Running Head: ParSE_Cameron et al.

The Parsing Syllable Envelopes (ParSE) test for assessment of amplitude modulation discrimination skills in children: Development, normative data and test-retest reliability studies

Sharon Cameron*
Nicky Chong-White *
Kiri Mealings*
Tim Beechey *
Harvey Dillon *
Taegan Young *

* National Acoustic Laboratories, Sydney, Australia

Parts of the research described in this paper were presented at the Audiology Australia National Conference, Melbourne, Australia, May 2016.

Acknowledgments. The authors would like to thank Mark Seeto for statistical advice. Thanks are also due to the NSW Department of Education and the primary school who took part in this research. The participation of the children and their families are also appreciated. This research is funded by the Australian Government through the Department of Health.

Corresponding Author:

Sharon Cameron, PhD
Senior Research Scientist
National Acoustic Laboratories
Australian Hearing Hub
16 University Avenue
Macquarie University NSW 2109
Australia
Phone: +61 2 9412 6851
Fax: +61 2 9412 6769
e-mail: Sharon.Cameron@nal.gov.au

Abstract

Background: Intensity peaks and valleys in the acoustic signal are salient cues to syllable structure, which is accepted to be a crucial early step in phonological processing. As such, the ability to detect low-rate (envelope) modulations in signal amplitude is essential in order to parse an incoming speech signal into smaller phonological units.

Purpose: The Parsing Syllable Envelopes test (ParSE) was developed to quantify the ability of children to recognize syllable boundaries using an amplitude modulation detection paradigm. The envelope of a 750 ms steady-state /a/ vowel is modulated into two or three pseudo-syllables using notches with modulation depths varying between 0-100% along an 11-step continuum. In an adaptive three-alternative forced-choice procedure the participant identified whether 1, 2 or 3 pseudo-syllables were heard.

Research Design: Development of the ParSE stimuli and test protocols, and collection of normative and test-retest reliability data.

Study Sample: Eleven adults (23, 10 (yr, mo) to 50, 9, mean 32, 10) and 134 typically-developing, primary-school children (6, 0 (yr, mo) to 12, 4, mean 9, 3). There were 73 males and 72 females.

Data Collection and Analysis: Data were collected using a touch screen computer. Psychometric functions (PF) were automatically fit to individual data by the ParSE software. Performance was related to the modulation depth at which syllables can be detected with 88% accuracy (referred to as the Upper Boundary of the Uncertainty Region, or UBUR). A shallower PF slope reflected a greater level of uncertainty. Age effects were determined based on raw scores. Z-scores were calculated to account for the effect of age on performance. Outliers, and individual data for which the confidence interval of the UBUR exceeded a maximum allowable value, were removed. Non-parametric tests were used as the data were skewed toward negative performance.

Results: Across participants, the performance criterion (UBUR) was met with a median modulation depth of 42%. The effect of age on the UBUR was significant ($p < 0.00001$). The UBUR ranged from 50% modulation depth for six year olds to 25% for adults. Children aged six to ten had significantly higher uncertainty region boundaries than adults. A skewed distribution towards negative performance occurred ($p = 0.00007$). There was no significant difference in performance on the ParSE between males and females ($p = 0.60$). Test-retest z-scores were strongly correlated ($r = 0.68$, $p < 0.0000001$).

Conclusions: The ParSE normative data shows that the ability to identify syllable boundaries based on changes in amplitude modulation improves with age, and that some children in the general population have performance much worse than their age peers. The test is suitable for use in planned studies in a clinical population.

Key Words: Amplitude modulation, central auditory processing disorder; dyslexia, speech envelope.

Abbreviations: 3AFC = three-alternative forced-choice; ANOVA = Analysis of variance; CAPD = Central Auditory Processing Disorder; CI UBUR = Confidence Interval of the upper boundary of the uncertainty region; GUI = graphical user interface; MCL = most comfortable listening level; ms = milliseconds; ParSE = Parsing Syllable Envelopes Test; PIT = Phoneme Identification Test; RMS = root mean square; TMTF: Temporal modulation transfer function; UBUR = upper boundary of the uncertainty region.

INTRODUCTION

Accurate speech perception relies on efficient processing of temporal cues in the speech signal (Specht, 2014). There are two types of temporal cues present in a speech signal: the slowly changing temporal amplitude envelope of speech which facilitate accurate perception of syllable boundaries and the fine structure formant changes which are needed for accurate phoneme discrimination (Goswami, 2011; Goswami et al., 2011; Rosen, 1992). The goal of our research was to develop two tests that assess children's processing of each of these temporal cues. The current paper describes the development of the Parsing Syllable Envelopes test (ParSE) which assesses children's detection of amplitude modulations. The description of the Phoneme Identification Test (PIT) assessing children's categorical perception of formant frequencies is described in Cameron et al. (submitted).

Identifying smaller units from a continuous acoustic stream is essential in the process of parsing, or structuring, speech. The incoming acoustic signal contains peaks and valleys of intensity which define syllabic boundaries. While speech recognitions involves both bottom up and top down processes, syllable envelopes provide one of the most prominent acoustic landmarks in the continuous speech signal. They are therefore likely to be important for lower level segmental processing (Stevens, 2002).

The amplitude changes related to the slow modulations of the incoming acoustic signal are one of the primary cues that young children first attend to when they hear speech (Nittrouer, 2006). Infants as young as one month old are better able to discriminate syllable-like stimuli than non-syllable like stimuli (Bertoncini and Mehler, 1981). As infants gain more experience with their native language, they start to discover the other acoustic properties of the phonetic identity for their language (Nittrouer, 2006). As long as temporal cues are available within each of several frequency bands, speech can be perceived with 90% accuracy by adults even when there is no fine spectral information available within each of

these frequency bands (Shannon et al., 1995). Therefore, having the ability to process temporal cues is vital for a person to be able to perceive speech accurately (Poelmans et al., 2011). The amplitude envelope can be analyzed in terms of its constituent temporal modulation frequencies (Goswami, 2003). The modulation spectrum of speech in quiet and with low reverberation has a peak at around 4-5 Hz which reflects the rate at which syllables occur (Houtgast and Steeneken, 1985; Joris et al., 2004). When the speech envelope is low-pass filtered below 4 Hz speech becomes unintelligible due to the loss of syllable boundaries (Drullman et al., 1994).

In principle, detection of syllable boundaries should be affected by clarity of articulation, the listening environment, and the psychoacoustic abilities of the listener. Clearly articulated speech is inherently more understandable (Picheny et al., 1985). It has longer gaps between words and more intense, fully released obstruent sounds (Picheny et al., 1986), all of which should make syllable boundaries clearer. Noise and reverberation each inherently mask lower level sounds more than higher level sounds, which must make the detection of gaps, and hence syllable boundaries more difficult. Because the detection of syllable boundaries is essentially one of gap, modulation, and/or change detection (albeit with complex, changing signals), boundary detection should also depend on any relevant individual listener-related psychoacoustic abilities. These might include rate of decay of forward and backward masking, spectral resolution (because adjacent syllables will in general have different frequency spectra), the smallest change in intensity that the listener can detect for sustained signals, and the time constant with which the listener integrates intensity. For a given talker and listening environment, differences in any of these psychoacoustic characteristics could cause individual differences in the ability to segment syllables, and hence cause individual differences in speech intelligibility (Gordon-Salant and Fitzgibbons, 1993). Decreased ability to detect amplitude modulations (as assessed with the ParSE) may therefore indicate that a

child is more susceptible to the detrimental effects of noise, reverberation or rapidly produced speech.

