest. This research was supported by the National Institute on Deafness and Other Communication Disorders (NIDCD), grant DC-00421 to University of Texas at Dallas

Corresponding Author: Susan Jerger, School of Behavioral Brain Sciences, GR4.1, University of Texas Dallas, 800 W. Campbell Rd, Richardson, TX 75080, sjerger@utdallas.edu, Phone: 512-216-2961.

Abstract

Objectives. Efficient multisensory speech detection is critical for children who must quickly detect/encode a rapid stream of speech to participate in conversations and have access to the audiovisual cues that underpin speech and language development, yet multisensory speech detection remains understudied in children with hearing loss (CHL). This research assessed detection, along with vigilant/goal-directed attention, for multisensory vs. uni-sensory speech in CHL vs. children with normal hearing (CNH).

Design. Participants were 60 CHL who used hearing aids and communicated successfully aurally/orally and 60 age-matched CNH. Simple response times determined how quickly children could detect a pre-identified easy-to-hear stimulus (70 dB SPL, utterance "buh" presented in Auditory only (A), Visual only (V), or Audiovisual (AV) modes). The V mode formed two facial conditions: static vs. dynamic face. Faster detection for multisensory (AV) than uni-sensory (A or V) input indicates multisensory facilitation. We assessed mean responses and faster vs. slower responses (defined by 1st vs. 3rd quartiles of response-time distributions), which were respectively conceptualized as: faster responses (1st quartile) reflect efficient detection with efficient vigilant/goal-directed attention and slower responses (3rd quartile) reflect less efficient detection associated with attentional lapses. Lastly, we studied associations between these results and personal characteristics of CHL.

Results. ***Uni-sensory A vs. V Modes:*** Both Groups showed better detection and attention for A than V input. The A input more readily captured children's attention and minimized attentional lapses, which supports *A-bound* processing even by CHL who were processing low fidelity A input. CNH and CHL did not differ in ability to detect A input at conversational speech level. ***Multisensory AV vs. A Modes.*** Both Groups showed better detection and attention for AV than A input. The advantage for AV input was facial effect (both static and dynamic faces), a pattern suggesting that communication is a social interaction that is more than just words. Attention did not differ between Groups; detection was faster

in CHL than CNH for AV input, but not for A input. ***Associations Between Personal Characteristics/ Degree of Hearing Loss of CHL and Results.*** CHL with greatest deficits in detection of V input had poorest word recognition skills and CHL with greatest reduction of attentional lapses from AV input had poorest vocabulary skills. Both outcomes are consistent with the idea that CHL who are processing low fidelity A input depend disproportionately on V and AV input to learn to identify words and associated them with concepts. As CHL aged, attention to V input improved. Degree of HL did not influence results.

Conclusions. Understanding speech—a daily challenge for CHL—is a complex task that demands efficient detection of and attention to AV speech cues. Our results support the clinical importance of multisensory approaches in order to understand and advance spoken communication by CHL.

During early development, children learn to process multisensory inputs (e.g., auditory and visual speech) interactively, an advance which increases the likelihood that these inputs will be detected rapidly, identified correctly, and responded to appropriately (Lickliter 2011). Rapid detection of multisensory speech is particularly important because real-time speaking rates—140 to 180 words/minute—place significant demands on listeners' speed of processing (Wingfield et al 2005). Clearly children with hearing loss (CHL) who are processing lower fidelity speech could easily become lost in conversation if they cannot detect the speech input as rapidly as it occurs. Such an inability could be problematic because deficient lower-level skills, such as detection, can have cascading effects that produce higher-level difficulties, as illustrated by the speech, language, and educational difficulties observed in CHL of early onset and by the delayed expressive language skills observed in children with visual impairments of early onset (e.g., McConachie & Moore 1994; Briscoe et al 2001; Jerger et al 2006; Stevenson et al 2017).

In short, proficient multisensory speech detection is critical for CHL to have access to the audiovisual cues that underpin speech and language development, yet we lack evidence about multisensory speech detection by CHL. This research addresses this gap in the literature. Such information is critical for developing effective intervention strategies that mitigate the effects of hearing loss on spoken word recognition and language development. Below we review the literature on multisensory detection by CHL and children with normal hearing (CNH).

Multisensory Detection

Multisensory speech detection does not appear to have been studied previously in CHL. In CNH, one study reported that 6-8-yr-olds showed an adult-like detection advantage for audiovisual relative to auditory speech (LaLonde & Holt 2016). Finally, one study in infants/todders with mild-moderate HL indicated that they detect the correspondences between auditory and visual speech just as infants with NH (Bergeson et al 2010). Specifically, when infants/toddlers with HL heard a word while watching

images of two talkers, one mouthing the heard word and one mouthing a different word, they looked longer at the matching visual speech. Because few studies of multisensory speech detection exist, we also reviewed the literature on multisensory non-speech detection (e.g., a tone and a light presented simultaneously vs. alone). This literature utilized our experimental approach, detection as assessed by *simple response time*, so we will digress briefly to explain this concept.

Simple response time, or the minimal time needed to detect and respond to a stimulus, is a basic measure of speed of processing (Woods et al 2015). It requires participants to detect as quickly as possible the onset of a pre-identified stimulus at a pre-known location and execute a pre-programed motor response. Thus the only uncertainty involved is the time between stimulus presentations. Simple response time primarily involves sensory and motor factors, along with some influence of a participant's general alertness (Luce 1991; Seitz & Rakerd 1997; Woods et al 2015). A difference between detection as measured by simple response time vs. the more traditional *threshold* approach is that the stimulus is usually easy to hear or see. Understanding the speed of detection of conversational-level speech input seems a critical area of research for understanding everyday speech processing by CHL.

With regard to the findings for the non-speech inputs, CNH detected simultaneous auditory and visual inputs faster than either uni-sensory input—in a manner resembling adult-like multisensory facilitation by about 14-yrs of age (e.g., Brandwein et al 2011). Only one study exists in CHL, which observed multisensory facilitation of simultaneous auditory and visual non-speech inputs in early-implanted cochlear implant users of about 11-yrs (Gilley et al 2010). These results with auditory and visual non-speech inputs are important as a whole for understanding the multisensory interactions that can enhance detection. However, they are not directly relevant to this research because the detection of multisensory non-speech vs. speech is differentially influenced by the "unity effect" (e.g., Chen & Spence 2017). This effect indicates that—in many conditions—the multisensory interactions influencing detection occur significantly more often for inputs from a common origin (i.e., auditory + visual speech

dimensions united by properties of the same vocal tract) than from separate origins (i.e., tone + light).

In short, proficient multisensory speech detection is critical for CHL who must quickly detect and encode a rapid stream of speech to participate in everyday conversations and to have access to the audiovisual cues that underpin speech and language development. Yet we lack evidence about multisensory speech detection by CHL. Such information is critical for developing effective interventions that mitigate the effects of hearing loss on spoken word recognition and language development.

Current Study

Our research assessed detection as quantified by *simple response time* of uni-sensory (auditory or visual) vs. multisensory (audiovisual) speech in CHL vs. CNH. We hypothesized that some of the currently unexplained individual differences characterizing spoken word recognition and language development in CHL may reside in this foundational skill supporting speech perception. The stimulus in our study consisted of the single utterance "buh" presented in auditory (A) only, visual (V) only, and audiovisual (AV) modes. Our primary research questions were whether children would show enhanced detection of multisensory relative to uni-sensory speech and whether the relationship between the two uni-sensory speech modes would be altered in the CHL due to the degraded fidelity of the A mode.

Another aspect of this research was that our V input consisted of either the dynamic V speech that produced the utterance "buh," or the talker's static face. We included a static face not only as a control condition but also because previous studies have observed some differences between dynamic articulating vs. static faces. As examples: on *f*MRI scans, a dynamic face generates more extensive cortical activation than a static face (Calvert & Campbell 2003; Campbell et al 2001); adults with NH—viewing a talker's dynamic vs. static face—monitor for a syllable in the A mode significantly better when they view the articulating face (Davis & Kim 2004); and, although both a dynamic face and a V symbol enhance the detection of A speech in adults with NH, the dynamic face produces a relatively greater degree of multisensory facilitation (Bernstein et al 2004; see Tjan et al 2013, for qualifications).

Finally, we should note that dynamic faces are also more ecologically valid because they correspond to everyday social interactions. For example, adults with NH recognize emotional expressions and infants with NH recognize unfamiliar faces more accurately when the facial stimuli are dynamic rather than static (Alves 2013; Otsuka et al 2009), perhaps because motion may enhance the perceptual processing of faces and thus produce richer mental representations (e.g., O'Toole et al 2002). The V speech may also act as a type of alerting mechanism that boosts vigilant attention and helps children detect input faster (e.g., Campbell 2006). This overall evidence predicts that performance in children may benefit more from the dynamic articulating face than the static face and that we may observe some effects of vigilant attention on the dynamic vs. static faces. Vigilant attention for our task may be defined as the ability to sustain goal-directed attention on an unchallenging, monotonous task that involves simple cognitive abilities and a simple motoric response (e.g., Langner & Eickhoff 2013). Goal-directed attention may be defined as the ability to focus attention on a stimulus and/or location according to task demands (e.g., Corbetta & Shulman 2002). We aggregated these two interrelated varieties of attention into one construct to discuss how they may influence performance on our task.

Vigilant/Goal-Directed Attention

Attention affects performance on behavioral tasks (e.g., Whyte 1992). These attentional effects, however, can be challenging to assess directly because attention 1) can be difficult to separate from the other cognitive skills involved in the task, and 2) has a fluctuant nature that makes its effects variable (Fritz et al 2007; Cooley & Morris 1990). That said, simple speed of processing tasks, as used herein, can offer valuable insights about attention from the speed and variability of responses. Speed of processing tasks consistently have fluctuant responses (faster vs. slower), and fluctuations in vigilant/goal-directed attention are thought to be associated with this variability in responding (e.g., McVay & Kane, 2012). We elaborate subsequently (in data analysis section) the characteristics of response-time distributions and how researchers have conceptualized the faster vs. slower responses. Now, however, we consider only

the periodically slowed responses, which are thought to be associated with lapses of attention (e.g., Hervey et al 2006; Luce 1991; Whelan 2008; Langner & Eickhoff 2013).

