Analysis of phone-errors in Reading Disabled children

Jiachen Tu[1], Cynthia J. Johnson[2], Jont B Allen[3]

[1]ECE, Univ. of Illinois, 306 N Wright St, Urbana, IL, 61801, jtu9@illinois.edu

[2]Speech and Hearing, Univ. of Illinois, 901 S 6th St, Champaign, IL, 61820, cjj@illinois.edu

[3]ECE, Univ. of Illinois, 306 N Wright St, Urbana, IL, 61801, jontallen@ieee.org

Abstract

Objectives: Reading disability (RD) is a key obstacle in the development of literacy. Studies show that 15-20% of grade-school students have RD, and that this has lifelong consequences for the individual and the wider community. Based on our two experimental tasks (SCO and NSCM), the current study examines a key potential source of RD in young children (8-11 years old), namely that due to deficits in phone-level perception.

Design: The *Syllable-Confusion Oddball* (SCO) procedure is an 3-interval forced-choice (3-IFC) closed-set task, to determine which of more than 20 phones have perceptual errors. The *Nonsense Syllable Confusion Matrix* (NSCM) procedure is a 1-interval open set task, where the subject hears one of 20 consonant vowels (CV), and orally reports back what they heard. The NSCM task complements the SCO task by measuring the detailed map of phone confusions, as either a confusion count matrix or a directed graph. More than ten normal hearing children having fully-documented RD served as subjects. Their performance was compared to that of six normal-hearing and language control children. On average 1,500 trials were performed on each child, over a two-week period, for both the RC (20-40 trials per syllable) and RD groups (30-40 trials per syllable), for both tasks.

Results: The current study shows that the proportion of errors was between 3 to 5 times greater for RD listeners (30-50% error) compared with the RC listeners (10% error). The RC subjects show a greatly reduced intra-confusion variance and a similar inter-confusion patterns, allowing for the definition of an average RC normal (AN), which meaningfully characterizes the RC group. Unlike the RC, the RD subjects were highly idiosyncratic (they had large individual differences in their confusion pattern errors).

Conclusions: It was clear from these data that increasing the number of RD subjects would simply add more idiosyncratic subjects. Given the rather high RD confusions (error), it seems unlikely that patterns of RD confusions would emerge. Perhaps more important is that the individual confusions indicate a program of treatment, targeted at those sounds having the largest errors. We conclude that RD children have a significant idiosyncratic (intra-confusion) phone-level speech perception problem, captured in the confusion patterns. With the confusion matrix information, it should be possible to generate specific diagnostic feedback to improve phone recognition.

Key words: dyslexia, phonemic awareness, phone confusions, decoding, encoding

Running head: Analysis of RD phone confusions

1 Introduction

In typically developing (TD) children, speech perception happens naturally, seemingly without effort, as early at two years old. (See Table 7 for a glossary of abbreviations used in the present paper.) The ability to discriminate and identify speech sounds (phones) provides the foundation for learning to produce and comprehend spoken language and equally important, the ability to read visual letters and associate them with sounds. In contrast, learning to read by TD children requires considerable instruction and practice. Most children start to read at age 6 or 7. The earliest is 3 years old, but this is uncommon. Most disturbing, some children *never* learn to read, and as a result dropout of school at an early age. After more than a hundred years of research, we still do not fully understand why. We do know that its *not* a dysfunctional brain, or a low IQ (Torgesen, 2004; Wong, 2011).

Understanding why some children cannot learn to read is a century old mystery (Torgesen, 2004), critical for explaining the problems encountered by children with reading disabilities (RD). Here we investigate the relationship between *reading ability* and *speech perception*, and we argue that a strong parallel exists between the RD subject and the effects of early hearing impairment (HI).

For example, children born with a HI have consistent idiosyncratic consonant confusions. Today this problem has been partially mitigated with the early-placement of a cochlear implant. We shall show that nearly identical symptoms in speech perception exist for RD listeners. The reasons for this parallel are presently unknown, but are consistent with poor performance on *phonemic awareness* (PA), and therefore poor auditory phone encoding deficits (Torgesen, 2004; Tallal, 2000; Singh and Allen, 2012). When one is born HI, they fail to learn the phones, thus they have low PA. With the addition of a cochlear implant, the HI is mitigated. On the other hand, recent studies on speech envelope enhancement (EE), a speech perception improvement strategy (Van Hirtum et al., 2019), also demonstrated that students with dyslexia, a developmental disorder in learning to read, not only benefited from the EE technique, but significantly benefited from it more than TD readers. This result therefore supports a relation between speech perception abilities and reading skills.

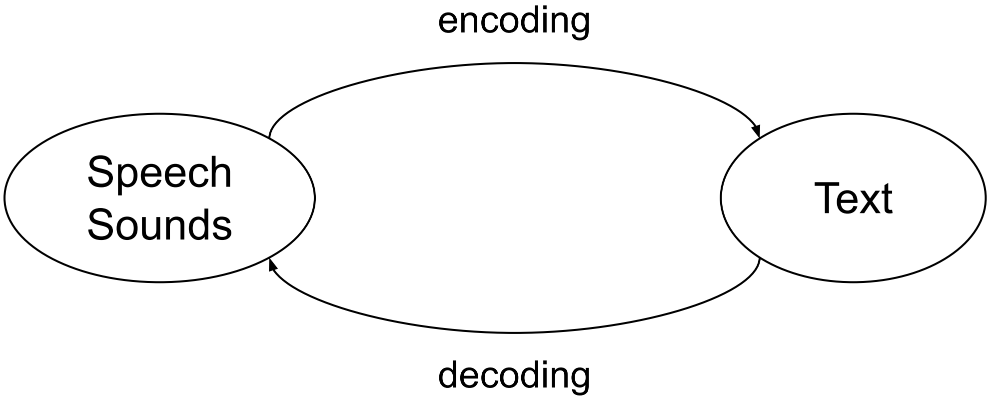


Figure 1:
relationship diagram between perceptual encoding and decoding

Reading requires *decoding*, the translation of printed words into speech sounds as shown in the Fig. 1 while *encoding* requires translation of speech sound into text. Learning to read requires *decoding*, which is the translation of the printed words to “unspoken” (perceptual) sound. Accurate decoding can be seen as fundamental to the reading process. As stated by

Hanford (2018): "The starting point for reading is sound. A child who can't decode will never become a reader." While decoding is central to current models of reading, in that it provides the child with the basic knowledge needed to map letters onto phones, and eventually directly to lexical representations (Seidenberg and McClelland, 1989), decoding is *not* the first step.

The decoding process must be contrasted with *encoding*, which involves mapping spoken speech sounds to phones, syllables, letters, words, and ultimately, meaning (i.e., information) (Allen, 2005a). This requires the construction of a brain-map of phones given speech sound stimuli.

More precisely, the *first* step toward reading is mastery of encoding, which ferments within the first two years of a TD child's life. Encoding allows the child to accurately recognize spoken speech sounds, internally representing them as phones (an some abstract representation in the auditory cortex, or more likely, Wenicke's area), but eventually phonemes.

Encoding must be mastered before decoding can begin. Decoding is a later step, and presumably is a step that must be taught in reading. Before teaching children to read, it has traditionally been *assumed* they have normal speech perception (encoding) skills. This assumed mastery of encoding appears to be the downfall in teaching reading skills. We hypothesize it is precisely the failure of this assumption which is the source of RD. In our view, this hypothesis naturally follows from Fletcher's 1921 model of speech perception (Allen, 1996, 2005a; Singh and Allen, 2012).

Issues with decoding, and possibly encoding, become relevant when we consider that more than 15% of children have difficulty learning to read, and 10% or more are diagnosed as having a RD (Torgesen, 2004). Understanding the source of RDs is critically important, as a lack of literacy skills is associated with a number of serious very negative outcomes (Torgesen, 2004, p. 25) (i.e., see the last chapter of Wong (2011)). For instance, according to national statistics, based on the 1994 Washington Summit on Learning Disabilities (Ellis and Cramer, 1994; NICHD/NRP, 2000a,b), 50% of inmates cannot read.

Moreover, an understanding of the precise source of a RD would likely impact the success rate of treatment. If true, and assuming that the encoding process is central to learning to read, one might naturally ask:

1. To what extent is accurate phone encoding (i.e. phonemic wareness) important for reading?

2. Does the disruptions in phone encoding during early childhood lead to RD?

Middle ear infections are common in early childhood, and these frequently lead to a temporary (typically less than one year), undetected 50 dB middle ear hearing loss (Williams and Jacobs, 2009). This can lead to a substantial disruption in exposure to speech sounds during the critical first year, possibly impacting speech development, thus contributing to RD.