Dimitrijevic et al. (2016) examined envelope following responses to monaural amplitude-modulated broadband noise carriers, with varying modulation depths. Participants were younger adults and older adults with varying pure-tone average hearing levels (19 dB HL versus 35 dB HL). Significant moderate correlations were found between electrophysiological and behavioral amplitude modulation detection thresholds across participants. Older adults had slightly higher amplitude modulation detection thresholds than younger adults, but the difference was not significant, and thresholds did not correlate with age. The authors concluded that the behavioral-physiological amplitude modulation depth threshold relationship was likely too weak to be clinically useful in the population assessed, who did not suffer from apparent temporal processing deficits.

Unfortunately, however, the ability to process slow-rate temporal auditory cues is impaired in some populations, such as those who have phonological dyslexia (Goswami et al., 2002; Hämäläinen et al., 2009). Children with phonological dyslexia have difficulty reading regular and/or non-words (e.g. “cat” or “gop”) which cannot be explained by low intelligence or neurological damage (Goswami et al., 2002; McArthur et al., 2013). An often co-occurring disorder with dyslexia is a central auditory processing disorder (CAPD) (King et al., 2003). CAPD refers to a variety of disorders characterized by difficulties in the central nervous system processing auditory information (American Speech-Language-Hearing Association, 2005). Children with suspected CAPD are highly heterogeneous in nature. Difficulty understanding speech when background noise is present is a commonly reported symptom of CAPD (Cameron and Dillon, 2014; Jerger and Musiek, 2000).

Despite 60 years of research, CAPD is still poorly understood and the definition, diagnosis, and treatment of the disorder is highly controversial (Vermiglio, 2014). There is a real need for new tests to be developed that can accurately diagnose additional specific

deficits causing individual children's listening problems (Dillon et al., 2012; Vermiglio, 2016). One potential deficit that Vermiglio (2016) suggests targeting is a test for a temporal resolution disorder. Temporal resolution can refer to the ability to detect changes in the duration of auditory stimuli, the presence of brief gaps in a stimulus, the masking of one sound by another that precedes or follows it, or the detection of rapid variations in intensity. While there are a number of paradigms used to characterize temporal resolution, one of the most widely adopted methods is the gap detection task (Buss et al., 2014). Gap detection tests in common use clinically include the Gaps-In-Noise test (GIN; Musiek et al., 2005) and the Random Gap Detection test (RGDT; Keith, 2000). Such tasks measure an individual's ability to detect a silent gap between two stimuli in milliseconds (ms). The stimuli bounding the silent gap can be spectrally identical narrow band markers (within-channel gap detection), identical broadband acoustic markers (across-channel gap detection), or acoustic markers that differ in frequency, ear stimulated, or location in free-field space (between-channel gap detection). Decreased gap detection thresholds are found for across-channel tasks, as the auditory system can integrate spectral information across very wide frequency ranges in order to detect the gap (Phillips, 1999; Phillips and Hall, 2000).

The acoustic markers that bound the gap in traditional gap detection tasks are identical in spectrum and have very short rise and fall times. The temporal task is therefore simply the detection of discontinuity in the activity aroused in the peripheral auditory neurons and the perceptual channel supported by that representation (Phillips and Hall, 2000). Bellis (2003) questions the clinical utility and ecological validity of such traditional gap detection techniques.

The detection of a temporal gap requires the listener to monitor stimulus intensity over time (Buss et al., 2014). The goal of this paper was to develop the Parsing Syllable Envelopes test (ParSE) to assess children's ability to perceive temporal envelope cues using a controlled, but highly realistic stimulus. In contrast to traditional gap detection tasks, the ParSE

investigates children's ability to recognize syllable boundaries using an amplitude modulation detection paradigm. Perceiving where syllables start and finish based on amplitude modulations at syllable boundaries is needed for developing accurate speech perception and reading skills (Goswami, 2011; Poelmans et al., 2011; Stevens, 2002). Consequently, we hypothesize that an underlying deficit in temporal resolution processing may be the cause of some children's listening and reading difficulties. This paper describes the development of the ParSE test and the analysis of normative test and retest data. It was hypothesized that children's ability to identify syllable boundaries based on changes in amplitude modulation would be poorer than that of adults, but that it would improve with age.

METHOD

Approval for the study was granted from the Australian Hearing Human Research Ethics Committee and the New South Wales Department of Education.

Participants

A total of 158 participants were initially assessed. There were 12 adults (23, 10 (yr, mo) to 50, 9, mean 32, 4) of which nine were female and three were male. All had normal hearing defined as equal to, or better than, 20 dB HL at all octave frequencies from 250 Hz to 8000 Hz measured bilaterally using an Interacoustics AC40 audiometer (Middelfart, Denmark) with Telephonics TDH 39P audiometric headphones (Huntington, NY) in H7A Peltor cups (3M, St. Paul, MN). The child participants were recruited from a Sydney primary school. Children whose parents reported they had an attention, language, or learning problem in the study consent form were excluded from participating. Children's hearing was tested on the day and only those who passed the pure tone audiometric screening test participated in the study. Audiometric testing was as for the adult participants, with the exception that an

Interacoustics Audio Traveller A222 portable audiometer was used. Data were collected from 146 children. Eight children were excluded post-testing on the ParSE due to inconsistent performance, as documented in the following section (Exclusions Based on Confidence Intervals). Further, five outliers (four children and one adult) were excluded, as documented in the section on calculation of z-scores. As such, following exclusions, data were analysed from a total of 145 participants. There were 134 children aged (6, 0 (yr, mo) to 12, 4, mean 9, 3), of which 64 were female and 70 were male, as well as 11 adults (23, 10 (yr, mo) to 50, 9, mean 32, 10) of which eight were female and three were male.

Software Development

The ParSE graphical user interface (GUI) and signal processing application were developed in MATLAB programing language (MathWorks Inc., 2014), and compiled for use on a touchscreen computer. Three screens were developed: a data capture screen for collection of client information and activation of reference tone; an operations screen for activation of practice, familiarization and test materials; and a test screen for the participant to respond to the ParSE stimuli. An image of the test screen appears as Figure 1.