Historically, researchers have viewed these slowed responses as "noise" and have discarded them from data analysis. More recently, however, studies have emphasized that the slowed responses can be informative about attention: for example, the number of slowed responses can serve as an index of the number of momentary attentional lapses (e.g., Weissman et al 2006; Key et al 2017; Lewis et al 2017). Such studies in CHL have indicated that—relative to a pre-test baseline—both CNH and CHL exhibited more slowed response times and thus more lapses of attention after effortful A speech tasks (Key et al 2017; Gustafson et al 2018). Hearing status did not differentially affect the slowed responses. Age, however, did: younger children found it more difficult to maintain vigilance and task goals. Younger children may find a simple response task particularly taxing because their immature frontal-cortex function may limit the use of more automatic strategies (Thillay et al 2015). Children's capacity to maintain vigilance and task goals improves up to the preteen/teenage years, with much of the developmental change occurring before 10–11 years (e.g., Betts et al 2006; Thillay et al 2015). Thus, we predict that age, but not hearing status, will affect vigilant/goal-directed attention on our task: Younger children, re: older children, will show more lapses of attention and thus more slowed responses. In addition to investigating how uni-sensory vs. multisensory speech detection and vigilant/goal-directed attention may be altered in CHL, we also assessed how degree of hearing loss and personal characteristics of CHL were related to vigilant/goal-directed attention and detection.

Individual Variability in Detection and Vigilant/Goal-Directed Attention

To analyze effects of the degree of hearing loss, we determined the difference in performance between HL subgroups with poorer *vs.* better hearing sensitivity. Further, we investigated the relation between detection and vigilant/goal-directed attention *vs.* A word recognition, vocabulary knowledge, V perception, age, and degree of hearing loss. We are not aware of any previous research on the

associations between multisensory speech detection and personal characteristics of CHL. However, our program of research has shown some relevant related associations concerning word identification and vocabulary.

First, a previous study in CHL, which evaluated whether the influence of V speech on discrimination predicted the influence of V speech on identification, revealed that discrimination scores were associated with the CHL's ability to identify speech onsets and—to a lesser extent—A words, even after the variation due to other relevant variables was controlled (Jerger et al 2017a). We qualified the latter association because it did not achieve statistical significance (p = .06), but it seems relevant because our statistical approach was stringent and constrained prediction to only that variance which was *uniquely* shared between discrimination and A word identification. Such results extended the findings of A-only studies that observed an association between phoneme discrimination and phoneme identification/ vocabulary skills in CHL and CNH/infants with NH (Jerger et al 1987; Briscoe et al 2001; Tsao et al 2004; Lalonde & Holt 2014). This evidence suggests that we may see an association between another lower-level process, detection, and word identification.

Second, a study with a picture-word naming task documented that the mode of input (A vs. AV) influenced semantic access in CHL (Jerger et al 2013). We found that semantic access by A speech in CHL was deficient. However, when V speech was added to the A speech, results changed and semantic access by AV speech in CHL now showed the normal pattern. Our study of speech discrimination in CNH (Jerger et al 2018b) found that the influence of V speech on discrimination uniquely predicted receptive vocabulary skills. These results suggest that we may see an association between the influence of V speech on detection and vocabulary knowledge. Below we elaborate how the uni-sensory vs. multisensory response times were assessed with two complementary analyses.

Data Analyses

The analysis of simple response times traditionally relies on a measure of central tendency, typically

the mean (e.g., Laurienti et al 2006; Balota et al 2008). Thus, in the first analysis, we analyzed mean response times in the CHL vs. CNH. Subsequently, however, we augmented this traditional approach with an analysis of the *faster* vs. *slower* response times. Multiple researchers have begun to consider the rich information provided by distributions of response times (Whelan 2008, illustrations in Appendix). Researchers have analyzed these distributions with the ex-Gaussian approach, which yields three measures (Parris et al 2013): *Tau* which indexes distributional differences in the skewed long tail of the right side (i.e., slower response times) and can be used as a measure of the lapses of attention, and *Mu* and *Sigma* which index distributional shifts in the more rapidly rising left side (i.e., faster response times) and can be used as a measure of task performance. The following results illustrate the value of this approach:

> In a neuropsychological study, mean response times on a Go/No Go task were slower in individuals with Attention Deficit/Hyperactivity Disorder (ADHD) than in the control group (Hervey et al 2006). Ex-Gaussian analysis of response time distributions, however, revealed that individuals with ADHD did not respond slower than the control group when only the faster response times were considered; instead the difference between groups occurred in the long tail of the right side (i.e., more slowed responses in individuals with ADHD). Results were interpreted as indicating that individuals with ADHD are not slower in responding but instead are more prone to attentional lapses.
>
> In a psycholinguistic project, participants named pictures (e.g., camel) in the presence of semantically-related words (e.g., donkey) vs. semantically-unrelated words (e.g., biscuit, Scaltritti et al. 2015). As expected, mean picture-naming times were slower in the presence of the semantically-related words (called semantic interference effect). Ex-Gaussian analysis of response time distributions indicated that the semantic interference effect was significantly reduced in the faster responses (when attention was operating efficiently) and significantly enlarged in the slower responses (when attention was not operating efficiently). Results were interpreted as indicating that attention is critical for resolving semantic interference.

Results such as the above support the following: The faster responses (left rising side of distribution) reflect efficient task behavior with efficient vigilant/goal-directed attention and the slower responses (right tail of distribution) reflect less efficient task behavior associated with attentional lapses (see

Scaltritti et al 2015; Tse et al 2010; Zhou & Krott 2016).

A limitation of the application of the ex-Gaussian analysis is that a large number of trials per participant and per condition are required (e.g., Heathcote et al 1991). Thus some researchers have valued an alternative approach that does not have this limitation: quantile analysis, in which conditions/groups of interest are compared at specific quantiles (Balota et al 2008). That is the approach of the current research and is detailed later. Our analyses are introduced by Data Analytic Sections and Research Questions.

Method

Participants

Participants were 60 CHL with early-onset sensorineural loss (47% boys) and 60 CNH (51% boys). The CNH group—with a corresponding mean and distribution of ages—was formed from a pool of 115 typically-developing children from associated projects (e.g., Jerger et al 2016; Jerger et al 2017a & b; Jerger et al 2018a & b). Ages (yr;mo) ranged from 4;3 to 14;9 (*M* = 9;2, *SD* = 3;1) in CHL and 4;2 to 14;6 (*M* = 9;3, *SD* = 3;1) in CNH. The racial distributions in CHL-CNH were, respectively, 71%-87% Whites, 22%-03% Blacks, and 7%-8% Asian in CHL and 0%-2% Multiracial. All participants met the following criteria: 1) English as native language, 2) ability to communicate successfully aurally/orally, and 3) no diagnosed or suspected disabilities other than HL and its accompanying speech and language problems.

Audiological Characteristics. Hearing sensitivity in the CNH at hearing levels (HLs) of 500, 1000, and 2000 Hz (pure-tone average, PTA; ANSI 2010) averaged 2.53 dB HL (*SD* = 4.31, right ear) and 3.67 dB HL (*SD* = 5.24, left ear). The PTAs in the CHL averaged 45.11 dB HL (better ear) and 57.47dB HL (poorer ear). The PTAs on the better/poorer ears respectively were distributed as follows: ≤ 20 dB (10% / 03%), 21-40 dB (30% / 23%), 41-60 dB (35% / 36%), 61-80 dB (22% / 20%), 81-100 dB (03% / 10%), and greater than 100 dB (0% / 8%). The CHL with PTAs of ≤ 20 dB had losses in restricted frequency regions. Hearing aids were used by 88% of the CHL. Participants who wore amplification were tested while wearing their

devices, which were mostly self-adjusting digital aids with the volume control either turned off or nonexistent. The estimated age at which the CHL who wore amplification received their first aid averaged 2.65 yrs (*SD* = 1.75); the estimated duration of device use averaged 7.80 yrs (*SD* = 3.40). The aided PTA averaged 20.16 dB HL; the aided PTAs were distributed as follows: ≤ 10 dB (8%), 11-20 dB (49%), 21-30 dB (34%), and 31-40 dB (9%). Seventy-six percent (76%) of CHL were mainstreamed in a public school setting and 24% were enrolled in an aural/oral school.

Comparison of Groups. Table 1 compares performance in the CHL vs. CNH on a set of verbal and nonverbal measures. A subset of the measures (vocabulary, V perception, and lipreading onsets) was analyzed with Mann Whitney *U* tests (Hettmansperger & McKean 1998), which were applied because the variances of the groups differed significantly (Levene test, National Institute of Standards & Technology [NIST] 2012). We did not include articulatory proficiency and A word recognition in the analyses because more than half of the CHL and CNH had few errors: respectively ≤ 1 error and > 90% correct. Numerically, average results for articulatory proficiency and A word recognition were poorer in CHL than CNH, a result consistent with previous findings (e.g., Jerger et al 2002). Results of the *U*-tests indicated that the CNH had significantly better vocabulary skills and V perception. The difference between groups in verbal skills was expected, but the difference in V perception was unexpected and is not easily explained. Note, however, that V perception in *both* groups was within the average normal range, and lipreading the onsets did not differ between groups.

Materials and Instrumentation: Stimuli and Response Times

Recording. The stimulus /buh/ was recorded—as a Quicktime movie file—by an 11-yr-old boy actor with clearly intelligible speech. His full facial image and upper chest were recorded, and he started and ended each utterance with a neutral face/closed mouth. The color video signal was digitized at 30 frames/s with 24-bit resolution at a 720 × 480 pixel size. The A signal was digitized at a 48 kHz sampling rate with 16-bit amplitude resolution. The video track was routed to a high-resolution computer

monitor, and the A track was routed through a speech audiometer to a loudspeaker atop the monitor. The stimulus was edited to begin with the frame containing the A onset. The talker's lips in this beginning frame remained closed but were no longer in a neutral position.

Stimuli. The stimulus /buh/ was presented in three modes: AV, A, and V. For the AV mode, children saw and heard the talker; for the A mode, the computer screen was blank; and for the V mode, the loudspeaker was muted. Testing with these modes was carried out in two separate conditions: 1) a dynamic face articulating the utterance and 2) a static face (i.e., the video track was edited, with Adobe Premiere Pro, to contain only the talker's still face and upper chest; the A track remained the same). Hence, the two conditions consisted of: 1) AV dynamic face, V dynamic face, and A (no face); 2) AV static face, V static face, and A (no face). The A stimuli are the same in both facial conditions, thus allowing us to estimate test-retest reliability.