Present goals: In this current study, we investigate whether reading development depends on the seemingly-easy task of understanding speech (i.e., phone encoding), and we will show that RD children do not have TD speech perception. Below we investigate these issues by examining the speech perception abilities of RD children and TD *reading control* (RC) children. Specifically, we examine how accurately children recognize and classify speech sounds in a quiet environment (i.e., with no added noise), using a large database (18 talkers) of naturally spoken (CV, VC) speech sounds, that capture the natural variation in speech observed across a diverse set of talkers.

This analysis provides a quantitative measure of the listener's sensitivity to fine phonetic detail in the speech signal, which is necessary for accurate (TD) encoding (Allen, 2005a; Phatak and Allen, 2007; Toscano and Allen, 2014).

Organization: We first briefly review previous work on speech perception in RD children, as well as current models of reading that provide the basis for the experiment presented here. Given the problem formulated here, there are two phonetic perception tasks involved in this study – the Syllable Confusion Oddball (SCO) Task, a speech perception discrimination task; and the Nonsense Syllable Confusion Matrix (NSCM) Task, a speech identification and production task – to evaluate our hypothesis that reading development depends on phone-level perception abilities. From the SCO task, we found that the proportion of errors incurred in the RD group was between 3 5 times greater for RD children when compared to the RC children. (The details regarding the SCO task will be presented in another manuscript which will be submitted in the near future.) Next, we present results from the NSCM task designed to measure listeners' speech perception abilities. Finally, implications for models of reading and approaches for RD interventions are discussed.

2 Previous Research

Many studies have investigated speech perception deficits as a possible source of RD. Most of these studies have used the categorical perception paradigm, where speech sounds are varied along specific acoustic-phonetic continua and listeners' identification and discrimination responses are compared (Liberman et al., 1957). These studies have often used synthetic speech to control for variability between speech sounds (Liberman et al., 1967). Recent work has explored natural speech sound continua Toscano and McMurray (2012). In our current study, natural speech was chosen over synthetic speech, as it provides a much richer set of acoustic cues used by the auditory system, distinct from those found in synthetic speech sounds (Li et al., 2010; Li and Allen, 2011). While this provides a high degree of control over the stimuli, their experimental design does not capture the large and highly relevant variability present in natural speech. Moreover, these studies have often focused on only a small subset of sounds, often examining only stop consonants, which are more accurately recognized than other types of natural speech sounds (Phatak and Allen, 2007; Singh and Allen, 2012).

Brandt and Rosen (1980) used synthetic speech to measure perception of speech sounds in 12 RD children and four TD children, who served as RC subjects. They concluded there was no significant difference in categorical perception between the RD and RC children.

Manis et al. (1997) investigated 25 dyslexic (DYS) children (4th-10th grade), whose performance was compared with 25 chronological age (CA) matched children (5th-8th grade) and 24 reading level (RL) matched children (2nd-3rd grade).

Joanisse et al. (2000) looked at phoneme categorization with 61 DYS 3rd graders (7-10 years old), 52 CA matched 3rd graders, and 37 RL matched 1st and 2nd graders (6-8 years old), testing phoneme categorization with the word pairs *dug-tug* and *spy-sky*.

Results: These studies all found no overall difference in categorization between the DYS and control groups, but a subset of DYS subjects showed shallower categorization functions for both speech sound contrasts.

Messaoud-Galusi et al. (2011) carried out consonant identification and discrimination experiments using several different tasks with 62 dyslexic (DYS) and 51 control children, examining perception of the contrast between /b/ and /p/ in quiet and in 20-talker babble noise. This experiment was a follow-up of a similar study done with adults (Hazan et al., 2009).

Results: The authors conclude that there were no consistent speech perception deficits associated with dyslexia.

Other work has aimed at measuring speech perception in children with RD using larger sets of speech sounds in other types of tasks, that may allow us to better measure perception of sub-phonemic (i.e., phone-level) differences in speech. Hazan and Adlard (1998) measured speech sound discrimination in 13 children with reading delays, 12 reading-age (RA) matched controls, and 12 chronological-age (CA) matched controls.

Results: There was no overall observed difference between the groups, but there was an interaction between group and consonant, such that children in the RD group made more errors for stop consonants. A subset of RD children were also found to make more errors overall compared with RC children across different speech sounds.

Hazan et al. (2013) further investigated identification of consonants from the set /p, b, t, d, f, v, s, z, m, n, sp, st/ and discrimination of sounds varying in in place of articulation (/b/ vs. /d/) and voicing (/b/ vs. /p/). The study included 34 DYS subjects (mean age: 147.3 months) and 25 control subjects (mean age: 146.8 months). Children in the DYS group made more errors in identification, but only for a subset of speech sounds, and they made more errors overall in the discrimination task.

In contrast to some of these studies, Ziegler et al. (2009) found speech perception deficits in a group of 19 DYS subjects compared with 18 chronological-age matched control subjects and 19 reading-level matched control subjects.

Finally, White-Schwoch et al. (2015) tested 112 children using an electrophysiological measure of phone processing (using the syllable /dɑ/ as testing material) with a group of 4-year-old children (N=37; mean age: 54.41 months) and a group of 3-year-old children (N=20; mean age: 43.35 months). They found that poor processing was related to differences in PA empirical scores (PA-ES). In addition, for a subset of children (N=34) who returned a year later, the earlier neuro-physiological measure predicted performance in measures of reading and literacy, again suggesting a link between phone processing and reading ability.

Summary: Work investigating the relationship between speech perception and reading development, while variable, has found no consistent differences between RD and TD children. However, many of these studies have used paradigms such as categorical perception tasks, that do not accurately capture sensitivity to low-level (sub-phonemic) differences in speech, and many have only investigated a small subset of synthetic speech sounds, such as stop consonants. Therefore, a more critical investigation is needed.

The role of two classic studies: As demonstrated by Miller et al. (1951), it is difficult (i.e., it is a very serious mistake) to analyze a speech perception problem using meaningful speech as testing material, due to the large influence of contextual information (Lu, 2018). The importance of the context channel was understood during his development of the *articulation index*

(AI) model of speech perception (Fletcher, 1995). The AI model decomposes the speech perception into a cascade of sound to neural processing elements (Allen, 1996, 2005b). Contextual information processing is the final stage in this cascade (Allen, 2005a).

Given this fundamental understanding of speech perception, phone recognition (encoding) is a more basic (earlier) layer of speech perception. It naturally follows that accurate phone recognition encoding is key to word encoding, and therefore reading skills. This strongly suggests that phone perception determines the success of subsequent decoding and reading comprehension.

Although phonological awareness and decoding—both of which assume accurate encoding speech perception—are viewed as causal factors in RD, as discussed above, many studies showing that speech perception deficits do not seem to be involved. However, few studies (if any) specifically map out the early phonetic encoding abilities of children with reading difficulties.

Hence, in the present study the following questions are addressed:

1. Without access to visual (i.e., letter) and contextual (i.e., word) information, do children with RD show a phonetic encoding deficit?

2. If yes, what type of tasks are most effective in diagnosing RD?

3. What task is most informative about RD regarding perception of specific speech sounds?

4. Do RD and TD children have common, or even unique phonetic perception patterns (i.e., do they have similar or dissimilar patterns of speech sound confusions)? Alternatively are RD and TD children idiosyncratic? If so, to what extent are they consistent in their confusions?

3 Method

There are two phonetic perception tasks involved in this study: the Syllable Confusion Oddball (SCO) Task, a speech perception discrimination task; and the Nonsense Syllable Confusion Matrix (NSCM) Task, a speech identification and production task. During each of these two tasks, the children were given game breaks (five minutes of break for every ten minutes of testing), and enough rest and treats to avoid possible fatigue and boredom. Each child participated in the study for a total of up to 10 weeks, for two 1-hour sessions per week whenever possible. The child first participated in approximately 10 sessions (5 weeks) of the SCO task. Once the SCO sessions were completed, the child participated in 10 more sessions (5 weeks) of the NSCM task. On average, a child performed 1,500 or more trials for each task, in both the RC (a total of 20-40 trials per syllable) and RD groups (a total of 30-40 trials per syllable).

3.1 Participants

The RD group had nine children (six girls), aged 8 to 11 years. The RC group had six children (two girls), aged 8 to 11 years. This is a typical age range during reading development where deficits are discovered, but rarely overcome. Initial interviews were conducted to obtain a family's informed consent for their child to participate in the study. The child was paid a nominal amount at the end of each session in which he or she participated. The child's parent also filled out a comprehensive questionnaire about the child's developmental, health, and educational history, including hearing, speech, language, reading, and writing abilities and any related clinical diagnoses pertaining to RD or dyslexia. All the children with RD were recruited from the Urbana *Reading Group* (3011 Village Office Pl, Champaign, IL 61822). The study was approved by the University of Illinois at Urbana-Champaign IRB.