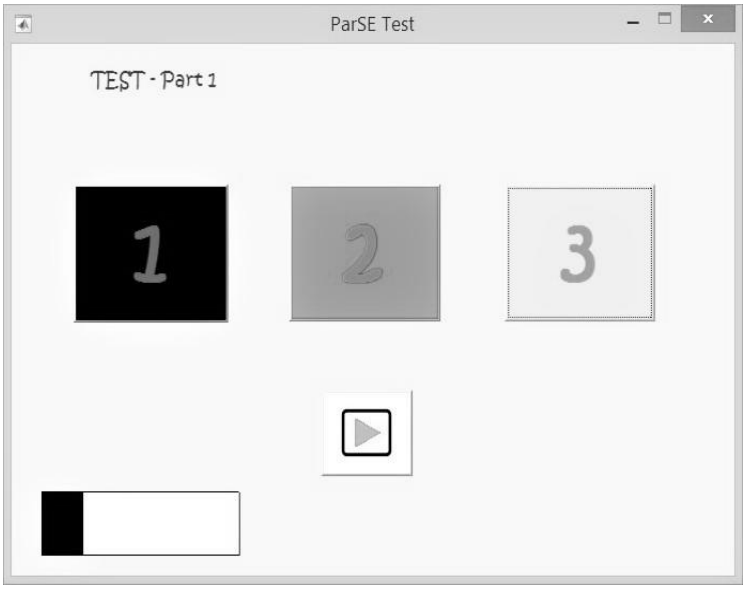

Figure 1. ParSE test screen.

Stimuli

A synthesized low central /a/ vowel, one second in duration, was generated with a sampling rate of 44100 Hz using the source-filter model provided by Praat (version 5.4.04) (Boersma and Weenink, 2014). The voicing source consisted of a pulse train with fundamental frequency (F0) of 110 Hz. A filter, representing the shape of the vocal tract, consisting of five formants with steady state frequencies of 750, 1200, 2350, 3300, and 4000 Hz was created based on frequencies used by Blomert and Mitterer (2004). Formant bandwidths of 50, 60, 110, 160, and 210 Hz were used, based on bandwidths reported by Fant (1962). The filter was applied to the voicing source to produce a low, central /a/ vowel. The resulting vowel has the temporal properties of the source with the spectral properties of the filter. The middle 750 millisecond (ms) portion was selected as the carrier wave. The carrier wave was then normalized to a root mean square (RMS) value of -20 dB (ref. full-scale square wave, 50 ms window) in Adobe Audition CS6. The onset and offset of the carrier was then ramped using MATLAB (MathWorks Inc., 2014), to zero amplitude using a 50 ms linear ramp function (see Figure 2).

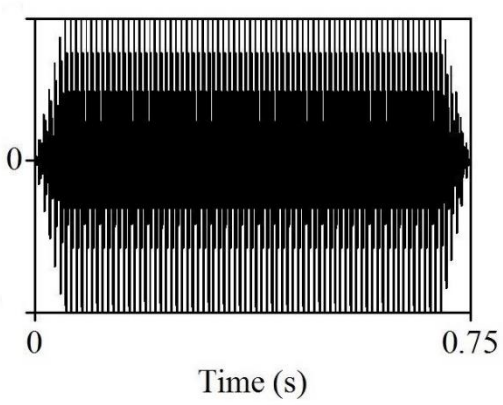

Figure 2. The 750 ms carrier /a/ vowel with 50 ms linear ramped onset and offset. Equivalent to the 0% modulated token.

A total of 16 different stimuli sets were constructed, including eight two-syllable sets and eight three-syllable sets. Each set consisted of 11 tokens of different modulation depths. For each token in the set the modulation depth decreased by 10%, from 100% modulated to 0% modulated (0% modulation represented a one pseudo-syllable token). The width of each notch in the token was fixed at 100 ms (50 ms linear down and up ramps), hence the slope of the linear ramp decreased as the modulation depth decreased. The fifty millisecond rise/fall times were selected to be broadly representative of syllable envelopes in natural speech (e.g. Howell and Rosen (1983) found average rise times of 33 ms for affricates and 76 ms for fricatives). Given the broadband nature of the carrier, amplitude modulation did not introduce spectral cues.

The location of each modulation was randomized within a specified range to ensure that modulation notches did not occur with a high degree of regularity and thus predictability. For the two-syllable stimuli, the duration of each syllable was always between 350 and 450 ms. For the three-syllable stimuli, the syllables were always between 225 and 300 ms. These durations are within the range of syllable durations observed in natural language (Crystal and

House, 1990; Greenberg et al., 2003). Examples of the two and three-pseudo-syllable tokens can be seen in Figure 3.

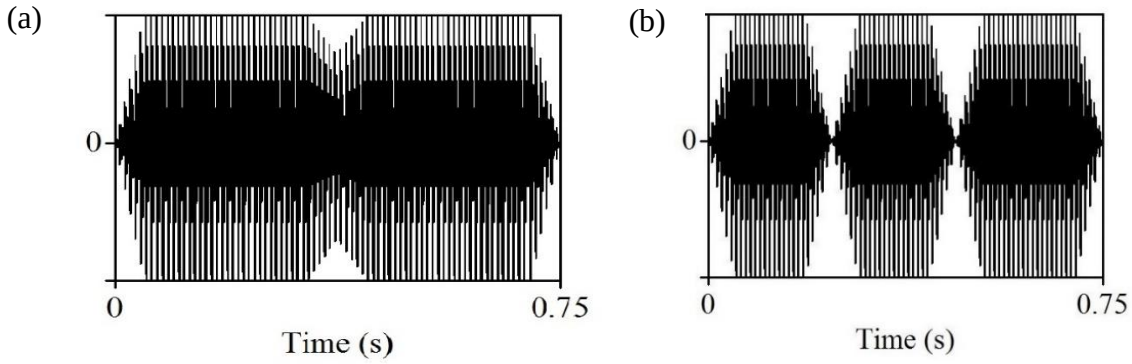

Figure 3. Examples of (a) a two pseudo-syllable token with a 30% modulation depth and (b) and three pseudo-syllable token with a 100% modulation depth.

Task

In an adaptive three-alternative, forced choice (3AFC) procedure the listener's task was to indicate whether each trial contained one, two, or three pseudo-syllables by pressing the corresponding numbered button on the touchscreen. The response buttons became active 300 ms after the offset of the stimulus to ensure that participants did not respond before hearing the complete stimulus. Participants did not receive any feedback following responses. A total of 112 tokens were presented in two blocks. The first block contained 64 randomly ordered presentations consisting of eight presentations of the odd continuum steps (0, 20, 40, 60, 80, 100%). At the end of the first block the threshold (inflection point) of the psychometric function was automatically calculated by the ParSE software. The second block contained 48 randomly ordered presentations consisting of four presentations of each end point token and eight presentations each of the five tokens clustered around the threshold of the psychometric function. These five tokens included the tokens nearest to 5% and 95% of

the psychometric function and the nearest token above and below the threshold plus the next nearest token (either above or below the threshold).

Scoring was weighted so that a score of 1 was given if the correct number of syllables (i.e. 1, 2 or 3) was identified; a score of 0 was given if modulation or the lack of modulation was not identified (i.e. a response of 1 instead of 2 or 3, or a response of 2 or 3 instead of 1); a score of 0.5 was given if modulation was detected but the incorrect number of syllables were identified (i.e. 2 instead of 3, or 3 instead of 2).