We formed one list of 39 test items (13 in each mode) for each facial condition (each list was presented forwards and backwards to yield two variations). The items of each list were randomized with the constraint that /buh/ was presented once in each mode for each triplet of items (e.g., two-triplet sequence = A/ AV/ V/ V/ A/ AV). This design assured that any changes in performance due to personal factors (e.g., fatigue, practice) were distributed over all modes equally.

Response Times. The computer triggered a counter/timer (resolution less than one ms) at the initiation of each stimulus. The stimulus continued until pressure on a response (telegraph) key stopped the counter/timer. The response board contained two keys separated by approximately 12 cm. A green square beside each key designated the start position for the child's hand, assumed before each trial. The key corresponding to the response (right vs. left) was counterbalanced across participants; a small box covered the unused key.

Procedure

These data were gathered as part of a larger protocol with three testing sessions of about one hour

each. The three sessions occurred on 3 separate days for 100% of CNH and on 1 (16%), 2 (40%), or 3 (44%) days for CHL. The interval between sessions averaged 12 days in each group. The current data were gathered in one session, with the presentation order of the two facial conditions counterbalanced across participants within groups and separated by about 30 minutes. For this testing, a tester sat at a computer workstation and initiated each trial, in an arrhythmic manner, by pressing a touch pad (out of children's sight). The children sat at a distance of 71 cm directly in front of an adjustable height table containing the computer monitor and loudspeaker. A co-tester sat alongside to keep the children on-task: operationally defined as seated erect and alert in the chair without shuffling, head and body oriented toward the monitor/loudspeaker with a visible focus on the monitor, and hand on the start position poised to respond. The co-tester encouraged the children's alertness, focus, and response readiness with a posture of interest in their performance and occasional comments (e.g., "nice"). No trial was initiated until both the tester and co-tester agreed that the child appeared on-task. Flawed responses were deleted online and re-administered at the end of the list (rarely, the equipment or child did not function properly, e.g., child removed hand from start position to scratch).

The children were told that they would sometimes hear, sometimes see, and sometimes hear and see a boy. When they heard the boy, he would always be saying /buh/. When they saw the boy, however, they would either see a movie or photo (i.e., dynamic or static face) of the boy. Before each facial condition, the children were shown the stimulus in each mode (A, V, and AV). The children were told to push the key as fast as possible to the onset of any of these targets with a whole hand response. Each child was told to always start with his or her hand on the green square and, after each trial, to put his or her hand back on the square to get ready for the next trial. Prior to the administration of each facial condition, practice trials were administered until response times had stabilized across a two-triplet sequence. The children's view of the talker's face subtended a visual angle of 7.17° vertically (eyebrow-chin) and 10.71° horizontally (eye level). The children heard the A input at a conversational intensity

level, approximately 70 dB SPL.

Finally, all trials were completed by 100% of CNH and 70% of the CHL. The CHL with incomplete data had, on average, 2.63% missing trials. The missing trials were distributed as follows: 49% (static face) and 51% (dynamic face); 32% (V mode), 32% (A mode), and 35% (AV mode). This research was approved by Institutional Review Boards of University of Texas at Dallas and Washington University in St. Louis.

Mean Performance

Data Analysis

We compared mean response times in the three modes for each facial condition in the CNH and CHL. This traditional measure of response times is shown in Figure 1 because it clearly portrayed how performance differed between the groups and the modes. However, for all statistical analyses, the response times of each participant were rank transformed because the variances of the groups differed significantly (Levene test, NIST 2012). The value of the rank transformation is that it provides the general applicability of non-parametric procedures to parametric procedures such as the analysis of variance (Hettmansperger & McKean 1998). To control for the possibility of false-positive findings (i.e., Type 1 errors), we adjusted the alpha levels for *all* of the subsequent statistical procedures with the Bonferroni correction (Abdi 2007).

Our research questioned whether the children's response times differed 1) for the two uni-sensory inputs and 2) for the AV input vs. the fastest uni-sensory input (as per the model for multidimensional stimuli, e.g., Biederman & Checkosky 1970; Mordkoff & Yantis 1993). Both types of faces were viewed as multidimensional stimuli because individuals can accurately match unfamiliar voices to both dynamic and static unfamiliar faces well above chance, which demonstrates that voices share source-identity information with both types of faces (e.g., Mavica & Barenholtz 2013; Smith et al 2016). Further, our participants were familiar with the talker's face and voice from the other tasks of our protocol. Research questions were: 1) Do response times differ for A vs. V uni-sensory inputs? 2) Do children respond faster

to multisensory AV input than the fastest uni-sensory input? 3) Does the facial condition affect performance? 4) Does performance differ in CNH and CHL? And 5) Do the children respond reliably?

Results

Figure 1 compares mean response times in the A, V, and AV modes for the static and dynamic faces in CHL vs. CNH. Statistical results (Table 2) revealed a significant effect of Facial Condition and Mode. The Facial Condition effect occurred because response times (collapsed across Group and Mode) were slightly but reliably faster for the dynamic than static face (600 ms vs. 630 ms). The Mode effect occurred because response times (collapsed across Group and Facial Condition) were significantly faster for the A and AV modes (582 ms and 554 ms) than the V mode (713 ms). A straightforward interpretation of these general results was complicated, however, because the Facial Conditions affected results for some Modes but not Others, producing a significant Mode × Facial Condition interaction. More specifically, whereas mean response times (collapsed across Group) for V input were faster for the dynamic than the static face (691 ms – 735 ms), response times for the A and AV inputs did not differ in the facial conditions (584 – 580 ms for A and 552 – 557 ms for AV). No other significant difference was observed. Below, we analyzed whether the uni-sensory inputs differed (V vs. A) and whether the addition of visual speech influenced performance (AV vs. fastest uni-sensory input). The above statistical results allowed us to address the relation between uni-sensory inputs.

V vs. A Modes. The above significant Mode effect indicated that both groups responded faster to A than V input (see Figure 1). The above finding of significantly faster responses for the dynamic than static face for V input but not for A input (Mode × Facial Condition interaction) also produced a smaller difference between V and A response times for the dynamic than the static face in both groups: difference scores (V – A) for dynamic vs. static faces respectively of 118 ms vs. 173 ms (CNH) and 96 ms vs. 137 ms (CHL). These data indicated that A responses were the fastest uni-sensory mode in both groups and, thus, the A mode served as our uni-sensory baseline for determining whether multisensory

input influenced performance.

***AV vs. A Modes*.** To address this question, we carried out paired *t* tests on the A vs. AV response times in each group for each facial condition. The results, summarized in Table 3, revealed a different pattern in the CHL and CNH. Specifically, CHL showed faster detection of the AV input for both the static and dynamic faces: a general facial effect. In contrast, CNH showed faster detection of the AV input only for the dynamic face.

Reliability. To assess test-retest performance for A response times, we reformatted the data to represent the first vs. second tests (the two facial conditions were counterbalanced such that each occurred as the first test ½ of the time). Rank transformed response times were statistically evaluated with a mixed-design analysis of variance with one between-participant factor (Group: CHL, CNH) and one within-participant factor (Test: first, second). Results did not show any significant effects or interactions. The mean A response times for the first vs. second tests were respectively 619 ms vs. 581 ms (CHL) and 568 ms vs. 560 ms (CNH). A follow-up simple regression in each group indicated that the children's A response times for the first vs. second tests were significantly correlated, CHL: $r = .780$, F 1, 58 = 90.34, $p < .0001$; CNH: $r = .814$, F 1, 58 = 113.58, $p < .0001$.

Faster vs. Slower Response Times

Data Analysis

We explored the faster vs. slower times with response time distributions computed by Vincentile analysis, a nonparametric technique that preserves the component distributions' shapes and does not make any assumptions about underlying distributions (e.g., Ratcliff 1979). Vincentile analysis is recommended for data such as ours because it yields stable estimates even when there are only 10–20 responses per participant/mode/condition. To obtain the Vincentile distributions, each child's response times—for each mode/condition—were rank-ordered. For illustrative purposes, we initially divided the rank-ordered response times into sequential bins of 10% (deciles) and obtained a ***cumulative***

distribution function (CDF) for each group by averaging each of the bins across its participants for each facial condition/mode. Figure A1 (Appendix) illustrates these CDFs for the A, AV, and V modes in the static and dynamic facial conditions for CHL and CNH. Conversely, for data analytic purposes—in which we compared the conditions/modes at two specific locations on the distribution—we divided each child's rank-ordered response times into quartiles or sequential bins of 25%. We analyzed the children's response times at the 1st and 3rd quartiles because the interquartile range is considered a robust measure of the dispersion of a distribution (Whelan 2008).

This quantile approach allowed us to assess whether the effects produced by the conditions/modes changed as a function of their location on the distribution (e.g., Balota et al 2008). And, because our data are simple response times (wherein fluctuations in the speed of responding are associated with fluctuations in the effects of attention on performance), a quantile analysis also provided the opportunity to investigate our questions with the assumption that: The faster responses (1st quartile) reflect efficient detection with efficient vigilant/goal-directed attention and the slower responses (3rd quartile) reflect less efficient detection associated with attentional lapses. Research questions were: 1) Do the A vs. V uni-sensory inputs differ at one or both quartiles? 2) Do the multisensory AV vs. fastest uni-sensory inputs differ at one or both quartiles? 3) Does the facial condition affect results? And, 4) Does hearing loss affect results?

V vs. A Modes. Figure 2 shows V vs. A response times in the CHL and CNH for the static and dynamic faces at the 1st and 3rd quartiles. Statistical results (Table 4) revealed a significant main effect for Quartile and Mode. The main effect of Quartile was not of interest because results at the 3rd quartile would, by definition, be slower than results at the 1st quartile, but the main effect of Mode strongly supported the previous results for mean performance: the children consistently responded faster to A than to V input. The current analysis, however, indicated significant interactions between the Quartile × Group and Mode × Group. These interactions were probed with Mann-Whitney *U* tests, which indicated

the following: The Quartile × Group interaction occurred because response times (collapsed across Mode and Facial Condition, see "All," Figure 2) were significantly faster in the CHL than in the CNH at the 1st quartile, but did not differ in the groups at the 3rd quartile. The Mode × Group interaction occurred because response times (collapsed across Quartile and Facial Condition) were significantly faster in the CHL than in the CNH for the V input, but did not differ in the groups for A input. No other significant effect was observed.