All parents reported that they had no concerns about their child's hearing. In addition, all but one child passed a pure-tone hearing screening at the beginning of the study (500Hz, 1kHz, 2kHz, and 4kHz in each ear, at 20 dB SPL), indicating normal hearing ability. If at any time during the study an upper respiratory infection was apparent, the screening was repeated before the day's session continued. For the one child who did not pass the initial screening bilaterally, this test was repeated at the beginning of each visit. She passed her screening at every subsequent visit, ensuring that she was not experiencing temporary hearing loss that day. Parents reported no known visual, neurological, cognitive, or emotional problems for these subjects. All the children had a nonverbal IQ in the normal range for their age. Language abilities, including comprehension vocabulary, grammar, and phonemic awareness (word and nonword segmentation) were measured using standardized language tests, as was articulation ability. Additionally, a nonstandardized, widely used measure or nonword repetition was administered. The RC and RD groups differed significantly on all these measures except for articulation. To assess their reading abilities, a battery of reading tests were administered, including the Woodcock Reading Mastery Tests-Revised (WRMT-R), specifically the Word Identification subtest (WI) and Word Attack subtest (WA), and the Grey Oral Reading Test, 4th edition (GORT-4), which included a Fluency score (R-FLU) and Comprehension score (R-COMP; Johnson et al., 2015). For each measure, the children in the RD group scored significantly lower than the children in the RC group (using Welch's t-test): WI (t(9)=4.34, p=0.002), WA (t(11)=5.18, p<0.001), R-FLU (t(7)=4.03, p=0.005), and R-COMP (t(13)=4.64, p<0.001). For all four reading measures, the RD group's mean score was at least one standard deviation below the mean reported in the administration manual of the standardized test.

3.2 Stimuli

Natural speech sounds have more subtle, variable, and realistic perceptual cues than synthetic sounds, thus are considered superior for human speech perception tests (Li, 2009). The set of natural sounds that were used for both tasks came from the commercial Linguistic Data Consortium LDC-2005S22 database (Fousek et al., 2004). The database contains a set of all diphone syllables allowed in English (i.e., CV and VC syllables) in both CI/CF and VI/VF order, formed from 24 consonants and 15 vowels, spoken by 18 talkers.

Stimuli were presented without background noise, in random sequence, at the listener's most comfortable loudness level, chosen by the subject at the beginning of each session. Although sounds could be replayed as many times as a child desired, children requested more than three presentations on fewer than 0.1% of the trials. The child was not given feedback about his or her response accuracy on either task. The children listened through AKG K240 Monitor headphones (circumaural, 600 Ω) via the laptop's 24 bit sound card. Sounds were preprocessed by LDC to remove artifacts (e.g., lip smacks) and loudness irregularities.

3.3 SCO design

For the SCO task, on each trial, three naturally-produced speech syllables (CV or VC) were presented, where two sounds were the same C (or V) and one was different. Only a C or V was modified on a given trial. For example, listeners might hear /ka, ʒa, ka/. They were asked to identify the position of the oddball syllable (second for this example). The oddball was

Table 1: *Average number of trials per child (standard deviation) for the CI, CF, VI, and VF in* **Left:** *SCO Task and* **Right:** *NSCM task. For the RC-SCO-CI task, there are 42 trials and 6 subjects, for a total of 42*6 = 252 presentations in total. For the RC-NSCM-CI task there are 31*6 = 186 trials. An average trial time is less than 10 [sec/trial]. These trials are generated randomly, so the exact number of times that each consonant was presented was not precisely controlled. This number may be computed from the row-sum of the cluster confusion matrix for each child (see Tables. 2, 3 and 4).*

`SCO`	Initial	Final
C	42(19)	44(17)
V	43(19)	45(15)

`NSCM`	Initial	Final
C	31(3)	30(5)
V	42(3)	47(4)

always chosen randomly to occur in one of the three positions. The three sounds were either CV or VC, and were always spoken by three different talkers, chosen randomly from a set of 18 mixed gender talkers. Thus, the three CV/VC tokens were *always* different, due to the talker differences. Based upon their performance, the children understood the task was to identify the oddball syllable based on the C or V difference and understood that they should ignore talker and gender differences.

Similar oddball tasks have been used in previous work on RD (e.g., Bradley and Bryant, 1978, 1983), but these studies used meaningful words. As discussed above, the use of real words as testing materials can lead to significant influence from the context channel (i.e., word meaning) in addition to the auditory channel (what we wish to measure). In the present study, the testing materials are all maximum entropy (*MaxEnt*) in the SCO task, defined as syllables consisting of all possible combinations in English, with equal probability (Singh and Allen, 2012). MaxEnt syllables represent the full range of phonological differences that may occur. Using materials with no meaningful linguistic content allows us to focus entirely on the contribution of the auditory channel to estimate children's phone identification and discrimination abilities.

Our SCO task has extensive coverage of consonants and vowels in English: 24 consonants spanning 15 vowel contexts. The sounds were produced by 18 talkers to better represent the natural variation encountered with speakers of English. Because the children only needed to respond based on the information in the specific tokens, there was no need for them to segment phones from the syllables, or to identify/label them. Hence, the task should be easier, thus more accurate, than a phonological awareness task. In fact experimentally turned out to be more difficult. Moreover, there is no influence of the visual channel in these results, since printed materials were not used. Thus, in this experiment the children were forced to rely exclusively on the auditory channel.

As shown in Table 1 (left), the number of trials averaged across consonants in initial position $\mu_{ci} = 42$ for all RC children with a standard deviation σ_{ci} of 19; for VI the mean is $\mu_{vi} = 43$ with a standard deviation of $\sigma_{vi} = 19$; $\mu_{cf} = 44$ consonants in final position4 with a standard deviation of $\sigma_{cf} = 17$; and $\mu_{vf} = 45$ for vowels in final position with a standard deviation of $\sigma_{ff} = 15$.

3.4 NSCM design

The NSCM task complements the SCO task in that it provides confusion data, which are unavailable from the SCO task (since it only identifies which sounds are confusable, not which sounds they are confused with). In the NSCM task, the listener hears a single CV or VC and is instructed to orally repeat the syllable. The downside of this test is that it requires two transcribers to code the verbal report. This is error prone since the children do not always clearly articulate their response, in fact they frequently misarticulate because they are unsure of the identity of the spoken token. Nonetheless, this task provides useful data, missing from the SCO task. In particular, the NSCM task adds value by providing confusion matrix information (Miller and Nicely, 1955), which can be used during training sessions and as diagnostic feedback on any change in the child's status. Knowledge of these confusions would also be useful to a speech therapist.

Based on Table 1, the average number of trials for consonants in syllable-initial position was 31, with a standard deviation of 3; the average number of trials for vowels in syllable-initial position was 42, with a standard deviation of 3; the average number of trials for consonants in syllable-final position is 30, with a standard deviation of 5; and the average number of trials for vowels in syllable-final position was 47, with a standard deviation of 4.

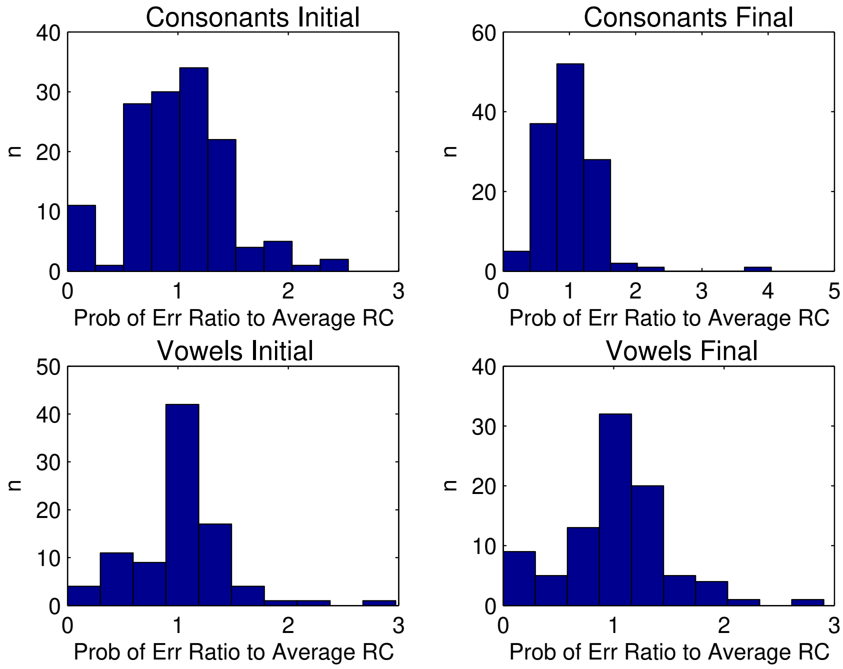


Figure 2: *Histograms of the frequency of error ratios of the six individual RC errors to the average RC normal (AN) for the NSCM data. In all four cases, the error ratio of 1 is the mode of the histogram distribution, subjects on the left of the mode represent better than average performance than the average RC normal (AN), and subjects on the right represent worse than average error rates. For example, for the case of Consonants Initial (CI), for the six RC subjects, there are 10 consonants where the error ratio (compared with the AN subject) is among the range between 0 and 0.25, while 35 consonants have an error ratio around 1 (approximately equal to the average). Four consonants for the six RC children have twice the CI error relative to the average performance. Note that Consonants Final (CF) and Vowels Initial (VI) sounds are nearly identical to the group average (the distributions are all close to the mode at 1). We conclude that the AN subject does an excellent job of representing the six RC subjects. We shall show that is* not *the case for the idiosyncratic RD subjects.*

4 Results

4.1 RC subjects span the NSCM data set

4.1.1 Error Analysis

On average, the six RC subjects have 15.44% error for consonants, 22.04% error for vowels, and 18.11% for all the phones. By comparing the ratio of the individual errors to the average RC normal (AN) from Fig. 2, we found that all the histograms are unimodal for the four consonant cases. RC children as a group performed similarly, with few outliers. The ratio ranges from 0.5 to 2. A profile of the hypothetical control subject Average Normal (AN) representing the average performance of the control group is thus created to characterize the space of the data set. It serves as the average normal (AN) control subject inside the RC group, allowing us to identify the general confusion patterns among the normal children, and to compare the RD and RC groups. Additionally, the AN subject is a useful representation of the RC group.