Practice and familiarization conditions were presented orally by the examiner prior to testing, as described below. The practice, familiarization and test instructions are provided in Appendix A. Including practice and familiarization, the ParSE test took approximately six minutes to complete for children aged eight to 12 years, and approximately eight minutes to complete for six and seven year olds.

Practice and Familiarization Procedure

Participants completed a brief practice task prior to the test condition. Twelve practice stimuli were presented. These comprised three repetitions each of the 100% modulated two and three pseudo-syllable endpoint tokens and six repetitions of a one pseudo-syllable (0% modulated) token. Practice stimuli were presented in random order. Following each response the participant was provided with visual feedback (the words *correct* or *incorrect* appearing on the screen).

Following the practice task participants (children only) performed a brief training task to ensure familiarity with the ambiguous tokens and to provide an understanding of how the ambiguous tokens relate to the endpoint tokens. Each continuum step was presented once in order from 0 % modulation to 100% modulation for the two syllable tokens only. After each token the participant selected whether they heard one or two syllables. No feedback was given following responses.

Reporting

Psychometric functions (PF) were automatically fitted to individual data by the ParSE software using a logistic curve. A graph of the results was displayed on the computer screen following testing. Figure 4 provides an image of a typical result for a child on the ParSE (z-score 0.0).

- The y-axis corresponds to the weighted proportion of multi-syllabic responses (range 0 to 1). Thus, for a specific modulation depth, 0 signifies that no modulation had been detected, whereas 1 indicated that a participant identified the correct number of modulated syllables 100% percent of the time.
- The x-axis corresponds to the depth of modulation (range 0% to 100%).

The following performance criteria were calculated by the software and stored in a spreadsheet.

a. Threshold: The threshold is the point at which a participant detects a 2 or 3 syllable token 50% of the time (weighted for accuracy of number of syllables), and no modulation 50% of the time.

b. Upper boundary of the uncertainty region (URUR): The UBUR corresponded to the value of the psychometric function evaluated at the point at which the asymptotic (mid-point) slope of the function intersects 100% correct. At this point, the psychometric function crosses the y-axis at 0.88. Thus, the UBUR represents the corresponding point on the x-axis (modulation depth) at which syllables can be parsed with 88% accuracy (as shown by the dissection lines in Figure 4.)

c. Confidence Interval of the UBUR (CI UBUR): The CI UBUR was obtained using a non-parametric bootstrapping technique with B=400 simulations. The observed proportions correct at each stimulus step (from 0% to 100% modulation) were used directly to generate the bootstrap simulations. The UBUR is obtained from the threshold and slope of the best-fitting logistic function for each set of simulated responses and is sorted in order from highest to lowest. The 95% confidence interval of the UBUR is determined using the (B*2.5%)th and (B*97.5%)th values. The width of the confidence interval of the uncertainty region (CI UBUR) is equal to the range of stimulus percentage values from the corresponding 2.5th percentile UBUR to the 97.5th percentile UBUR. Note that although the stimuli can have values only between 0 and 100%, both the UBUR and CI UBUR can exceed 100%.

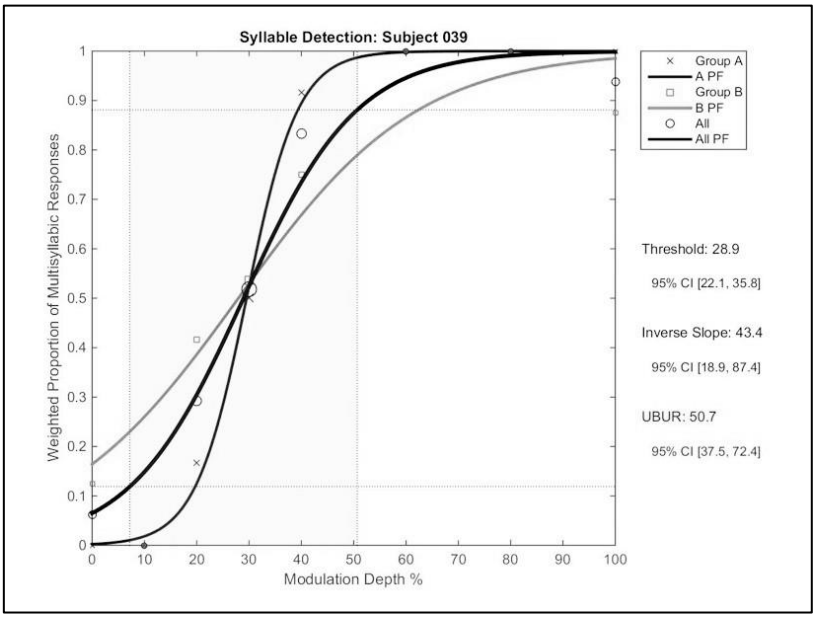

Figure 4. Image of ParSE results screen for a participant 8 yr, 2 mo of age with a z-score of 0.00. The curved lines represent the psychometric functions (PF) fitted to the data. The dark-colored central curved line is the curve fitted to all measured data, and the other two lines are

the curves fitted to odd-numbered responses of the data and even-numbered responses of the data, respectively. The right-hand dashed line shows the upper boundary of the uncertainty region (UBUR).

Procedure

Testing for the adult participants was conducted in a sound-attenuated room at the National Acoustic Laboratories using a Sony Vaio Duo 11 touchscreen computer (Sony, Japan). For the child participants, testing was completed in a quiet room at their primary school. Sound levels in the school testing rooms were measured between 45-50 dBA using a Q1362 digital sound level meter (Dick Smith Electronics, Australia). Data was collection using a Microsoft Surface Pro 3 touchscreen computer (Microsoft, China).

The pre-generated stimuli were presented binaurally, using the ParSE software, through Sennheiser HD 215 circumaural headphones. All tokens were presented at a volume control setting calibrated to 77 dB SPL during the steady-state /a/ vowel using a Brüel & Kjær Head and Torso Simulator Type 4128C (Naerum, Denmark). This level corresponded to a volume level of 40 on the Sony computer and 18 on the Surface computer. This level was selected based on the average most comfortable listening level (MCL) of the test stimuli chosen by four normal-hearing adult listeners.

Exclusions based on Confidence Intervals

As noted in the participant section, only children whose parents reported no attention or learning deficits on the study consent form were assessed with the ParSE. However, an additional inclusion criteria was implemented post-testing based on the width of the confidence interval of the upper boundary of uncertainty region (CI UBUR) recorded for the ParSE for each participant. If the CI UBUR was greater than 200, then the result was rejected

as invalid as the reliability of the fitted psychometric function was very poor. In such cases, the slope of the fitted function was extremely flat and the measured data were consistent with the responses being essentially random. Participants were also excluded if their threshold – the modulation depth at which a participant detects a 2 or 3 syllable token 50% of the time - was less than 0. Based on these criteria, and as noted in the participant section, of the 146 children assessed, data were excluded for eight children. The majority of children excluded (63%) were aged 6, 0 (yr, mo) to 6, 12.