AV vs. A Modes. Figure 3 shows the AV vs. A response times in the CNH and CHL for the static and dynamic faces at the 1st and 3rd quartiles. Statistical results (Table 5) again revealed a significant main effect for Quartile and Mode. The main effect of Quartile was, as noted previously, predictable, but the main effect of Mode yielded new information, which indicated that the children responded faster to AV input than A input (imagine results for each mode collapsed across Quartile and Facial Condition, Figure 3). The interpretation of these overall effects was again complicated, however, by significant interactions between the Quartile × Group and the Mode × Group. These interactions were explored with Mann-Whitney *U* tests, which indicated the following: The Quartile × Group interaction occurred because response times (collapsed across Mode and Facial Condition, see "All") were significantly faster in the CHL than in the CNH at the 1st (detection) quartile, but not at the 3rd (attention) quartile. The Mode × Group interaction occurred because response times were significantly faster in CHL than CNH for AV input, but did not differ in the Groups for A input (imagine results collapsed across Quartile and Facial Condition, Figure 3).

Effect of Degree of Hearing Loss. To address whether results in the CHL differed as a function of the degree of HL, we divided the CHL into better vs. poorer hearing sensitivity subgroups based on the PTA score on the best ear. The better vs. poorer subgroups (*N*=30 each) had average PTA scores as follows: Best Ear: 1) 29.55 dB HL (*SD* = 11.09) vs. 60.67 dB HL (*SD* = 12.66); Worst Ear: 2) 43.44 dB HL (*SD* = 23.01) vs. 71.50 dB HL (*SD* = 18.22). The age in the better vs. poorer subgroups averaged 9.23 yrs (*SD* = 3.07) vs.

9.19 yrs (*SD* = 3.00). To analyze effects of the degree of hearing loss, we determined the difference between the mean response times in the poorer minus better HL subgroups: for the A, V, and AV modes at the 1st and 3rd quartiles in the static and dynamic facial conditions. Figure 4 portrays these results. The error bars are the 95% Confidence Intervals (CIs), or the range of plausible values, for the difference scores between the two independent means (Sullivan 2017). If the 95% CI contains zero, performance does not differ significantly in the subgroups. As seen in Figure 4, all of the CIs contained zero. To supplement these findings, we carried out a mixed-design ANOVA with the A, V, and AV response times (for both facial conditions and both quartiles) in the better vs. poorer HL subgroups, which also did not reveal any significant differences between the subgroups nor any significant interactions. Thus, analyses from two approaches showed that differences in the degree of hearing loss did not influence findings.

Associations Between Personal Characteristics of CHL and Uni-Sensory / Multisensory Effects

We carried out separate multiple regression analyses to probe possible *unique* associations between selected descriptors of the CHL and the effects of V or AV input relative to A input at the 1st and 3rd quartiles. We defined "unique" statistically by the part correlations, which express the independent contribution of a variable after controlling for all the other variables (Abdi et al 2009). The dependent variable was the difference (in ms) between the V – A response times or the AV – A response times; the independent variables were the standardized scores for age, vocabulary, visual perception, A word recognition, and degree of hearing loss (PTA) on the better ear. Table 6 summarizes statistical findings.

The multiple correlation coefficients and omnibus *Fs* indicated significant associations between the omnibus analyses and all of the descriptors considered simultaneously (excepting AV – A: detection), with the significant multiple correlation coefficients explaining about 20% – 26% of the variability. These multiple correlation coefficients were of less interest, however, than the part correlation coefficients and partial *F* statistics, which evaluated the variation in the difference scores *uniquely* associated with each individual descriptor.

The part correlations for V – A: detection indicated that these difference scores were uniquely associated with the CHL's ability to identify A words. The unique *t* value associated with the partial *F* was negative, which indicated an inverse relation between A word recognition and V – A: detection. In other words, the CHL who showed the largest positive V – A: detection difference scores (greatest slowness in detecting V relative to A input) had the smallest (poorest) word recognition scores.

For V – A: attention, the part correlations indicated that age was uniquely associated with the V – A difference scores. The unique *t* value associated with this partial *F* was also negative, indicating an inverse relation between age and V – A: attention. In other words, the CHL with the smallest (youngest) ages showed the largest positive V – A difference scores, which occurred because the younger children showed more unusually slowed responses for the silent V input due to more attentional lapses.

For AV – A: detection, the part correlations did not reveal any significant associations with the personal descriptors. For AV – A: attention, the part correlations again indicated that age was uniquely associated with the AV - A difference scores, but this time, the *t* value unique to the partial *F* was positive. The CHL with the smallest (youngest) ages showed the largest negative AV – A difference scores (largest AV benefit), which resulted from AV input minimizing attentional lapses and the unusually slowed responses more than A input.

In addition to age, the vocabulary of the CHL was also uniquely associated with the AV – A: attention difference scores. The *t* value unique to the partial *F* was positive, which indicated a direct relation between vocabulary and AV – A: attention. In other words, the CHL with the highest negative AV – A difference scores (greatest reduction in attentional lapses from multisensory AV input) possessed the lowest vocabulary scores.

Discussion

Understanding conversational speech—a daily challenge for CHL—is a complex task that requires listeners to detect and process a rapid stream of speech or become lost in conversation. For CHL, this

not only demands efficient detection skills but also efficient vigilant/goal-directed attention because the perception of degraded speech requires attention (Wild et al 2012). Despite the importance of these efficiencies, however, we know little about how CHL detect and attend to uni-sensory and multisensory speech cues. Thus, this research studied speech detection and vigilant/goal-directed attention for the utterance "buh" presented in A, V, or AV mode in CHL who used hearing aids and communicated successfully aurally/orally. Our V input consisted of both static and dynamic faces, which allowed us to determine whether effects on performance reflected a facial effect (influenced by both faces or only the static face) or an articulating-face-specific effect.

We should note that our task offered some advantages for studying the effects of attention on uni-sensory and multisensory speech detection. As previously mentioned, the effects of attention can be difficult to assess because: 1) attention sometimes cannot be differentiated from the other cognitive skills of a task, and 2) attention fluctuates so its effects are not consistent over time (references above). With regard to the first difficulty, a simple response time is considered one of the simplest measures of processing. A participant is instructed to respond as quickly as possible to the occurrence of the stimulus, and the stimulus, its location, and the response are known *a priori* and do not vary. Thus a simple response time depends mostly on sensory and motor factors rather than cognitive skills. With regard to the second difficulty, a simple response time behavioral task is indeed susceptible to fluctuations in the effects of attention over time as are behavioral tests in general. However, in our research, these fluctuations were of primary interest because fluctuations in the speed of responding are associated with fluctuations in the effects of attention on performance. Thus our experimental design assessed not only traditional mean response times but also the faster vs. slower response times. The faster vs. slower responses were conceptualized as: Faster responses (1st quartile) reflect efficient detection with efficient vigilant/goal-directed attention and slower response (3rd quartile) reflect less efficient detection associated with attentional lapses.

In addition to these advantages, we also want to acknowledge some limitations. One is that we had only 13 trials per participant/condition/mode (78 trials total) due to the limited testing time available with young children. Importantly, however, we analyzed our data with a technique (Vincentizing) that is considered especially well-suited for data with only a few observations per participant/condition/mode (references above). As noted previously, parametric analyses (e.g., ex-Gaussian approach) provide alternatives to Vincentizing for research with hundreds of observations per participant/condition/mode. It is interesting to note, however, that researchers who conduct ex-Gaussian analyses may follow up with quantile analyses to examine the extent to which the ex-Gaussian parameters capture the empirical response time distributions (e.g., Tse et al 2010; Zhou & Krott 2016). Finally, another consideration to note is that some of the slower responses may have been reflecting motivational factors rather than attentional lapses (e.g., Reinvang, 1998). We minimized this possibility, however, by having a co-tester who tried to keep the children engaged in the task. We will discuss the overall results in terms of the uni-sensory inputs (V vs. A), the multisensory vs. the fastest uni-sensory input (AV vs. A), and the association between these results vs. the personal characteristics/degree of hearing loss of CHL.

Mean Performance

Both Groups responded faster to A than V input—a pattern consistent with the non-speech literature indicating that simple response times are faster for the A than V mode (e.g., Woodworth & Schlosberg 1954; Vickers 2007), with no significant difference in results between CHL vs. CNH (e.g., Jerger et al 2016). A silent articulating face (i.e., mouthing) also improved detection in the V mode (relative to a static face) in both Groups. In contrast to these effects, a difference between Groups emerged with regard to whether children responded faster to AV than A input. Whereas CHL showed improved performance (i.e., benefit) from AV input for both static and dynamic faces (a facial benefit), CNH showed improved performance from AV input only for the dynamic articulating face. Responses for A speech in both Groups were reliable. The below results refined these results.

Faster vs. Slower Response Times

V vs. A Modes. Both Groups showed poorer detection and poorer vigilant/goal-directed attention for V than A input. That said, the CHL detected V input significantly faster than CNH, a pattern that may reflect the CHL's educational training and their greater dependence on V input for communication. This significant difference in the detection of V input by CHL vs. CNH was not revealed in the analysis of mean performance. Finally, CHL detected A input at a conversational speech level just as well as CNH.

If we view response times for A input as a baseline, both groups detected V input more efficiently than they sustained attention to this V input. Poorer attention for V input (or better attention for A input) indicated that A input in both Groups more readily captured the children's attention and minimized attentional lapses. This capture of attention by A speech may be particularly helpful in nurturing speech and language development because it would help children perceive talkers' rapidly spoken words, for which they cannot "take another listen." Overall these results strongly endorsed stimulus-bound A processing by these children, even the CHL who were processing lower fidelity A input and who had experienced early A deprivation.

AV vs. A Modes. Both Groups demonstrated consistently better detection and better attention for AV than A input. That said, the CHL benefited more from AV multisensory input (i.e., larger differences between AV and A responses) than the CNH. This outcome is consistent with the long-held idea that V speech benefits low fidelity A speech more than high fidelity A speech. Two other findings were: 1) general overall detection was faster in CHL than CNH whereas attention did not differ between Groups, and 2) general overall response times were generally faster in CHL than CNH for AV input but not for A input.

Finally we should note that the above AV results in these children were facial effects (i.e., no significant difference between the dynamic vs. static face), which implies that the benefit from AV input in these children was a redundancy effect: an effect that may reflect the simultaneous or correlated

onsets interacting to produce a more emphatic onset. This outcome is also consistent with the idea that communication is a social interaction that is more than just words. Children use both perceptual and social cues to learn word and meaning relationships, and facial expressions have an important communicative function (e.g., Rollins 2016). Eye-tracking studies have documented a "social-tuning" pattern (Worster et al 2018, p. 169) in which children look at the eyes before and after speech utterances and at the mouth during utterances. These different areas of the face convey social and emotional cues (e.g., Lansing & McConkie 2003), which may be particularly important to CHL who may have less access to such cues (e.g., intonation) in the lower fidelity A input.