Table 2: *Clustered Confusion Matrix for RC-Anton who has two cluster groups. Clustering depends on a threshold. Here errors of 4 or less are not considered significant. Note that /k/ was presented 95 times, so 4 is assumed to be well below chance. /p/ on the other hand was presented 68 times. For this case we define 4 as the empirical threshold.*

	p	t	k	f	θ	ð	v	w	d	g	z	b	ʃ	ʒ	ʤ	ʧ	m	n	h	s	j	r	L
p	63	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.
t	2	71	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
k	.	.	93	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.
f	.	.	.	65	24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.
θ	.	.	.	6	50	6	2	.	2	.	.	.	.	.	.	.	.	.	.	2	.	.	.
ð	.	.	.	.	10	47	6	4	3	.	2	2	.	.	.	.	.	.	.	.	.	.	2
v	.	.	.	4	2	8	50	13	.	.	.	4	.	.	.	.	.	.	.	.	.	.	2
w	.	.	.	.	.	.	11	72	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
d	.	.	.	.	.	.	.	.	86	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	.	.	2	.	.	.	.	.	2	86	.	.	.	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	.
b	.	.	.	.	.	.	2	.	2	.	.	80	.	.	.	.	.	.	.	.	.	.	.
ʃ	.	.	.	.	.	.	.	.	.	.	.	.	73	.	.	.	.	.	.	.	.	.	.
ʒ	.	.	.	.	.	.	.	.	.	.	2	.	6	65	7	.	.	.	.	.	.	.	.
ʤ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	85	4	.	.	.	.	.	.	.
ʧ	.	2	.	.	.	.	.	.	.	.	.	.	.	.	2	75	.	.	.	.	.	.	.
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	47	.	.	.	.	.	.
n	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	81	.	.	.	.	.
h	2	.	2	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	62	.	.	.	.
s	.	.	.	.	2	.	.	.	.	.	2	.	.	.	.	.	.	.	.	103	.	.	.
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	73	.	.
r	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	82	.
L	.	.	.	.	.	.	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	59

4.1.2 NSCM Confusion Matrix Analysis

To better explore the confusion patterns in perceiving phones, confusion matrices are generated for the NSCM task to show a) with which sound and b) how many times each subject is confused. In each matrix, the rows are spoken sounds and the columns are heard sounds.

Table 2 shows the clustered Confusion Matrix for RC-Anton. His largest error was to respond 24 times with /θ/ when presented with /f/. The highlighted blocks are Anton's two confusion groups, which are '/f/-/θ/-/ð/-/v/-/w/' and '/ʃ/-/ʒ/-/ʤ/'.

Table 3 shows the clustered confusion matrix for subject RC-Evan who seems to have 3 confusion groups, '/f/- /θ/-/ð/-/s/', '/f/-/v/-/b/ and '/ʃ/-/ʒ/-/ʤ/'. RC-Evan and RC-Anton

	p	b	v	f	θ	ð	s	d	k	t	z	g	ʃ	ʒ	ʤ	n	m	ʧ	h	w	j	r	L
p	66	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	4	.	.	.	.
b	.	68	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.
v	1	13	31	6	.	5	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	.	.	.	61	9	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
θ	.	.	.	10	44	11	8	1	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.
ð	.	.	2	1	7	39	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	59	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.
d	.	.	.	.	.	.	.	72	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.
k	.	.	.	.	.	.	.	.	60	2	.	.	.	.	.	.	.	.	.	.	.	.	.
t	.	.	.	.	.	.	.	.	2	70	.	.	.	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	.	4	.	.	.	58	.	.	2	.	.	.	.	.	.	.	.	.
g	.	.	.	.	.	.	.	.	2	.	.	54	.	.	.	.	.	.	.	.	.	.	.
ʃ	.	.	.	.	.	.	.	.	.	.	.	.	64	4	.	.	.	2	.	.	.	.	.
ʒ	.	.	.	.	.	.	.	.	.	.	1	.	8	59	6	.	.	.	.	.	.	.	.
ʤ	.	.	.	.	.	.	.	.	.	.	1	.	.	.	51	.	.	.	.	.	.	.	.
n	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	63	.	.	.	.	.	.	1
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	43	.	.	2	.	.	.
ʧ	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	56	.	.	.	.	.
h	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	70	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	58	.	.	.
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	56	.	.
r	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	53	1
L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	63

Table 3: *Clustered Confusion Matrix for RC-Evan who has three well defined confusion groups, based on a threshold of 4 (≤ 4 errors are considered chance)*

share a fraction of the confusions, yet have different confusion pairs.

Table 4 is the clustered confusion matrix for the representative matrix of the RC subjects (the AN subject) where we averaged all the errors for CI sounds across the RC subjects. Once again, '/f/-/v/-/θ/-/ð/' is one of the confusion groups. The other confusion group is '/s/-/z/-/ʒ/-/ʤ/' which also shares '/ʒ/-/ʤ/' with Anton and Evan.

With the average RC normal (AN) subject having performance similar to the other RC subjects, we found that the AN subject have a significant overlap in confusion groups, which suggested that six RC subjects are sufficient to draw the pattern of confusion groups of RC subjects and that RC subjects can be well-represented by AN subject with the similar confusion groups and errors. More subjects would include more small idiosyncratic errors, but in general the average RC normal (AN) would not change.