RESULTS

Statistical analysis was performed using Statistica version 10. For each participant, performance on the ParSE was determined by the upper boundary of the uncertainty region (UBUR) score, which was converted to a z-score, as described below. Effects of age were calculated using raw scores. All other analyses, including correlations between measures, were calculated using the z-scores. Use of z-scores removes the contribution that age makes to correlations between the measures (because all test scores on average improve with age). Correlation coefficients based on z-score data will therefore be smaller than those based on raw scores. As the data were skewed towards negative performance non-parametric analyses were utilized.

Calculation of Z-Scores and Removal of Outliers

The upper boundary of the uncertainty region (UBUR) scores, collected from the 150 children and adults remaining following exclusion based on confidence intervals, was used to create equations that allow the expression of individual scores in age-corrected population standard deviation (SD) units (z-scores).

The raw UR scores were regressed against age using the exponential formula UBUR=a+b*exp (-age/c) where a, b and c are the coefficients that determine the curve. These three coefficients determine the asymptotic value applicable to adults (a); the rate of change with age (b), and the age above which the effect of age starts to diminish (c). This equation calculated the predicted UBUR score for participants as a function of age.

Residual scores were calculated as the difference between each participant's actual score and the predicted score for participants of that age. The squared residual scores were regressed against using the formula described earlier, but with new values of a, b and c. The square root of this second regression formula was used to predict the standard deviation of UBUR scores at any age. The coefficients are reported in Table 1.

Table 1. ParSE upper boundary of the uncertainty region (UBUR) (% modulation depth), regression coefficients used in the creation of z-scores.

Mean			SD		
a_1	b_1	c_1	a_2	b_2	c_2
24.3	72.6	7.76	30.7	1218	5.28

UBUR scores for each participant were then were standardized to a mean of zero and unity standard deviation and reported as z-scores, using the formula:

$z = (a+b*\exp(-Age/c) - score)/SD_{predicted}$

The UBUR z-scores were examined to determine if they deviated significantly from a normal distribution. The Shapiro-Wilk *W* value was 0.91 ($p < 0.000001$) with z-scores

ranging from 1.7 to -3.1 (mean -0.0002). Outliers (z-scores poorer than 2.5 standard deviations below the mean) were removed from the normative data. At this cut-off point it would be expected that approximately one participant would be removed were the distribution to be normal. However there were five outliers removed (four children and one adult), demonstrating a skew towards decreased precision of syllable boundary perception.

Z-scores were recalculated for the remaining 145 participants. The Shapiro-Wilk *W* value was 0.95 (p = 0.00007), with z-scores ranging from 2.3 to -3.0. A histogram is provided as Figure 5.

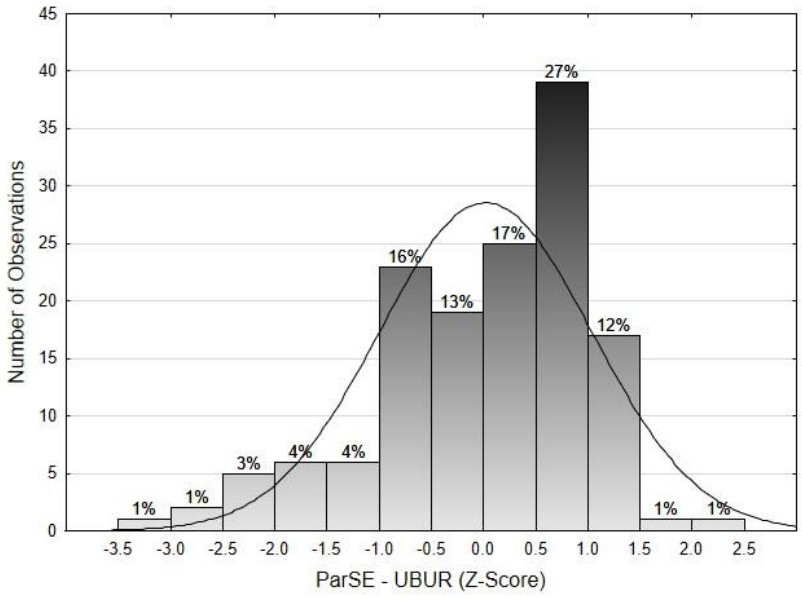

Figure 5. Histogram of upper-boundary of the uncertainty region (UBUR) z-scores for the 145 participants on the ParSE.

Gender Effects

The median upper boundary of the uncertainty region (UBUR) was 42% for males (n = 73) and 42% for females (n = 72). The Kruskal–Wallis H test (one-way ANOVA on ranks) was used to determine if threshold differed significantly between groups. Age was controlled for by comparing z-scores. There was no significant difference between males and females (H (1) = 0.27, p = 0.60).

Age Effects – Thresholds

The mean and median ParSE thresholds, as a function of age, are provided in Table 2. As there were only a small number of 12 year-olds with a maximum age of 12, 4 (yr, mo), data from 11 and 12 year-olds were combined. Across age groups, the median threshold (% modulation depth) was 22%. Due to the skewed distribution of the data, the Kruskal–Wallis H test was used to determine if threshold differed significantly between age groups. Overall, there was no significant effect of age on ParSE threshold (H (6) = 12.465, p = 0.052.

Age Effects – UBUR

The mean and median ParSE UBUR scores, as a function of age, are provided in Table 2. Children aged 11 and 12 years were combined, as noted above. Across age groups, the median UBUR (% modulation depth) was 42%. There was a trend of decreased UBUR with age. The Kruskal–Wallis H test revealed a significant effect of age on the ParSE UBUR (H (6) = 35.7, $p < 0.00001$). Children aged six to ten had significantly higher uncertainty region boundaries than adults (see Figure 6). A scatterplot of individual raw ParSE UBUR data for the 145 children and adults as a function of age is provided as Figure 7. The ± 2 standard deviation limits, calculated from the regression equations, is delineated.

Table 2. Median, mean and standard deviations for ParSE threshold and upper boundary of the uncertainty region (UBUR) (% modulation depth), as a function of age.

		ParSE					
		Threshold (%)			UBUR (%)		
Age	N	Median	Mean	SD	Median	Mean	SD
Overall	145	22	23	10	42	45	17
6	19	24	24	8	50	55	19
7	18	22	23	11	45	47	17
8	19	25	30	14	54	58	20
9	22	22	22	11	41	42	13
10	25	19	22	8	39	43	15
11-12	31	20	21	9	38	40	12
Adult	11	19	17	5	25	26	8