Associations Between Results and Personal Characteristics/Degree of Hearing Loss of CHL. The CHL who showed the greatest deficits in the detection of silent V input had the poorest word recognition skills and the CHL who showed the greatest reduction of attentional lapses from AV input had the poorest vocabulary skills. Both of these outcomes are consistent with the idea that CHL (who are listening to lower fidelity A input) benefit from V and AV input to learn to identify words and associate them with concepts. When the CHL had unusual difficulty detecting V input (larger V – A difference), their ability to learn to identify words was hampered. This finding supports our hypothesis that some of the individual differences in speech recognition by CHL may reside in differences in detection skills. When the CHL had an unusual reduction of attentional lapses by AV input (larger AV – A difference), their ability to learn the meanings of words was hampered. A relation between poorer vocabularies and the greater reduction of attentional lapses by AV input may result from the fact that lower fidelity A input produces more effortful listening (Tharpe et al 2002), which can affect alertness and reduce the stimulation for attention (Nissen 1977); this, in turn, can produce greater attentional lapses (that impair word learning) for uni-sensory A input. Our previous research in CHL clearly revealed that semantic access by A speech was deficient whereas semantic access by AV speech was typical of that in CNH (Jerger et al 2013). The degree of hearing loss did not influenced results.

In short, attention was captured and attentional lapses were minimized more readily by A than V input and by AV than A input, especially in younger children, a pattern which yielded a significant effect of age. As the CHL aged (and perhaps as they received more educational training), they learned to minimize attentional lapses and improve vigilant/goal-directed attention to V input (both uni-sensory and multisensory inputs). Such results are consistent with the literature (see Introduction).

In conclusion, this research investigated detection and attention for multisensory vs. uni-sensory input in CHL and found that 1) AV input improved the speed of detection and reduced attentional lapses in CHL and 2) AV input and V input benefited CHL's ability to learn words. Such findings support the importance of multisensory assessment and intervention strategies to mitigate the effects of hearing loss on spoken word recognition and language development.

Acknowledgements

This research was supported by the NIDCD, grant DC-00421 to University of Texas at Dallas (UT-D). Dr. Abdi acknowledges the support of an EURIAS fellowship at the Paris Institute for Advanced Studies (France), with the support of the European Union's 7th Framework Program for research, and funding from the French State managed by the "*Agence Nationale de la Recherche* (program: *Investissements d'avenir, ANR-11-LABX-0027-01 Labex RFIEA+*)." We thank Dr. Nancy Tye-Murray, Washington University School of Medicine (WUSM), for supervising data collection in CHL, the children and parents who participated, and the research staff who assisted: Aisha Aguilera, Carissa Dees, Nina Dinh, Nadia Dunkerton, Derek Hammons, Scott Hawkins, Brittany Hernandez, Demi Krieger, Rachel Parra McAlpine, Michelle McNeal, Jeffrey Okonye, and Kimberly Periman of UT-D (data collection, analysis, stimuli editing, computer programming) and Drs. Nancy Tye-Murray and Brent Spehar, WUSM (stimuli recording, editing).

References

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 103-107). Thousand Oaks, CA: Sage.

Abdi, H., Edelman, B., Valentin, D., et al (2009). *Experimental design and analysis for psychology*. New York: Oxford University Press.

Alves, N. (2013). Recognition of static and dynamic facial expressions: A study review. *Estudos de Psicologia, 18*, 125-130.

American National Standards Institute (ANSI). (2010). *Specifications for audiometers*. ANSI/ASA S3.6-2010 (R2010). New York: American National Standards Institute.

Balota, D., Yap, M., Cortese, M., & Watson, J. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *J Mem Lang, 59*, 495-523.

Beery, K., & Beery, N. (2004). *The Beery-Buktenica developmental test of visual-motor integration with supplemental developmental tests of visual perception and motor coordination.* (5th ed). Minn: NCS Pearson.

Bergeson, T., Houston, D., & Miyamoto, R. (2010). Effects of congenital hearing loss and cochlear implantation on audiovisual speech perception in infants and children. *Restor Neurol Neurosci, 28*, 157-165.

Bernstein, L., Auer, E., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Commun, 44*, 5-18.

Betts, J., McKay, J., Maruff, P., et al (2006). The development of sustained attention in children: The effect of age and task load. *Child Neuropsychol, 12*, 205-221.

Biederman, I., & Checkosky, S. (1970). Processing redundant information. *J Exp Psychol, 83*, 486-490.

Brandwein, A., Foxe, J., Russo, N., et al (2011). The development of audiovisual multisensory integration across childhood and early adolescence: A high-density electrical mapping study. *Cereb Cortex, 21*, 1042-1055.

Briscoe, J., Bishop, D., & Norbury, C. (2001). Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment. *J Child Psychol Psychiatry, 42*, 329-340.

Brownell, R. (2000). *Expressive one-word picture vocabulary test*, 3rd ed., Acad. Ther. Pub., Novato, CA.

Calvert, G., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *J Cogn Neurosci, 15*, 57-70.

Campbell, R. (2006). Audio-visual speech processing. In K. Brown, A. Anderson, L. Bauer, M. Berns, G. Hirst, & J. Miller (Eds.), *The encyclopedia of language and linguistics* (pp. 562-569). Amsterdam: Elsevier.

Campbell, R., MacSweeney, M., Surguladze, S., et al (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cogn Brain Res, 12*, 233-243.

Cooley, E. & Morris, R. (1990) Attention in children: A neuropsychologically based model for assessment, *Dev*

Neuropsychol, 6, 239-274

Corbetta, M. & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci, 3*, 201-215.

Chen, Y. & Spence, S. (2017). Assessing the role of the 'unity assumption' on multisensory integration: A review. *Front Psychol*, 8, 445, doi: 10.3389/fpsyg.2017.00445.

Dunn, L., & Dunn, D. (2007). *The peabody picture vocabulary test-IV* (4th ed.). Minneapolis, MN: NCS Pearson.

Fritz, J., Elhilali, M., David. S., & Shamma S. (2007). Auditory attention—focusing the searchlight on sound. *Curr Opin Neurobiol, 17*, 1–19.

Gilley, P., Sharma, A., Mitchell, T., et al (2010). The influence of a sensitive period for auditory-visual integration in children with cochlear implants. *Restor Neurol Neurosci, 28*, 207-218.

Goldman, R. & Fristoe, M. (2000). *Goldman Fristoe 2 test of articulation*, Amer. Guidance Ser., Circle Pines, MN.

Gustafson, S., Key, A., Hornsby, B., et al (2018). Fatigue related to speech processing in children with hearing loss: Behavioral, subjective, and electrophysiological measures. *J Speech Lang Hear Res, 61*, 1000-1011.

Heathcote, A., Popiel, S., & Mewhort, D. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychol Bull, 109*, 340-347.

Hervey, A., Epstein, J., Curry, J., et al (2006). Reaction time distribution analysis of neuropsychological performance in an ADHD sample. *Child Neuropsychol, 12*, 125–140.

Hettmansperger, T. & McKean, J. (1998). *Robust nonparametric statistical methods*. Wiley, New York.

Jerger, S., Damian, M., Karl, C., et al (2018a). Developmental shifts in detection and attention for auditory, visual, and audiovisual speech. *J Speech Lang Hear Res, 61,* 3095-3112.

Jerger, S., Damian, M., McAlpine, R., et al (2018b). Visual speech fills in both discrimination and identification of non-intact auditory speech in children. *J Child Lang,* 45, 392-414.

Jerger, S., Damian, M., McAlpine, R., et al (2017a). Visual speech alters the discrimination and identification of non-intact auditory speech in children with hearing loss. *Intl J Pediatr Otorhinologyngol, 94*, 127-137.

Jerger, S., Damian, M., Tye-Murray, N., et al (2017b). Children perceive speech onsets by ear and eye. *J Child Lang, 44*, 185-215.

Jerger, S., Damian, M., Tye-Murray, N., et al (2006). Effects of childhood hearing loss on organization of semantic memory: Typicality and relatedness. *Ear Hear, 27*, 686-702.

Jerger, S., Lai, L., & Marchman, V. (2002). Picture naming by children with hearing impairment: Effect of phonologically-related auditory distractors. *J Am Acad Audiol*, 13:478-492.

Jerger, S., Martin, R., & Damian, M. (2002). Semantic and phonological influences on picture naming by children and teenagers. *J Mem Lang, 47*, 229-249.

Jerger, S., Martin, R., & Jerger, J. (1987). Specific auditory perceptual dysfunction in a learning disabled child. *Ear Hear*, 8, 78-86.

Jerger, S., Tye-Murray, N., Damian, M., et al (2013). Effect of hearing loss on semantic access by auditory and audiovisual speech in children. *Ear Hear*, 34, 753-762.

Jerger, S., Tye-Murray, N., Damian, M., et al (2016). Phonological priming in children with hearing loss: Effect of speech mode, fidelity, and lexical status. *Ear Hear, 37*, 623-633.

Key, A., Gustafson, S., Rentmeester, L., et al (2017). Speech-processing fatigue in children: Auditory event-related potential and behavioral measures. *J Speech Lang Hear Res,*60, 2090-2104.

Kim, J.& Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Commun, 44*, 19-30.

Lalonde, K., & Holt, R. (2016). Audiovisual speech perception development at varying levels of perceptual processing. *J Acoust Soc Am,* 139, 1713–1723.

Lalonde, K., & Holt, R. (2014). Cognitive and linguistic sources of variance in 2-year-olds' speech-sound discrimination: a preliminary investigation. *J Speech Lang Hear Res, 57*, 308-326.

Langner, R., & Eickhoff, S. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychol Bull, 139*, 870–900.

Lansing, C., & McConkie, G. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Atten Percept Psychophys, 65*, 536-552.

Laurienti, P., Burdette, J., Maldijian, J., et al (2006). Enhanced multisensory integration in older adults. *Neurobiol Aging, 27*, 1155-1163.

Lewis, F., Reeve, R., Kelly, S., et al (2017). Sustained attention to a predictable, unengaging Go/No-Go task shows ongoing development between 6 and 11 years. *Atten Percept Psychophys, 79*, 1726-1741.