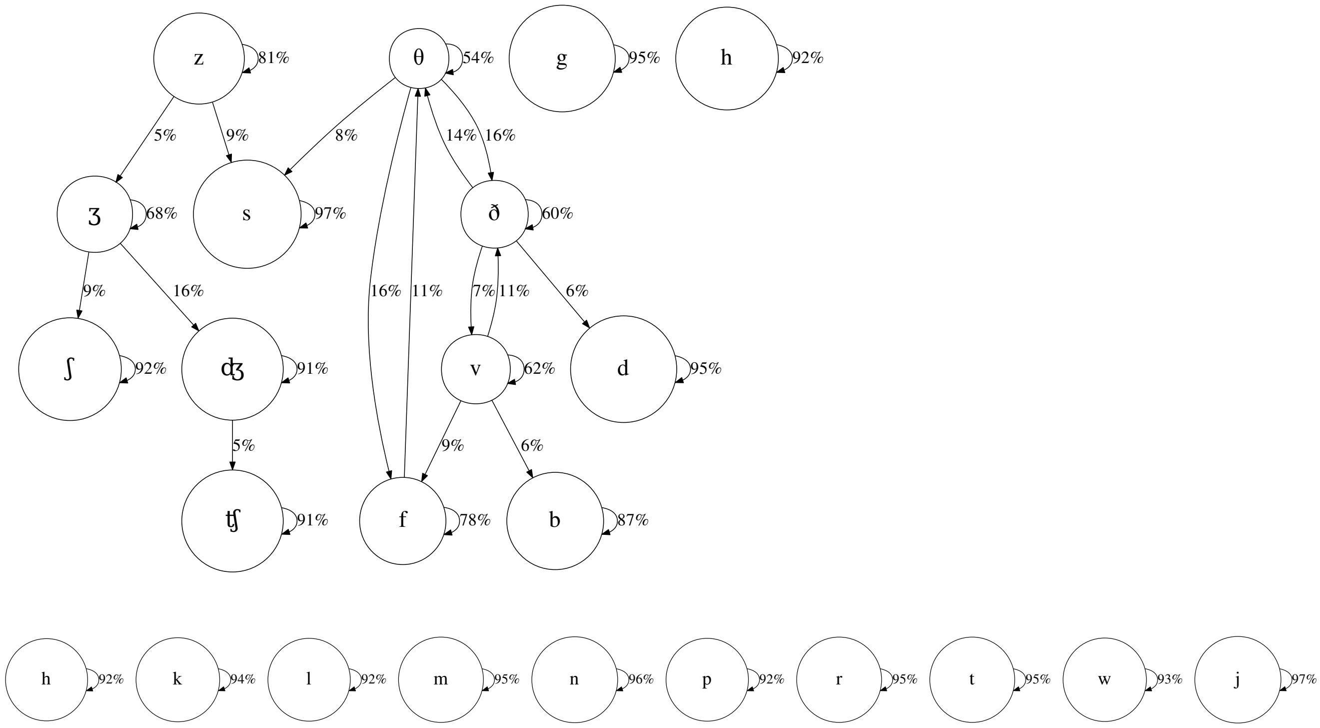
4.1.3 Directed Graph Analysis

To better visualize the perceptual confusion patterns, the confusion matrix data may be analyzed as directed graphs, which provides a graphical method for summarizing a confusion matrix. In a graph, nodes represent individual phones that the listener hears, and arrows between nodes depict the listener's confusions. Loops from the node back to itself represent correct responses. Connections to other nodes are errors, with the percentage written above each connection indicating how often that particular confusion was made. In this way the confusing alternative phones are captured as transition paths from the spoken phone. The benefit of using this type of visualization is that it offers a direct view of the child's high-error phones, which show where to focus RD intervention.

To understand the average phonetic perception for normal reading children, the mean values of confusion ratios between phones were extracted from the RC data as the AN subject in the directed graph of Fig. 4. As shown in the figure, children with normal reading abilities could identify all presented phones with at least 54% accuracy, and had at least 91% accuracy in identifying phones such as /g/, /h/, /k/, /l/, /m/, /n/, /p/, /r/, /t/, /w/, /s/, /ʃ/, /tʃ/,

Table 4: **Top:** *Clustered Confusion Matrix for the average RC normal (AN) subject in CI position. Assuming that chance performance is ≤ 5 trials, we have clustered the sounds as shown here. The resulting clusters above chance are /f/, /θ/, /ð/, /v/, and /z/, /ʒ/, /ʤ/. If we assume 5 and 4 responses are above chance, then we would have more one-way confusions between /θ/ → /s/, /ð/ → /d/ and /ʤ/ → /ʧ/, which have been shaded green.*
Bottom: *Phonetic Confusion Patterns for the average RC normal (AN) subject in CI position. At the top are the directed graphs of the sounds with confusions greater than 8%. Note how the center of the confusions is Θ (Th). Also shown are /g,h/ with 5% and 8% error each. Below are the sounds with confusions less than 9% error (/h/ is included in this group for continuity).*

	p	t	k	f	θ	ð	v	b	d	g	ʃ	s	z	ʒ	ʤ	n	m	ʧ	h	w	y	r	ʎ
p	61	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.
t	.	64	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
k	.	1	58	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	.	.	.	53	8	2	1	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.
θ	.	.	.	9	33	9	.	.	1	.	.	5	.	.	.	.	.	.	.	.	.	.	.
ð	.	.	.	.	10	42	5	1	4	.	.	.	1	.	.	.	.	.	.	.	.	.	2
v	.	.	.	6	1	7	43	4	1	.	.	.	.	.	.	.	.	.	.	2	.	1	1
b	.	.	.	.	.	.	2	61	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.
d	.	.	.	.	.	1	.	.	61	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	.	.	1	.	.	.	.	.	.	62	.	.	.	.	.	.	.	.	.	.	.	.	.
ʃ	.	.	.	.	.	.	.	.	.	.	64	1	.	1	.	.	.	2	.	.	.	.	.
s	.	.	.	.	1	.	.	.	.	.	.	72	1	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	.	.	.	.	.	.	6	55	5	1	.	.	.	.	.	.	.	.
ʒ	.	.	.	.	.	.	.	.	.	.	5	.	2	42	10	.	.	1	.	.	.	.	.
ʤ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	63	.	.	4	.	.	.	.	.
n	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	63	1	.	.	.	.	.	.
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	56	.	.	.	.	.	.
ʧ	.	2	.	.	.	.	.	.	.	.	2	.	.	.	1	.	.	70	.	.	.	.	.
h	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	58	.	.	.	.
w	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	59	.	.	1
y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	58	.	.
r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	55	.
ʎ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	57



/d/, /ʤ/, and /j/ when these consonants were put in the syllable-initial position. Confusion patterns that appeared more than 10% of the time were generally for fricative and affricate targets: θ→ð (16%), θ→f (16%), ʒ→ʤ(16%), ð→θ (14%), f→θ (11%), v→ð (11%). Confusion patterns that appeared less than 10% of the time but were still notable happened between obstruents (sibilants, fricatives, affricates, and stops): v→f (9%), z→s (9%), ʒ→ʃ (9%), θ→s (8%), ð→v(7%), ð→d(6%), v→b (6%), ʤ→tʃ (5%), and z→ʒ (5%). Based on the connections depicted in Fig. 4, there was a clear separation between these confusing phones and those intact phones like nasals, glides and stops; however, inside the confusion patterns, there were no obvious boundaries for phones with different features. They all belong to a large group. Phones at the center of confusions were /θ, ð, f, v/ and /ʒ/. From these confusion patterns, we can understand the obstacles that exist in perceiving natural English for normal reading children. Alternatively these errors could be due to talker errors, detected by our average RC normal (AN) subject. In fact it was noted by Phatak et al. (2008); Phatak and Allen (2007) that the talker error in this database is 20%, which is larger than the average listener error for the case of no added noise.

It is interesting (and perhaps important) to see that even with a certain degree of perceptual errors, the speech is sufficiently robust for us to accurately communicate.

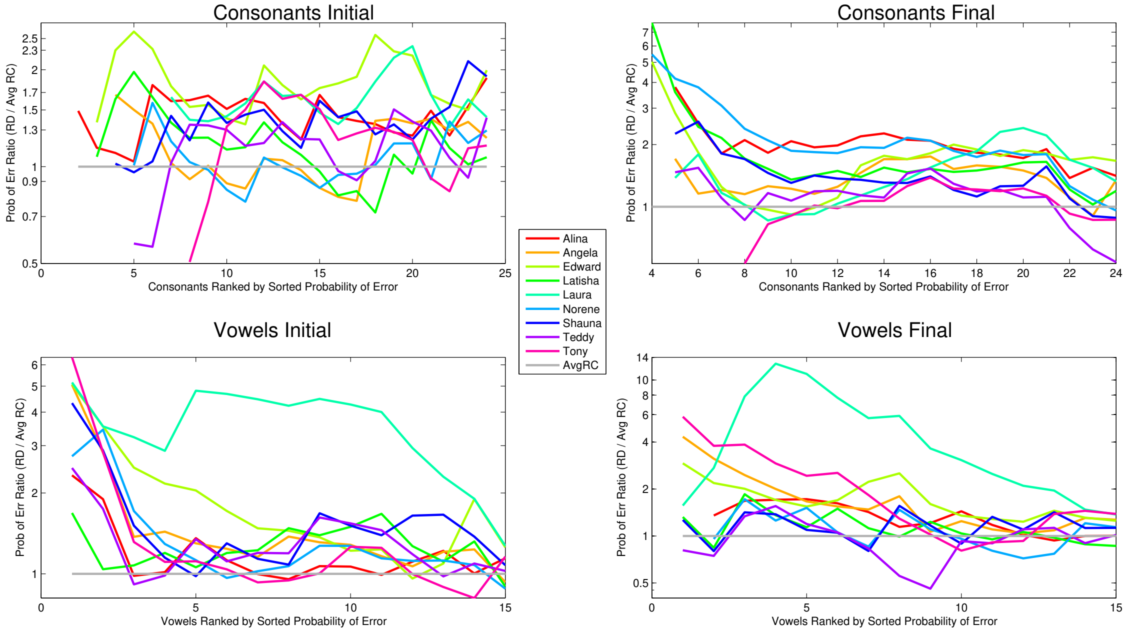


Figure 3: *Ratios of the 10 RD NSCM phone empirical scores (ES) divided by the average RC normal (AN) ES, as a function of the phones being tested, for the four conditions (CI, CF, VI, VF). These plots quantify the idiosyncratic nature of the RD scores relative to the low-error AN scores. For example, in the upper-left panel, each of the 10 subject lines, as indicated by the legend, corresponds the 25 CI phone scores divided by the 25 CI AN sorted scores. The AN scores were sorted with increasing error (from smallest to largest). Thus each lines represents the AN normalized scores for the 10 RD children. The horizontal gray line at 1 indicates where the RD and AN errors are equal (which is rare). The few scores not plotted have a score of zero (subjects Teddy and Tony), which cannot be represented on the log scale. Since most of the RD errors are significantly large than the AN scores (smaller ES), the ratios are mostly above 1. In the three remaining panels (CF, VI, VF) the ratios general decrease due to the AN sorting.*

4.2 Idiosyncratic RD subject errors

Figure 3 is a plot of the ten NSCM RD subject empirical scores (ES normalized by the AN scores, on a log error scale). This chart demonstrate the degree to which the ten RD subjects are idiosyncratic. The four panels are CI/CF (upper) and VI/VF (lower). The abscissas for each panel are sorted by the small but systematic AN phone scores, for each of the four test conditions (CI, CF, VI, VF). The horizontal grey line is set to 1 for reference. Points below this line indicates the RD subject out-performed the AN subject (rare). Each of the ten subjects is shown as a line on the chart.