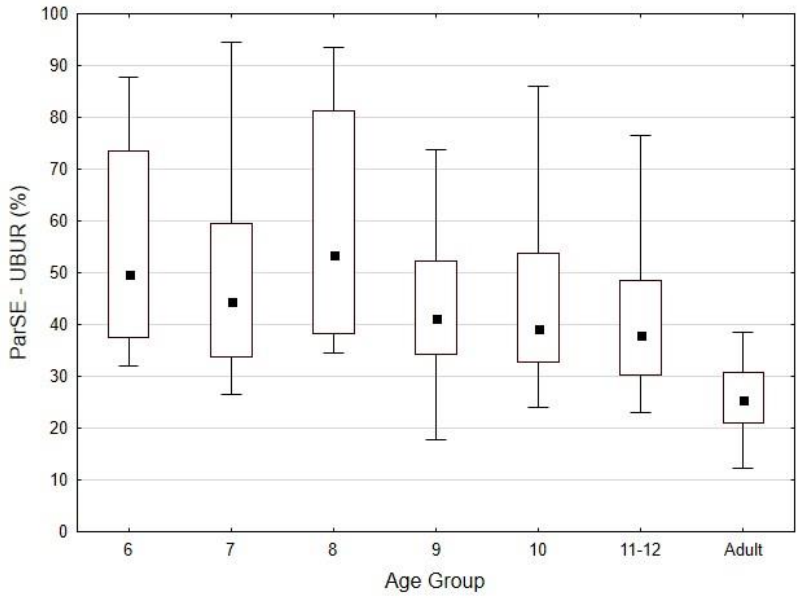

Figure 6. Box and whisker plot of upper boundary of the uncertainty region (UBUR) as a function of age for the 145 participants on the ParSE. The filled square represents the median UBUR, the open boxes represent the 25-75% boundaries, and the whiskers represent the minimum and maximum scores.

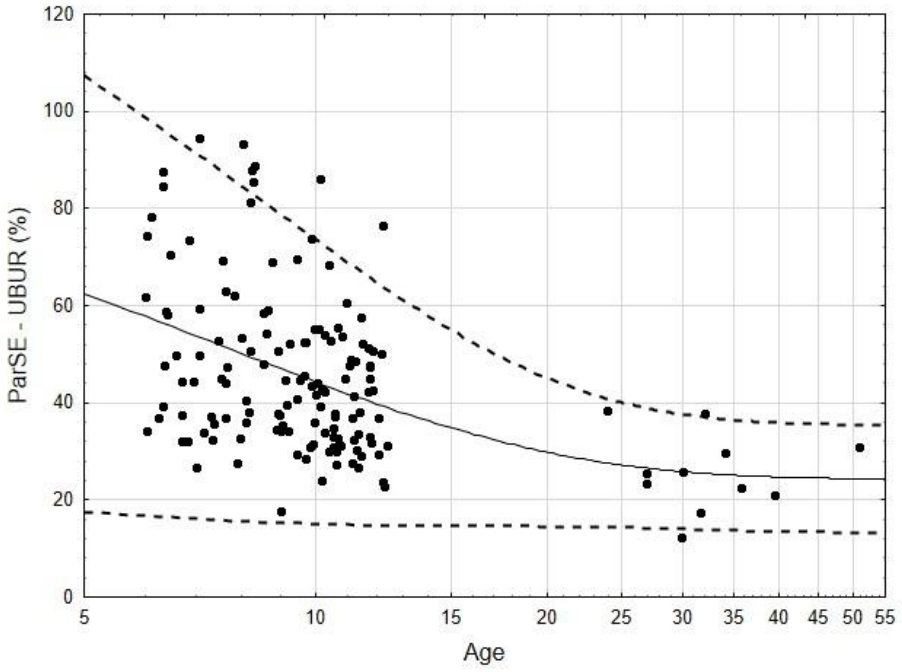

Figure 7. Scatterplot of the individual raw UBUR scores at for the 145 children and adults tested on the ParSE. The solid line indicates the mean score as a function of age, the dashed line shows the ± 2 SD limits.

Test-Retest Reliability

Test-retest reliability data was analysed for 106 children on the ParSE. Results for an additional 13 children were rejected prior to statistical analysis due to invalid confidence intervals. Participants were retested between 16 days and 44 days (mean 35 days) after their initial appointment. The median upper boundary of the uncertainty region (UR) z-scores were 0.11 SD at test and 0.56 SD at retest (mean -0.01 and 0.41 SD respectively). Repeated measures (Wilcoxon Matched Pairs) test revealed a significant difference between test and retest ($T = 1173$, $p < 0.0000001$). Spearman rank order correlations revealed a strong

correlation ($r = 0.68$, $p < 0.0000001$) between test and retest UBUR z-scores (Figure 6a). The mean difference between retest and test z-scores was 0.42 SD.

The Pearson's product moment UBUR z-score test-retest correlation was $r = 0.64$ ($p<0.00001$). Consequently the proportion of variance accounted for by measurement error in the test scores is estimated as 36% (equal to 1-r).

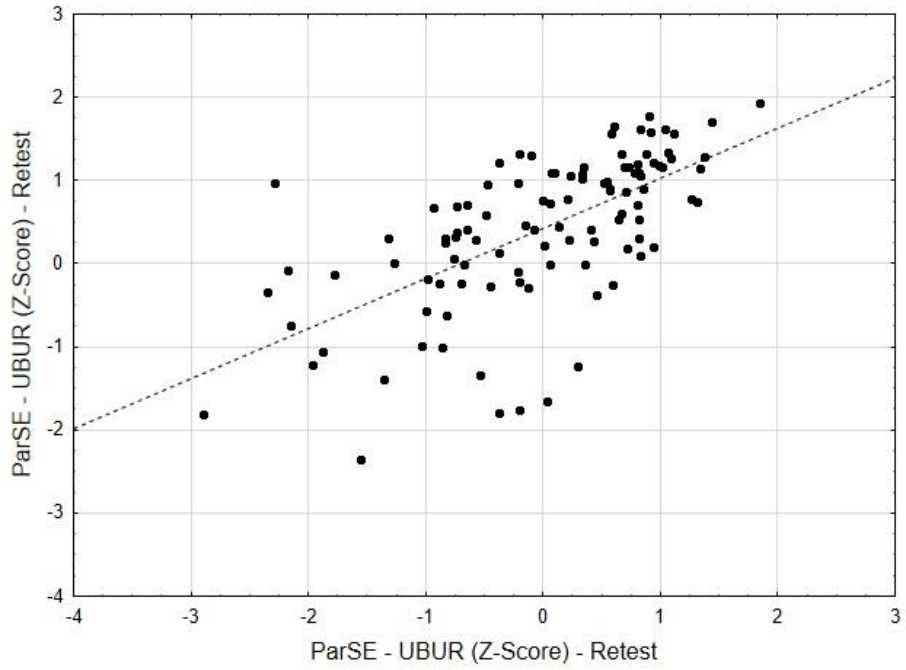

Figure 8. Scatterplot of the upper boundary of the uncertainty region (UBUR) scores at test and retest for the 106 children retested on the ParSE. The dashed line represents the least squares regression line.

DISCUSSION

The present paper documents the development of a new test of temporal resolution intended for future clinical use for children with suspected CAPD and/or reading deficits. The

Parsing Syllable Envelopes (ParSE) test uses an adaptive amplitude modulation detection paradigm to evaluate an individual's ability to analyse the low-rate peaks and valleys in the envelope of an incoming acoustic signal that provide cues to syllable boundaries. Stimuli were randomized one, two and three pseudo-syllable tokens - the multisyllabic tokens containing notch modulation depths ranging from 10 to 90 percent. Performance was determined by the individual's upper boundary of the uncertainty region (UBUR), being the modulation depth at which multi-syllabic tokens could be detected with 88% accuracy.