Lickliter, R. (2011). The integrated development of sensory organization. *Clin. Perinatol. 38,* 591-603.

Luce, R. (1991). *Response times: Their role in inferring elementary mental organization*. Oxford: Oxford University Press.

Mavica, L., & Barenholtz, E. (2013). Matching voice and face identity from static images. *J Exp Psychol Hum Percept Perform*, 39, 307–312.

McConachie, H. & Moore, V. (1994). Early expressive language of severely visually impaired children. *Dev Med Child Neurol 36*, 230-240.

McVay, J. & Kane, M. (2012). Drifting from slow to “D’oh!”: Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *J Exp Psychol Learn Mem Cognit, 38,* 525–549.

Mordkoff, T., & Yantis, S. (1993). Dividing attention between color and shape: Evidence of coactivation. *Atten Percept Psychophys, 53*, 357-366.

National Institute of Standards and Technology/SEMATECH (2002). *e-Handbook of statistical methods*. https://www.itl.nist.gov/div898/handbook/.

Nissen, M. (1977). Stimulus intensity and information processing. *Atten Percept Psychophys, 22*, 338-352.

O'Toole, A., Roak, D., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends Cogn Sci, 6*, 261-266.

Otsuka, Y., Konishi, Y., Kanazawa, S., et al (2009). Recognition of moving and static faces by young infants. *Child Dev, 80*, 1259-1271.

Parris, B., Dienes, Z., & Hodgson, T. (2013). Application of the ex-Gaussian function to the effect of the word blindness suggestion on Stroop task performance suggests no word blindness. *Front Psychol, 4*, 647, doi: 10.3389/fpsyg.2013.00647.

Ratcliff, R. (1979). Group reaction time distributions and analysis of distribution statistics. *Psychol Bull, 86*, 446-461.

Reinvang, I. (1998). Validation of reaction time in continuous performance tasks as an index of attention by electrophysiological measures. *J Clin Exper Neuropsychol, 20*, 885-897.

Rollins, P. (2016). Words are not enough. Providing the context for social communication and interaction. *Topics Lang Dis, 36*, 198-216.

Ross, M., & Lerman, J. (1971). *Word intelligibility by picture identification*. Pittsburgh: Stanwix House, Inc.

Scaltritti, M., Navarrete, E., & Peressotti, F. (2015). Distributional analyses in the picture–word interference paradigm: Exploring the semantic interference and the distractor frequency effects. *Quart J Exp Psychol*, *68*, 1348-1369.

Seitz, P. & Rakerd, B. (1997). Auditory stimulus intensity and reaction time in listeners with longstanding sensorineural hearing loss. *Ear Hear, 18*, 502-512.

Smith, H., Dunn, A., Baguley, T., et al (2016). Matching novel face and voice identity using static and dynamic facial images. *Atten Percept Psychophys*, 78, 868-879.

Stevenson, R., Sheffield, S., Butera, I., et al (2017). Multisensory integration in cochlear implant recipients. *Ear Hear, 38*, 521-538.

Sullivan, L. (2017). *Confidence intervals*. Biostatistics, Boston University School of Public Health. Retrieved from http://sphweb.bumc.bu.edu/otlt/MPHModules/BS/BS704_Confidence_Intervals

Tharpe, A., Ashmead, D., & Rothpletz, A. (2002). Visual attention in children with normal hearing, children with hearing aids, and children with cochlear implants. *J Speech Lang Hear Res, 45*, 403-413.

Thillay, A., Roux, S., Gissot, V., et al (2015). Sustained attention and prediction: Distinct brain maturation trajectories during adolescence. *Front Hum Neurosci, 9*, 519, doi: 10.3389/fnhum.2015.00519.

Tjan, B., Chao, E., & Bernstein, L. (2013). A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. *Eur J Neurosci, 39*, 1323 - 1331.

Tsao, F., Liu, H., & Kuhl, P. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Dev, 75*, 1067-1084.

Tse, C., Balota, D., Yap, M., et al (2010). Effects of healthy aging and early stage dementia of the Alzheimer's

type on components of response time distributions in three attention tasks. *Neuropsychology, 24*, 300-315.

Tye-Murray, N. & Geers, A. (2001). *Children's audio-visual enhancement test*. CID, St. Louis, MO.

Vickers, J. (2007). *Perception, cognition, and decision training: The quiet eye in action* (pp. 47 – 64). Champaign, IL: Human Kinetics.

Weissman, D., Roberts, K., Visscher, K., et al (2006). The neural bases of momentary lapses in attention. *Nat Neurosci*, 9, 971-978.

Whyte, J. (1992). Attention and arousal: Basic science aspects. Arch Phys Med Rehabil 73, 940-949

Wild, C., Yusuf, A., Wilson, D., et al (2012). Effortful listening: The processing of degraded speech depends critically on attention. *J Neurosci, 32*, 14010-14021.

Wingfield, A., Tun, P., & McCoy, S. (2005). Hearing loss in older adulthood: What it is and how it interacts with cognitive performance. *Curr Dir Psychol Sci, 14*, 144-148.

Woods, D., Wyma, J., Yund, E., et al (2015). Factors influencing the latency of simple reaction time. *Front Hum Neurosci 9*, 131. doi: 10.3389/fnhum.2015.00131

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. New York: Holt.

Worster, E., Pimperton, H., Ralph-Lewis, A., et al (2018). Eye movements during visual speech perception in deaf and hearing children. *Lang Learn, 68*(S1), 159-179.

Zhou, B., & Krott, A. (2016). Bilingualism enhances attentional control in non-verbal conflict tasks – evidence from ex-Gaussian analyses. *Biling: Lang Cogn, https://doi.org/10.1017/S1366728916000869*

Figure Legends

Figure 1. Mean response times in A, V, and AV modes for static and dynamic faces in CNH vs. CHL. Error bars are ± 1 standard error of mean.

Figure 2. Mean response times for *V vs. A modes* in CHL and CNH for static and dynamic faces at 1st (detection) and 3rd (attention) quartiles. "All" represents mean response times collapsed across Mode and Facial Condition. Error bars are ± 1 standard error of mean.

Figure 3. Mean response times for *AV vs. A modes* in CHL and CNH for static and dynamic faces at 1st (detection) and 3rd (attention) quartiles. "All" represents mean response times collapsed across Mode and Facial Condition. Error bars are ± 1 standard error of mean.

Figure 4. Difference (ms) between mean response times in Poorer minus Better HL subgroups of CHL: A, V, and AV modes at 1st and 3rd quartiles for static and dynamic facial conditions. Error bars are 95% CIs for differences between means. If 95% CI contains zero, performance does not differ significantly in subgroups.

Figure Legends: Appendix

Figure 1App. Cumulative distribution functions (CDFs) for A, AV, and V modes in static (1App_a) and dynamic (1App_b) facial conditions for CNH vs. CHL.

Figure 1

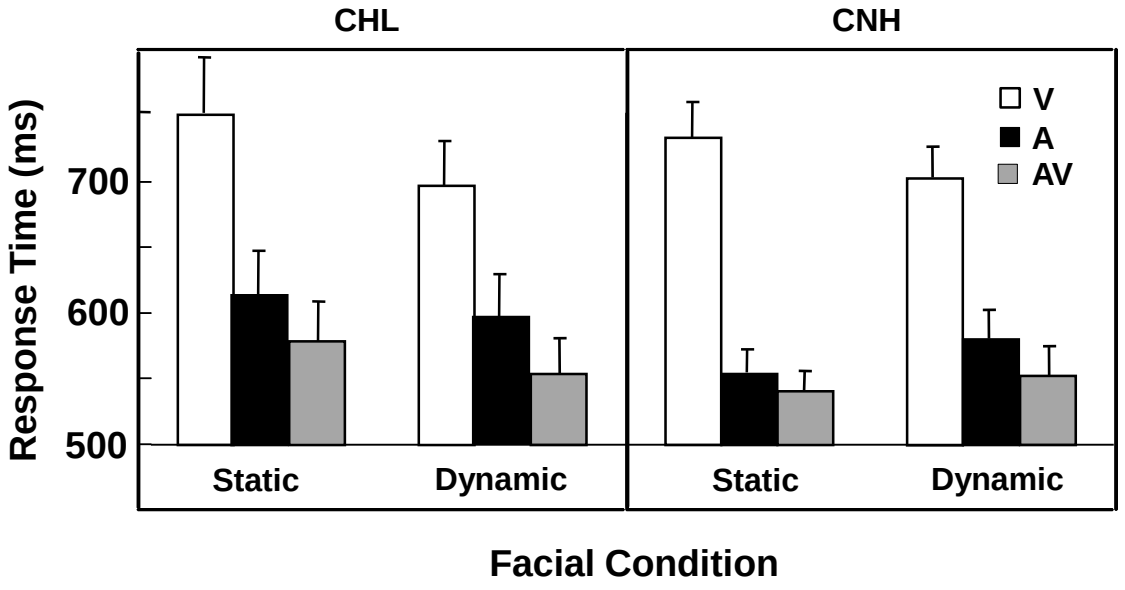

Figure 1. Mean response times in A, V, and AV modes for static and dynamic faces in CNH vs. CHL. Error bars are ± 1 standard error of mean.