For example, the upper-left panel for the CI case shows how the ten RD subjects compare to the AN subject. In this panel most score ratios are between 1 and 2.5. In a few cases the error drops below the AN error, but on average the error-ratio is between 1 and 2, with a maximum of 2.5. This chart shows, with only a small number of exceptions, that the RD error is between 1 and 2.5 times the average RC error, but otherwise random. The ratio is roughly uniformly distributed over this range.

Consonant final (upper-right) tells a similar story, but with a different distribution having a larger spread. A few CF sounds have scores, relative to AN, between 1.5 to 7. Subject Alna has 9 zero-error scores, with the remaining sounds having error ratios between 1 and 1.5.

The vowels tell a somewhat different story. One RD child (Latisha) has a huge relative ES around 5, with almost no high error phones. The remainder of the subject scores are mostly below between 1 and 1.5. The VF story is similar except that Latisha's maximum ES is 14. Three RD subjects show maximum errors between 6 and 2. The remainder have errors between 1.5 and 0.5 (i.e., Teddy).

Thus Fig. 3 shows that RD subjects in general tend to show a higher error especially on the similar confusion sounds when compared with RC and have poorly concentrated confusion groups. On average, RDs have 21.27% error for consonants, 29.77% error for vowels, and 24.71% averaged over all the phones. This is similar to the normal hearing error found in earlier studies for no added noise (Phatak and Allen, 2007; Phatak et al., 2008).

The distribution of probability of error ratios of RDs to the average RC level are presented in Fig. 4. The relative performance for RD children, when compared with average RC normal (AN), ranges mostly from one (same) to three. More to the point, RD children have unique highly idiosyncratic confusion scores, and have either much higher errors, or unique confusions, or both.

One significant point is that if we were to increase the number of RD subjects, we would not reveal distinct patterns, because of the idiosyncratic nature of the RD children (not observed in our six RC children).

In summary: Four major characteristics of the RD subjects may been identified:

1. RD subjects' confusion groups for errors are idiosyncratic.
2. RD subjects have much higher error on some confusion pairs when compared to RC subjects.
3. It is unusual for an RD subject to outperform the RC subjects. Given the numbers of subjects in our experiments, there is very little overlap in the RC and RD distributions.
4. An entropy analysis (Singh and Allen, 2012) may be used to further quantify the nature of subject idiosyncrasies.

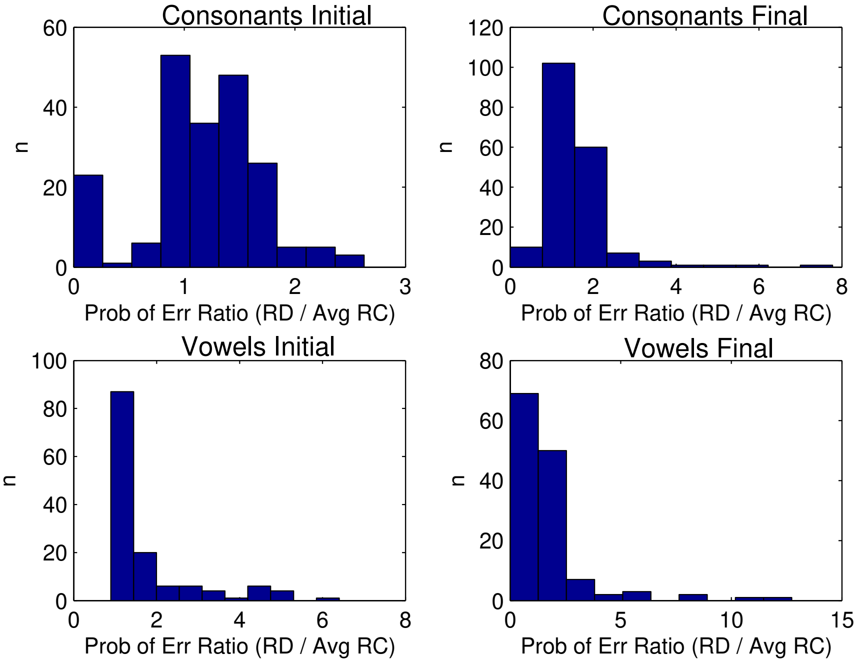


Figure 4: *Histograms of probabilities of error ratios of individual RD to the average RC normal (AN). The values of ratios for all phones are mostly higher than the AN. That is, RDs were performing at a level worse than the AN.*

In addition, the major characteristics identified for RC subjects are:

1. RC subjects as a group shows similar inter-confusion patterns and reduced intra-confusion variance.
2. A *fictitious* average RC normal (AN) subject has been defined to modeled the average RC performance.

4.2.1 Entropy Analysis

Entropies for the errors in all four conditions were calculated as an indicator of the degree of diversity or inconsistency in responses. The formal definition of entropy $\mathcal{H}$ is the expected value ($\mathcal{E}$) of the log of the information $\mathcal{I}_k = 1/p_k$, where p_k is the probability (i.e, empirical score) of the k phone. The units of probability is *certainty*, and the log is base-2. In terms of information $\mathcal{I}_k$ the entropy is

$$\begin{aligned} \mathcal{H} &\equiv \mathcal{E}(\log_2 \mathcal{I}_{n|j}) && (1) \\ &= \sum_{n=1}^{N} p_{n|j} \log_2 \frac{1}{p_{n|j}} && (2) \\ &= -\sum_{n=1}^{N} p_{n|j} \log_2 p_{n|j} && (3) \end{aligned}$$

which is measured in [*bits*]. Here $I_{n|j}$ is information density as $I_{n|j} = \frac{1}{p_{n|j}}$ and $p_{n|j}$ is the probability of reporting event n (phone) given (conditioned on) event j (Cover and Thomas, 2006).

The value of entropy is determined by the amount of information rooted in the source (in our case, the designed experiment) and the processing and delivery capability of the channel and output device (in our case, the speech unit perception capability of the children). So, a high entropy value may reflect that children received too few helpful cues for them to correctly identify the sounds. In other words, if the entropy is high, the child may be guessing about which phone he or she heard. Entropy can be calculated with Eq. 1. Here, N is the number of trials.

In Figure 5 (left) the points represent the RD entropy versus the probability of error data, while the panel on the right the points represent entropy for the RC children. The entropy for each child is shown as 2 letter acronyms for the test-child's name. The solid lines are for entropy reference curves represent 1, 2, 3 possible outcomes. The first (lowest) line which has it maximum at 50% (purple) shows is entropy as a function of probability p_k for the two-outcomes, or 1 [bit] case, which is maximum at 50% (equal error for two outcomes). The next line (blue) describes the condition for three outcomes (one correct and two wrong), corresponding to 1.5 [bits] (2/3 error at the maximum). The third line (cyan) describes the situation for 2 [bits] (four outcomes) which is maximum at 75%.

For the RC case (Figure 5, RIGHT) no subject comes close to the MaxEnt (peak) entropy. For the RD case however on the left, several subjects (Td, Lr, AI, NR, Sn, . . .) are near or beyond the MaxEnt point. Recall MaxEnt stands for the maximum possible entropy, consistent with idiosyncratic responses. When the error is greater than the MaxEnt value, the entropy returns to zero. This means that the wrong sound (or sounds for more than 1 [bit]) is consistently reported (the correct response is less likely to be called out, indicating that the child is guessing).