The threshold for syllable detection - that is, the modulation depth where an individual indicated that they had heard a two or three syllable token half the time – was 22%. There was no significant difference between age groups on the threshold for syllable detection. There was, however, a significant age effect on the UBUR, with younger children (six to ten years) needing a greater modulation depth in order to consistently detect a multi-syllabic syllable. From Table 2 and Figure 6 it can be seen that there is both a gradual variation of performance with age, and also year-to-year random fluctuations. The method we used to calculate z-scores involved regressing the test scores against age as a continuous variable. This method preserves the gradual variation while smoothing out the random fluctuations.

It is clear that the distribution of normative data test scores deviated from a normal distribution, with a skew towards below-average performance. That is, the worst performers were further below the mean performance than the better performers were above the mean. It is not possible to say whether this reflects an intrinsic skew in the range of abilities of syllable boundary perception, or is a product of the particular measure used to describe performance – in this case the amplitude modulation depth needed for the children to detect the modulation 88% of the time. While we could have performed a mathematical transformation that normalises the distribution we have chosen not to as there is no a priori reason why the ability to perceive syllable boundaries should be normally distributed. Importantly, any such transforms do not change the rank order of children's performance on

the task, so the only effect of such a transform is to change the z-score at which one considers performance is sufficiently different from the mean to represent a problem in real life perception of speech.

Repeated measures analysis revealed a small but significant average improvement on retest by 0.4 standard deviations. Correlational analysis revealed a strong relationship ($r = 0.68$) between test and retest z-scores. Test-retest correlations were used to determine the impact of random measurement error. The proportion of measurement error in the ParSE test, estimated as 36% of the variance in test scores, is considerably less than was found for the Phoneme Identification Test (PIT) (Cameron et al., submitted). Possibly there are bigger true differences in children's ability to identify syllable boundaries than differences in their ability to perceive categorically. Alternatively, perhaps the task of counting syllables maintains a more constant level of attention, or attracts a more constant criterion of the boundary between the different sounds, than the task of identifying which phoneme was heard.

Inspection of individual data prior to group analysis revealed greatest variation in performance in the youngest children. Confidence intervals were calculated on the slope parameter used to calculate the upper boundary of the uncertainty region (UBUR) and results were deemed invalid for individuals whose confidence intervals were substantially wider than those of the remaining children. Of the children excluded, the majority were in the six year age group. In clinical trials currently in progress participation has been restricted to children 7 yrs, 6 mths to 11 yrs, 6mths. If it is found during these studies that the ParSE has clinical validity in older children we may re-examine performance on 6 to 7.5 year olds using a verbal or pictorial response method whereby the child indicates to the audiologist the number of syllables perceived and the audiologist inputs the data. Younger children may be more motivated to attend to the test stimuli, and less distracted by outside influences, if an authority figure is more actively involved in the test administration.

Finally, even having accounted for any invalid results by excluding children based on UBUR confidence interval parameters, prior to analysis of the normative data, five outliers were removed from the ParSE results. Based on the sample size, the number of outliers would be predicted to be around one. Thus the number of children who exhibited decreased precision of syllable boundary detection outside the normal range was unexpected, and may be an indication of a distinct clinical population. Alternatively, they may be children who for some reason were not able to properly follow the test directions, despite provision of practice and familiarization procedures, or who otherwise did not respond in a way that reflected their true temporal resolution. It will require future research to resolve whether poor performance on the test always has adverse consequences for communication in challenging situations or whether poor performance can sometimes have no real-life consequences. The same statement can be made about many, if not all, clinical tests used to assess auditory processing disorders.

The present study forms the first steps toward developing a new and innovative test of temporal resolution ability suitable for clinical use. Studies are currently being undertaken in children who present with phonological dyslexia. Patterns of performance across a range of standardized assessment tasks and measures of cortical auditory evoked potentials will be documented. These results will be correlated with ParSE results, as well results on the PIT (Cameron et al, submitted.), which was developed for the study to investigate the ability of children to process rapid formant transitions. It would be desirable for future research to also examine the relationship between performance on this modulation-based measure of temporal resolution and performance on traditional rapid-onset/offset gape-detection tasks. It is anticipated that the results from the research presented here, as well as the clinical studies currently in progress, will contribute to the expansion of diagnostic targets for auditory processing assessment noted by (Vermiglio, 2016), and shed light on any auditory-specific contributions to other diagnostic entities such as phonological dyslexia.

REFERENCES

American Speech-Language-Hearing Association. (2005) *(Central) auditory processing disorders [Technical statement]*.

Bellis, TJ. (2003) *Assessment and management of central auditory processing disorders in the educational setting. From science to practice*. New York: Delmar Learning.

Bertoncini, J, Mehler, J. (1981) Syllables as units in infant speech perception. *Infant Behav Dev* 4(1):247–260.

Blomert, L, Mitterer, H. (2004) The fragile nature of the speech-perception deficit in dyslexia: Natural vs. synthetic speech. *Brain Lang* 89(1):21–26.

Boersma, P, Weenink, D. (2014) Praat: doing phonetics by computer (version 5.4.04).

Buss, E, Hall, JW, Porter, H, Grose, J. (2014) Gap detection in school-age children and adults: Effects of inherent envelope modulation and the availability of cues across frequency. *J Speech Lang Hear Res* 57(3):1098–1107.

Cameron, S, Dillon, H. (2014) Remediation of spatial processing issues in CAPD. In: Chermak, G, Musiek, F, eds. Handbook of Central Auditory Processing Disorder. Comprehensive Intervention (Vol. II). San Diego: Plural Publishing, 201–224.

Cameron, S, Chong-White, N, Mealings, KT, Beechey, T, Dillon, H, Young, T. The Phoneme Identification Test (PIT) for assessment of spectral and temporal discrimination skills in children: Development, Normative Data and Test-Retest Reliability Studies. *J Am Acad Audiol*.

Crystal, TH, House, AS. (1990) Articulation rate and the duration of syllables and stress groups in connected speech. *J Acoust Soc Am* 88(1):101–112.

Dillon, H, Cameron, S, Glyde, H, Wilson, W, Tomlin, D. (2012) An opinion on the assessment of people who may have an auditory processing disorder. *J Am Acad Audiol* 105(23):97–105.

Dimitrijevic, A, Alsamri, J, John, MS, Purcell, D, George, S, Zeng, F-G. (2016) Human envelope following responses to amplitude modulation: effects of aging and modulation depth. *Ear Hear* 37(5):e322–e335.

Drullman, R, Festen, JM, Plomp, R. (1994) Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am* 95:1053–1064.

Fant, G. (1962) Formant bandwidth data. Speech Transmission Laboratory Quarterly Progress and Status Report. Stockholm: Royal Institute of Technology, 1–2.

Gordon-Salant, S, Fitzgibbons, P. (1993) Temporal factors and speech recognition performance in young and elderly listeners. *J Speech Lang Hear Res* 36(6):176–1285.