Figure 2

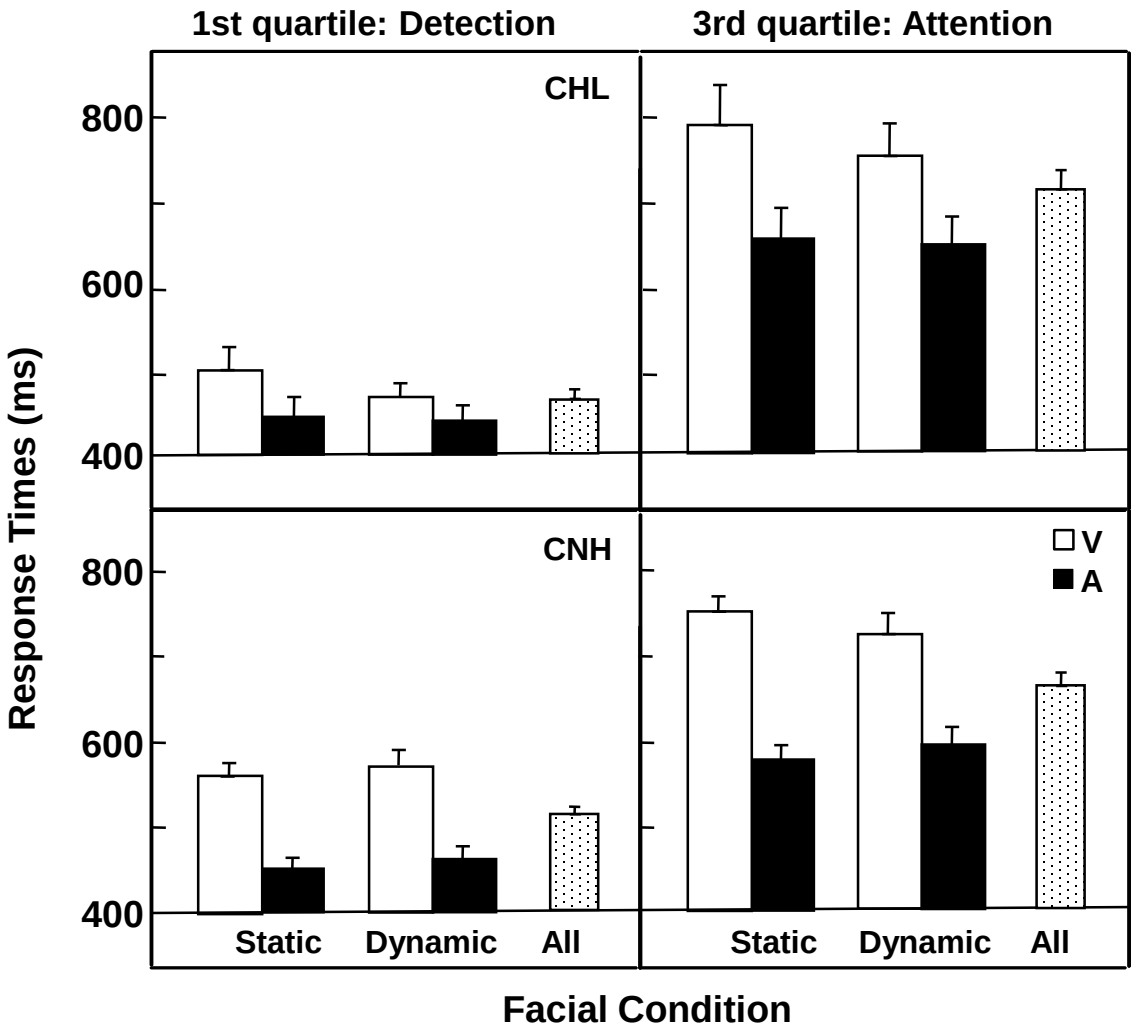

Figure 2. Mean response times for *V vs. A modes* in CHL and CNH for static and dynamic faces at 1st (detection) and 3rd (attention) quartiles. "All" represents mean response times collapsed across Mode and Facial Condition. Error bars are ± 1 standard error of mean.

Figure 3

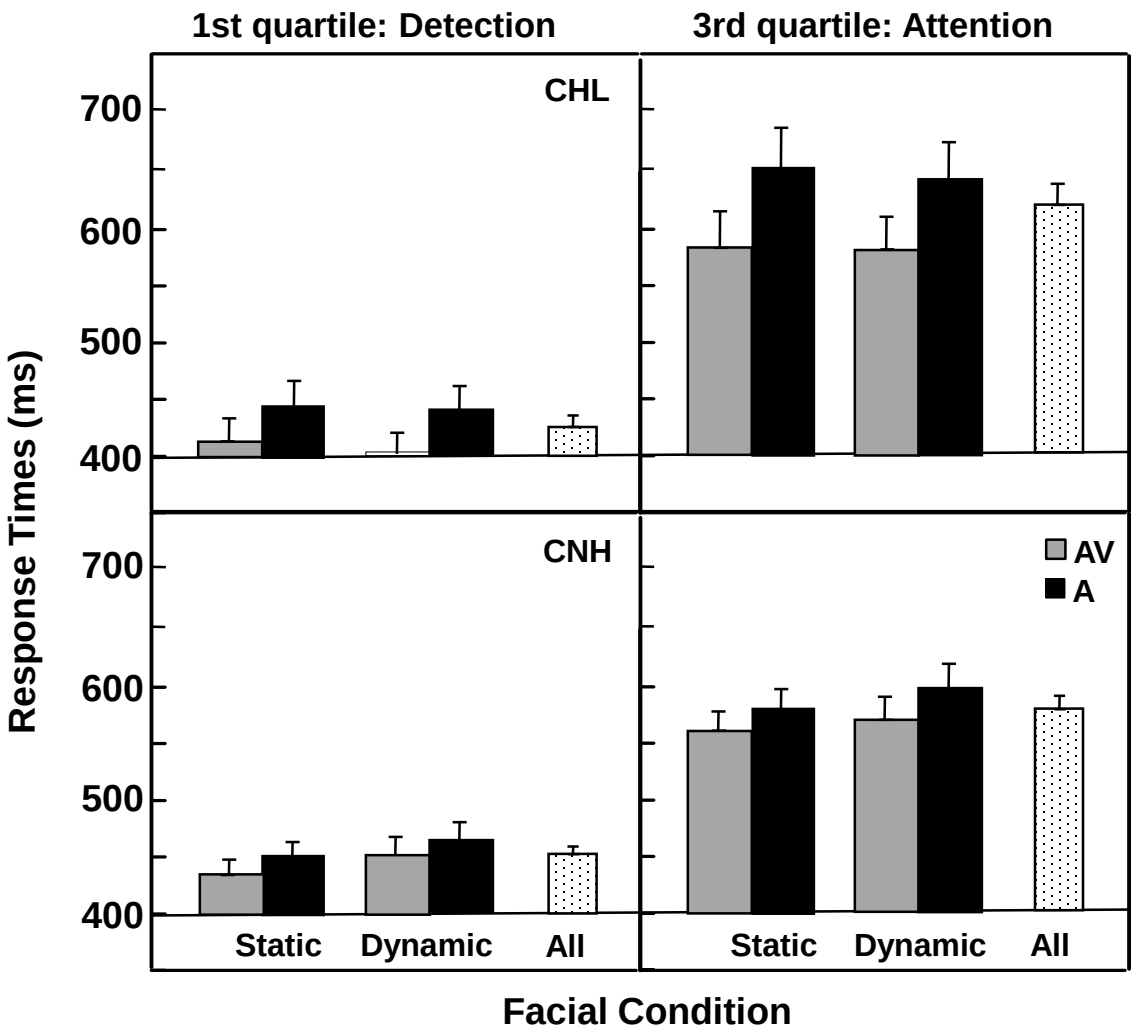

Figure 3. Mean response times for *AV vs. A modes* in CHL and CNH for static and dynamic faces at 1st (detection) and 3rd (attention) quartiles. "All" represents mean response times collapsed across Mode and Facial Condition. Error bars are ± 1 standard error of mean.

Figure 4

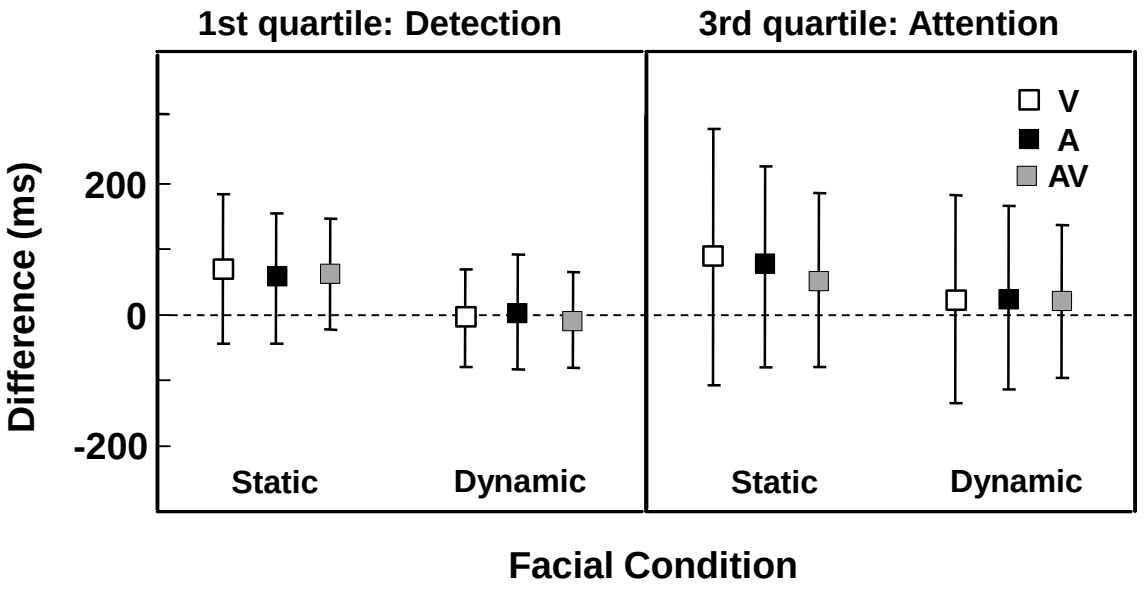

Figure 4. Difference (ms) between mean response times in Poorer minus Better HL subgroups of CHL: A, V, and AV modes at 1st and 3rd quartiles for static and dynamic facial conditions. Error bars are 95% CIs for differences between means. If 95% CI contains zero, performance does not differ significantly in subgroups.

Table All

Table 1. Average (standard deviation in parentheses) performance on a set of verbal and nonverbal measures in the CHL vs. CNH.

Measures	***Groups***	
	CHL N = 60	***CNH*** N = 60
Verbal Skills		
Vocabulary (standard score)		
Receptive*	94.67 (16.37)	122.08 (9.93)
Expressive*	93.92 (15.48)	121.90 (11.46)
Articulation Proficiency (# errors)	4.67 (7.86)	0.40 (1.72)
Nonverbal Skills		
Visual Perception (standard score)*	100.75 (15.95)	115.48 (12.86)
Word Recognition (%)		
Auditory	87.92 (10.78)	99.53 (1.30)
Audiovisual	94.83 (10.62)	---
Lipreading Onsets	67.92 (22.33)	62.90 (20.05)

Note: * Indicates performance in CNH vs CHL differed significantly (adjusted $p < .05$). Tests included in the statistical analyses were vocabulary, visual perception, and lipreading (see text).