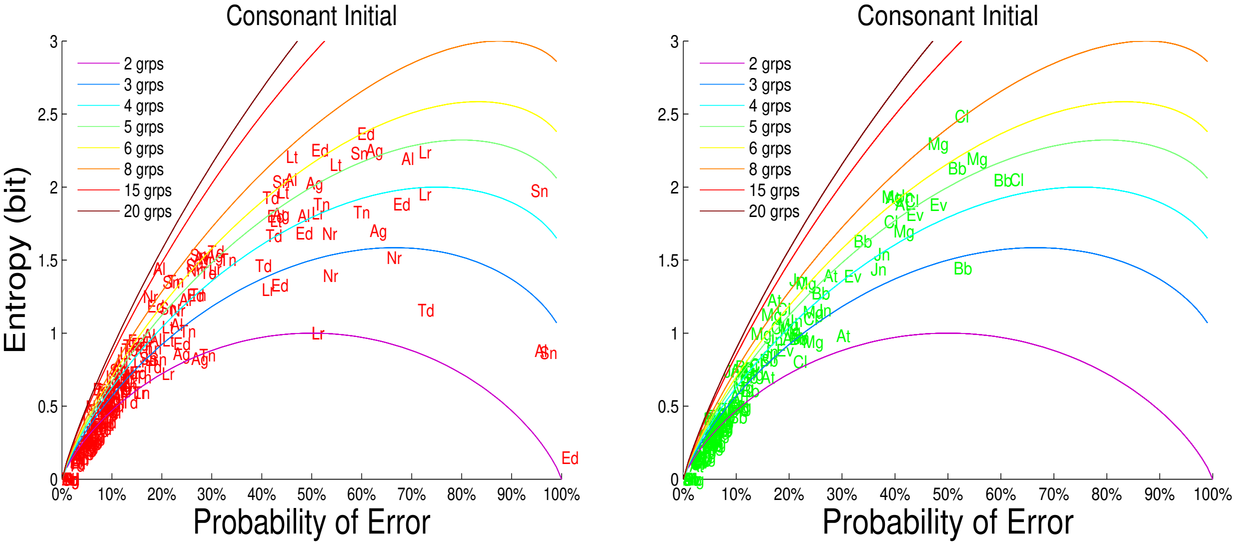


Figure 5: **LEFT:** *The performance of RDs in the Consonant Initial condition are plotted in red. The number of confusions ranges from one to seven. Not the increased concentration of errors in the 40-50% error range, and above 90%.* **RIGHT:** *The performance of RCs in the Consonant Initial condition are plotted in green. The number of confusions ranges from one to seven. Note that there are no errors above 60% and a reduced number above 30%. There are many fewer errors between 5 and 6 group errors*

To further explore the error patterns between RD subjects and RC subjects, we investigated the error distributions specifically for three major categories shown in Fig. 6. Figure 6 plots phonetic perception for phones with one confusions (1 bit]). Figure 6 plots phonetic perception for phones with three confusions (1.5 [bits]). For all conditions we see that the RD subjects

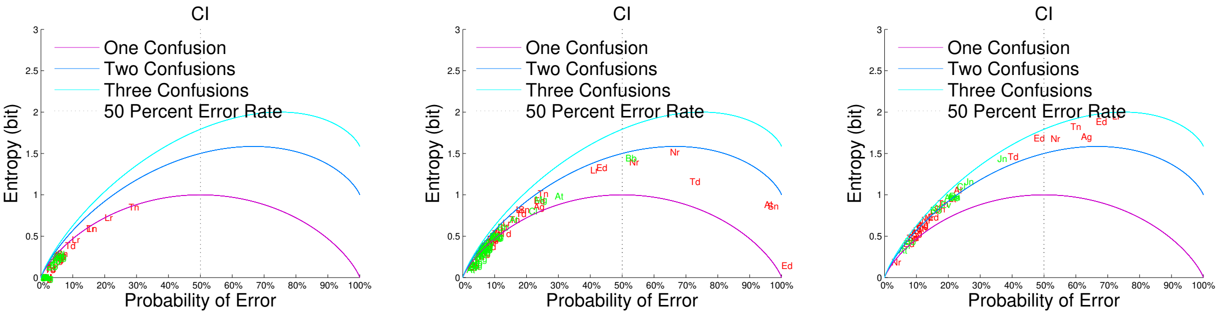


Figure 6: **Left:** *One-Confusion summary for CI: most one-confusion phones for the RCs have less than 10% error. RDs who have high error one-confusion phones are Tony, Laura, and Teddy.* **Center:** *Two-Confusion summary for CI: most two-confusion phones for the RCs have less than 30% error. RDs who have high error two-confusion phones are Norene, Edward, Tony, Laura, and Teddy.* **Right:** *Three-Confusion summary for CI: all three-confusion phones for the RCs have less than 40% error. RDs who have slightly high error three-confusion phones are Tony, Norene, Edward, Angela, Laura, and Teddy.*

(red points) have much higher errors than the RC subjects (green), thus are more idiosyncratic.

4.2.2 Confusion Matrix Analysis

To further illustrate the idiosyncratic error patterns, confusion matrix analysis is required. Table 5 shows the clustered confusion matrix for the RD nominal-error subject Angela. Consistent with RC subjects, Angela has confusions with '/f/-/v/-/θ/(T)-/ð/(D)', but shows a larger confusion between '/f/-/θ/(T)-/ð/(D)' with errors greater than 20%. Note that more than half of the responses she confused /θ/(T) with three other consonants, which is rare among RC subjects. Similarly, the confusion group consisting of /ʃ/(S), /ʒ/(Z), /ʤ/(J) and /ʧ/(C), which shows that Angela has very poor performance compared to the RC children. Angela also shows a mild confusion with /ð/(D) and /L/, which is rare for the RD subjects.

Table 5: *Clustered Confusion Matrix for RD-Angela. Here /θ/ is confused with three other sounds, forming a clear 3-groups with /f/, /ð/and /s/ while /ð/forms a 3-group with /ð/, /v/ and /L/. Also /ʒ/ is split equally with /ʤ/, forming a 2-group.*

	p	t	k	f	θ	ð	s	v	z	b	L	d	ʃ	ʒ	ʤ	ʧ	m	n	h	w	j	r	g
p	64	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1	.	.	.
t	.	44	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.
k	.	.	34	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	.	.	.	34	1	1	4	3	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.
θ	.	1	.	14	30	16	10	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
ð	.	.	.	.	16	29	1	7	2	.	7	2	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	45	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
v	.	.	.	7	.	6	.	33	.	8	.	.	.	.	.	.	.	.	.	.	.	1	.
z	.	.	.	.	.	.	11	.	36	.	.	.	1	.	.	.	.	.	.	.	.	.	.
b	1	.	.	1	.	.	.	1	.	53	.	.	.	.	.	.	.	.	.	.	.	.	.
L	.	.	.	2	.	.	.	.	.	.	44	.	.	.	.	.	.	.	.	.	.	.	.
d	.	1	.	.	1	.	.	.	.	.	.	44	.	.	.	.	.	.	.	.	.	.	2
ʃ	.	.	.	.	.	.	1	.	.	.	.	.	31	.	.	.	.	.	.	.	.	.	.
ʒ	.	.	.	.	.	.	.	.	1	.	.	.	6	23	22	3	.	.	.	.	.	.	.
ʤ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	39	12	.	.	.	.	.	.	.
ʧ	.	2	.	.	.	.	.	.	.	.	.	.	2	.	.	35	.	.	.	.	.	.	.
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	1	.	.	.	.	.
n	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	51	.	.	.	.	.
h	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	.	.	.	.
w	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	28	.	2	.
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	35	.	.
r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	34	.
g	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	47

Table 6 shows a second RD example. First Norene has smaller confusion groups than Angela. Besides the common confusion pair '/θ/(T)-/ð/', Norene performs much worse on consonant /v/, confusing with /ð/(D) approximately 50% of the time. For consonant /ʒ/(Z), is confused

Table 6: *Clustered Confusion Matrix for RD-Norene who confuses /θ/and /v/ with /ð/and /ʒ/ with /ʤ/, well above chance. Thus /ð/is split three ways, as /ð/, /θ/and /v/. None of these confusions are symmetric (/ð/is rarely confused with /θ/(below chance ≤ 6 trials.), yet there is a bias for /ð/given /v/ or /θ/.*

	p	t	k	f	θ	ð	v	s	d	g	S	b	z	ʒ	ʤ	n	m	ʧ	h	w	j	r	L
p	42	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
t	.	46	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.
k	.	.	77	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	1	.	.	56	8	3	2	2	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
θ	.	.	.	4	31	22	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ð	.	.	.	.	4	51	.	.	4	.	.	.	4	.	.	.	4	.	.	.	.	.	2
v	.	.	.	.	4	18	19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	34	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	.	.	.	.	.	1	.	.	30	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	.	.	2	.	.	.	.	.	.	39	.	.	.	.	.	.	.	.	.	.	.	.	.
S	.	.	.	.	.	.	.	5	.	.	65	.	.	.	.	.	.	2	.	.	.	.	.
b	.	.	.	.	.	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	3	.	.	2	.	.	.	32	1	.	.	.	.	.	.	.	.	.
ʒ	.	.	.	.	.	.	.	.	.	.	2	.	3	13	21	.	.	.	.	.	.	.	.
ʤ	.	.	.	.	.	.	.	.	.	1	.	.	.	3	28	.	.	2	.	.	2	.	.
n	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	35	.	.	.	.	.	.	3
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	67	.	.	.	.	.	.
ʧ	.	2	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	58	.	.	.	.	.
h	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	30	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	34	.	.	.
j	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	42	.	.
r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43	.
L	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	54

with /ʤ/(J) approximately two-thirds of the time. Norene performs better than the average RD subject most of the time, but shows major confusions on three pairs: '/θ/(T)-/ð/(D)', '/v/-/ð/(D)' and '/ʒ/(Z)-/ʤ/(J)'. Note that Angela confuses /θ/(T) with /f/, /ð/(D) and /s/, and confuses /ʒ/(Z) with /ʤ/(J) and /ʧ/(C). In addition, Angela has minor confusions on consonants /b/ and /L/ which are seldom confused by other RD subjects.