Goswami, U. (2003) Why theories about developmental dyslexia require developmental designs. *Trends Cogn Sci* 7(12):534–540.

Goswami, U. (2011) A temporal sampling framework for developmental dyslexia. *Trends Cogn Sci* 15(1):3–10.

Goswami, U, Thomson, J, Richardson, U, Stainthorp, R, Hughes, D, Rosen, S, Scott, SK. (2002) Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proc Natl Acad Sci U S A* 99(16):10911–10916.

Goswami, U, Fosker, T, Huss, M, Mead, N, Szucs, DD. (2011) Rise time and formant transition duration in the discrimination of speech sounds: The Ba-Wa distinction in developmental dyslexia. *Dev Sci* 14(1):34–43.

Greenberg, S, Carvey, H, Hitchcock, L, Chang, S. (2003) Temporal properties of spontaneous speech – a syllable-centric perspective. *J Phon* 31:465–485.

Hämäläinen, J a., Leppänen, PHT, Eklund, K, Thomson, J, Richardson, U, Guttorm, TK, Witton, C, Poikkeus, a.-M, Goswami, U, Lyytinen, H. (2009) Common variance in amplitude envelope perception tasks and their impact on phoneme duration perception and reading and spelling in Finnish children with reading disabilities. *Appl Psycholinguist* 30(3):511.

Houtgast, T, Steeneken, H. (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am* 77:1069–1077.

Howell, P, Rosen, S. (1983) Production and perception of rise time in the voiceless affricate/fricative distinction. *J Acoust Soc Am* 73(3):976–984.

Jerger, J, Musiek, F. (2000) Report of the consensus conference on the diagnosis of auditory processing disorders in school-aged children. *J Am Acad Audiol* 11:467–474.

Joris, PX, Schreiner, CE, Rees, A. (2004) Neural processing of amplitude-modulated sounds. *Physiol Rev* 84:541–577.

Keith, R. (2000) *Random Gap Detection Test*. St Louis: AUDiTEC.

King, WM, Lombardino, LJ, Crandell, CC, Leonard, CM. (2003) Comorbid auditory processing disorder in developmental dyslexia. *Ear Hear* 24(5):448–456.

MathWorks Inc. (2014) MATLAB (Release 2014b). [Computer Software]. Natick, Massachusetts: The MathWorks Inc.,.

McArthur, G, Kohnen, S, Larsen, L, Jones, K, Anandakumar, T, Banales, E, Castles, A. (2013) Getting to grips with the heterogeneity of developmental dyslexia. *Cogn Neuropsychol* 30(1):1–24.

Musiek, FE, Shinn, JB, Jirsa, R, Bamiou, D-E, Baran, J a, Zaida, E. (2005) GIN (Gaps-In-Noise) test performance in subjects with confirmed central auditory nervous system involvement. *Ear Hear* 26(6):608–618.

Nittrouer, S. (2006) Children hear the forest. *J Acoust Soc Am* 120(4):1799–1902.

Phillips, DP. (1999) Auditory gap detection, perceptual channels, and temporal resolution in speech perception. *J Am Acad Audiol* 10(6):343–354.

Phillips, DP, Hall, SE. (2000) Independence of frequency channels in auditory temporal gap detection. *J Acoust Soc Am* 108(6):2957–2963.

Picheny, MA, Durlach, NI, Braida, LD. (1985) Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *J Speech Lang Hear Res* 28:96–103.

Picheny, MA, Durlach, NI, Braida, LD. (1986) Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *J Speech Lang Hear Res* 29(4):434–446.

Poelmans, H, Luts, H, Vandermosten, M, Boets, B, Ghesquière, P, Wouters, J. (2011) Reduced

sensitivity to slow-rate dynamic auditory information in children with dyslexia. *Res Dev Disabil* 32(6):2810–2819.

Rosen, S. (1992) Temporal information in speech: Acoustic, auditory, and linguistic aspects. *Philos Trans* 336(1278):367–373.

Shannon, R V., Zeng, F-G, Kamath, V, Wygonski, J, Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science* 270(5234):303–304.

Specht, K. (2014) Neuronal basis of speech comprehension. *Hear Res* 307(378):121–135.

Stevens, KN. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *J Acoust Soc Am* 111(4):1872–1891.

Vermiglio, AJ. (2014) On the clinical entity in audiology: (Central) auditory processing and speech recognition in noise disorders. *J Am Acad Audiol* 25(9):904–917.

Vermiglio, AJ. (2016) On diagnostic accuracy in audiology: Central site of lesion and central auditory processing disorder studies. *J Am Acad Audiol* 27(2):141–156.

APPENDIX A

Instructions given to participants prior to undertaking the Parsing Syllable Envelopes (ParSE) practice, familiarization and test conditions.

Practice

"When you are ready to start press the play button. You will hear a voice saying either 'ah', 'ah ah', or 'ah ah ah'. Press the button on the screen that matches how many 'ah's you hear each time. If you are not sure how many 'ah's you hear please guess."

Familiarisation

"Now you will hear the voice again but the sound will start as one syllable and change to two. So at first it will sound like 'ah' and then it will start to sound more like 'ah ah'. Press the button that matches what you hear. If you're not sure, just guess. There are no 'ah ah ahs' this time."

Test

"Now you will hear the same type of sounds but some sounds may be less clear. If you are not sure whether you heard 'ah', 'ah ah', or 'ah ah ah' please choose the sound you think it was more likely to be. Half-way through you can have a break. Try to press the button as soon as you hear the sound but it is more important to be accurate than fast".

FIGURE CAPTIONS

Figure 1. ParSE test screen.

Figure 2. The 750 ms carrier /a/ vowel with 50 ms linear ramped onset and offset. Equivalent to the 0% modulated token.

Figure 3. Examples of (a) a two pseudo-syllable token with a 30% modulation depth and (b) and three pseudo-syllable token with a 100% modulation depth.

Figure 4. Image of ParSE results screen for a participant 8 yr, 2 mo of age with a z-score of 0.00. The curved lines represent the psychometric functions (PF) fitted to the data. The dark-colored central curved line is the curve fitted to all measured data, and the other two lines are the curves fitted to odd-numbered responses of the data and even-numbered responses of the data, respectively. The right-hand dashed line shows the upper boundary of the uncertainty region (UBUR).

Figure 5. Histogram of upper-boundary of the uncertainty region (UBUR) z-scores for the 145 participants on the ParSE.

Figure 6. Box and whisker plot of upper boundary of the uncertainty region (UBUR) as a function of age for the 145 participants on the ParSE. The filled square represents the median UBUR, the open boxes represent the 25-75% boundaries, and the whiskers represent the minimum and maximum scores.

Figure 7. Scatterplot of the individual raw UBUR scores at for the 145 children and adults tested on the ParSE. The solid line indicates the mean score as a function of age, the dashed line shows the ± 2 SD limits.

Figure 8. Scatterplot of the upper boundary of the uncertainty region (UBUR) scores at test and retest for the 106 children retested on the ParSE. The dashed line represents the least squares regression line.