---Audiovisual mode for word identification was not administered in CNH due to ceiling performance in auditory mode. We estimated: Vocabulary skills with Peabody Picture Vocabulary Test-III (Dunn & Dunn 2007) and Expressive One-Word Picture Vocabulary Test (Brownell 2000); Articulation proficiency with Goldman Fristoe Test of Articulation (Goldman & Fristoe 2000); Visual perception with Beery-Buktenica Developmental Test of Visual Perception [Beery & Beery 2004]; Spoken word recognition at 70 dB SPL with Word Intelligibility by Picture Identification Test (auditory mode, Ross & Lerman 1971) and Children's Audiovisual Enhancement Test (CAVET, auditory and audiovisual modes, Tye-Murray & Geers 2001); and Lipreading word-onsets with CAVET (visual mode with visemes counted as correct).

Table 2. *Results of mixed-design ANOVA with one between-participant factor (Group: CNH, CHL) and two within-participant factors (Mode: V, A, AV; Facial Condition: static, dynamic).*

Factors	***F value***	***p value***	***partial η^2***
Facial Condition	***362.27***	***< .0001***	***.756***
Mode	***84.67***	***< .0001***	***.420***
Mode x Condition	***409.03***	***< .0001***	***.778***
Group	0.10	ns	.000
Condition x Group	1.68	ns	.014
Mode x Group	3.75	ns	.030
Mode x Condition x Group	0.37	ns	.007

Note: Dependent variable: rank transformed response times. Significant results are bolded. ns = not significant.

Table 3. Results of paired t tests: Were responses faster to AV than A input?

Group ***Condition***	***AV***	***A***	***t value***	***p value***
CNH				
Static	539	551	2.14	ns
Dynamic	***550***	***577***	***4.56***	***<.0001***
CHL				
Static	***575***	***609***	***3.63***	***.004***
Dynamic	***552***	***591***	***3.95***	***.0007***

Note: Significant results are bolded. The *p* values were tested with the Bonferroni correction for multiple comparisons.

Table 4. *Results of mixed-design ANOVA with one between-participant factor (Group: CNH, CHL) and three within-participant factors (Quartile: 1st, 3rd; Mode: V, A; Facial Condition: static, dynamic).*

Factors	***F value***	***p value***	***partial η^2***
Quartile	***704.90***	***< .0001***	***.857***
Mode	***301.03***	***< .0001***	***.718***
Quartile x Group	***23.98***	***< .0001***	***.169***
Mode x Group	***45.24***	***< .0001***	***.277***
Group	2.12	ns	.018
Facial Condition	0.01	ns	.000
Facial Condition x Group	0.09	ns	.001
Quartile x Mode	0.30	ns	.003
Quartile x Mode x Group	5.58	ns	.045
Quartile x Facial Condition	0.14	ns	.001
Quartile x Facial Condition x Group	0.94	ns	.008
Mode x Facial Condition	1.13	ns	.009
Mode x Facial Condition x Group	0.22	ns	.002
Quartile x Mode x Facial Condition	0.14	ns	.001
Quartile x Mode x Facial Condition x Group	1.39	ns	.011

Note: Dependent variable: rank transformed response times.

Significant results are bolded. ns = not significant.

Table 5. *Results of mixed-design ANOVA with one between-participant factor(Group: CNH, CHL) and three within-participant factors (Quartile: 1st, 3rd; Mode: AV, A; Facial Condition: static, dynamic).*

Factors	***F value***	***p value***	***partial η^2***
Quartile	***751.39***	***< .0001***	***.864***
Mode	***84.03***	***< .0001***	***.416***
Quartile x Group	***14.21***	***.0003***	***.107***
Mode x Group	***13.81***	***.0003***	***.105***
Group	0.97	ns	.008
Facial Condition	0.40	ns	.003
Facial Condition x Group	0.15	ns	.001
Quartile x Mode	2.61	ns	.022
Quartile x Mode x Group	1.14	ns	.010
Quartile x Facial Condition	0.10	ns	.001
Quartile x Facial Condition x Group	1.61	ns	.013
Mode x Facial Condition	0.21	ns	.002
Mode x Facial Condition x Group	0.51	ns	.004
Quartile x Mode x Facial Condition	0.34	ns	.003
Quartile x Mode x Facial Condition x Group	3.83	ns	.031

Note: Dependent variable: rank transformed response times.

Significant results are bolded. ns = not significant.

Table 6. Multiple correlation coefficient and omnibus F for all variables considered simultaneously followed by the part correlation coefficients and partial F statistics evaluating the variation in performance uniquely accounted for by age, vocabulary, visual perception, auditory word recognition, or degree of hearing loss on better ear (after removing the influence of the other variables).

Variables	**1st Quartile: Detection**			**3rd Quartile: Attention**		
				V – A		
	Multiple R	***Omnibus F***	***p***	**Multiple *R***	***Omnibus F***	***p***
ALL	***.509***	***3.70***	***.006***	***.511***	***3.82***	***.005***
	Part r	***Partial F***	***p***	**Part *r***	***Partial F***	***p***
Age	.045	0.13	.721	***.316***	***7.31***	***.009***
Vocabulary	.045	0.15	.698	.114	0.93	.340
Visual Perception	.161	1.88	.176	.148	1.65	.205
Word Recognition	***.367***	***9.68***	***.003***	.167	2.04	.159
Degree of Loss	.197	2.75	.103	.084	0.50	.481
				AV – A		
	Multiple *R*	***Omnibus F***	***p***	**Multiple *R***	***Omnibus F***	***p***
ALL	.288	0.98	.440	***.513***	***3.71***	***.006***
	Part *r*	***Partial F***	***p***	**Part *r***	***Partial F***	***p***
Age	.242	3.47	.068	***.257***	***4.67***	***.035***
Vocabulary	.084	0.40	.532	***.253***	***4.49***	***.039***
Visual Perception	.000	0.01	.935	.170	2.04	.160
Word Recognition	.055	0.16	.694	.237	0.02	.893
Degree of Loss	.071	0.31	.578	.000	0.02	.882

Note: Data were collapsed across static and dynamic faces; dependent variable was difference in response times (ms). Significant results are bolded. Intercorrelations among set of standardized variables were: 1) Age vs. vocabulary (.070), visual perception (–.129) , word recognition (.457), and degree of loss (–.053), 2) Vocabulary vs. visual perception (.365), word recognition (.352), and degree of loss (–.094), 3) Visual Perception vs. word recognition (.193), and degree of loss (.163), and 4) word recognition vs. degree of loss (–.289).

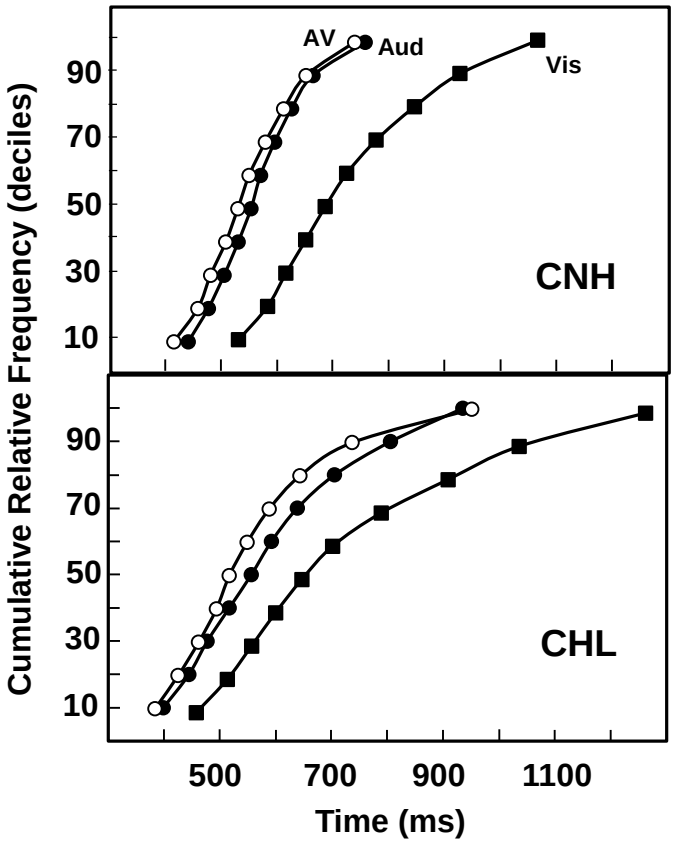

Figure Legends: Appendix

Figure 1App_a. The cumulative distribution functions (CDFs) for the A, AV, and V modes in the *static facial condition* for CNH vs. CHL.

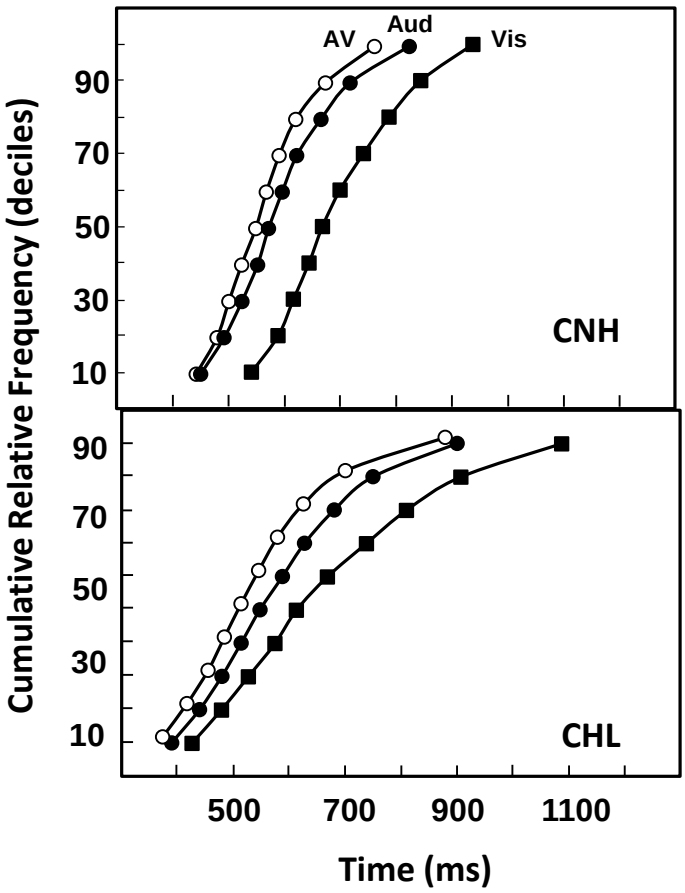

Figure Legends: Appendix

Figure 1App_b. The cumulative distribution functions (CDFs) for the A, AV, and V modes in the *dynamic facial condition* for CNH vs. CHL.