The comparison between two RD subjects supports our RD idiosyncratic error hypothesis. Unlike RC subjects, who are well represented by average RC normal (AN) there are no such average patterns.

5 Discussion and Conclusions

1. The group of 6 RC children has been accurately summarized as the single average RC normal (AN) subject. This AN subject was found to have only a few (two) minor confusions which are consistent across the six RC subjects. We then compared the RC-AN subject to the 10 RD subjects.

   These comparisons show that, unlike the AN subject, the RD subjects are highly idiosyncratic. We conclude that increasing the number of RD subjects would not prove useful, rather it would only create increased entropy (random examples), not order. While obvious patterns have been found in each of the 10 RD subjects, there seems to be only weak correlations between the RD subjects. Not obvious correlations of confusion between RD subjects have been found.

2. We also demonstrated the idiosyncratic nature of the RD children by the use of several other techniques. First we explored the nature of the random errors in Fig. 3. The distributions of normalized errors were found to be very different between the four types of syllables (CI, CF, VI, VF). In Fig. 4 we found the error distributions are close to the mode, meaning the errors were neither small nor patterned.

   We looked at the errors themselves, by plotting the phone entropy as a function the error.

By a direct comparison of the RCs (right) with the RDs (left) in Fig. 5 showed a large increase in the RD entropy.

The main cluster of low entropy is tightly grouped, making it difficult to see (all the points are on top of each other). For the higher entropy, high error tokens, there is a widely separated distribution. A further break down in Fig. 6 parses out the 1, 2 and 3 confusion groups, further verifying our idiosyncratic hypothesis.

3. In Tables 5 and 6 we compare RD subjects Angela and Norene who show conflicting confusion patterns, again supporting our conclusions.

4. While the points about encoding and decoding have been emphasized many times before, perhaps the point has not been emphasized that the difficulty with decoding (reading) has to do with how the brain develops. The neuroscience of speech processing has made great strides in the last decades and we expect the pace to accelerate in the near future. These studies will likely lead to some solutions to this century old problem.

5. Another important and perhaps related area is how the inner ear and early auditory brain decodes primitive speech sounds (phones). Our research shows that when important speech cues are missed, normal hearing listeners confuse CV and VC sounds (Allen and Li, 2009; Li and Allen, 2011).

6. While it seems unlikely that the SCO and NSCM could be used in the clinic, due to their complexity, some related simplified adaptive strategies could be developed.

7. In conclusion, we believe that the case for a strong causal correlation between phone recognition and reading disability is strongly supported. Yet there is a lot to do. Eventually one must show that after early detection of these RD errors, one can with feedback reduce them and that this is the path to normal reading.

6 Acknowledgments

A large portion of this work was done by Jie Lu as part of her PhD Thesis. Without Ms Lu, this would not have happened.

We are also grateful to JARO for considering this work, which is relevant to their goals because it is directly related to the neuro-science of how the auditory brain processes speech sounds.

Appendices

A Summary of sounds Used in this study

Table 7: Table of common abbreviations (acronyms).

AI	articulation index (measure of phone intelligibility)
AN	average RC normal (AN) subject
CA	chronological age matched subjects
CI/CF	consonant in initial/final syllable position
DYS	dyslexia or dyslexic subjects
decoding	word to speech sound (in auditory perception); print to sound in reading
encoding	speech sound to word (in auditory perception); sound to print (in spelling or writing)
HI	hearing impairment
NSCM	nonsense syllable confusion matrix task
phoneme	smallest unit of meaningful speech (i.e., distinct, contrastive speech sounds in the language)
phone	a speech sound (i.e., an individual instance of a spoken or heard phoneme in the speech stream)
MaxEnt	maximum entropy syllable (constructed by random selection of consonants and vowels)
RL	reading level (for matching the reading performance of subjects)
RC	reading control subject (with a typical RL)
RD	reading disabled subject/reading disability
SCO	syllable confusion oddball task
ES	empirical score
SNR	signal to noise ratio (dB)
TD	typically developing
VI/VF	vowel in initial/final syllable position
WI	Word Identification subtest
WA	Word Attack subtest
PA	phonemic awareness

B Summary of sounds Used in this study

Table 8: *Conversion from Darpabet to International Phonetic Alphabet for LDC unvoiced consonants, voiced consonants, and vowels.*

Unvoiced Consonants		**Voiced Consonants**		**Vowels**		
Dbet	IPA	Dbet	IPA	Dbet	IPA	L/T
C	ʧ	D	ð	@	æ	
S	ʃ	b	b	A	ʌ	T
T	θ	d	d	E	ɛ	L
f	f	g	g	I	ɪ	L
h	h	J	ʤ	O	ɔɪ	
k	k	l	l	R	ɝ	
p	p	m	m	U	ʊ	L
s	s	n	n	W	aʊ	
t	t	r	r	Y	aɪ	
H	ʍ	Z	ʒ	c	ɔ	T
		G	ŋ	a	ɑ	
		v	v	e	e / eɪ	T
		w	w	i	i	T
		y	j	o	o / oʊ	T
		z	z	u	u	T
					ə	L

C Subject Information

Table 9: *Originally there were 19 subjects. Data of four subjects were left out because of lack of data: Savannah and Lucas barely started the study and only did a few sessions, Matt and Tina only completed the SCO task but not the NSCM.*

Pseudonym	Acronym	Group	Age
Anton	At	RC	11;4
Bob	Bb	RC	9;10
Carly	Cl	RC	8;9
Evan	Ev	RC	11;6
Joanna	Jn	RC	10;3
Miguel	Mg	RC	10;6
Alina	Al	RD	10;8
Angela	Ag	RD	9;0
Edward	Ed	RD	8;5
Latisha	Lt	RD	8;5
Laura	Lr	RD	9;11
Norene	Nr	RD	9;10
Shauna	Sn	RD	10;1
Teddy	Td	RD	8;4
Tony	Tn	RD	9;0

References

Allen, J. B. (1996). Harvey Fletcher's role in the creation of communication acoustics. *Journal of the Acoustical Society of America*, 99(4):1825–1839.

Allen, J. B. (2005a). *Articulation and Intelligibility.* Morgan and Claypool, 3401 Buckskin Trail, LaPorte, CO 80535.

Allen, J. B. (2005b). Consonant recognition and the articulation index. *J. Acoust. Soc. Am.*, 117(4):2212–2223.

Allen, J. B. and Li, F. (2009). Speech perception and cochlear signal processing. *IEEE Signal Processing Magazine*, 26(4):73–77.

Bradley, L. and Bryant, P. E. (1978). Difficulties in auditory organisation as a possible cause of reading backwardness. *Nature*, 271(5647):746–747.

Bradley, L. and Bryant, P. E. (1983). Categorizing sounds and learning to read - a causal connection. *Nature*, 301(5899):419–421.

Brandt, J. and Rosen, J. (1980). Auditory phonemic perception in dyslexia: Categorical identification and discrimination of stop consonants. *Brain and Language*, 9(2):324–337. cited By 60.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory.* John Wiley and Sons, Inc.

Ellis, W. and Cramer, S. C. (1994). Learning disabilities: A national responsibility. report of the summit on learning disabilities. Report on learning disabilities ERIC Number: ED378783, ERIC, National Center for Learning Disabilities, 381 Park Ave., South, Suite 1420, New York, NY 10016.

Fletcher, H. (1995). Speech and hearing in communication. In Allen, J. B., editor, *The ASA edition of Speech and Hearing in Communication*, pages A1–A34,1–487. Acoustical Society of America, New York.

Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (2004). New nonsense syllables database: Analyses and preliminary asr experiments. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2749–2752.

Hanford, E. (2018). Hard words: Why aren't kids being taught to read? American Public Media: APM Reports.

Hazan, V. and Adlard, A. (1998). Speech perception in children with specific reading difficulties (dyslexia). *Quarterly Journal of Experimental Psychology: Section A*, 51(1):153 – 177.

Hazan, V., Messaoud-Galusi, S., Rosen, S., Nouwens, S., and Shakespeare, B. (2009). Speech perception abilities of adults with dyslexia: Is there any evidence for a true deficit? *Journal of Speech, Language, and Hearing Research*, 52:1510 – 1529.

Hazan, V., Messaoud-Galusi, S., Rosen, S., Schlauch, R., and Wright, B. (2013). The effect of talker and intonation variability on speech perception in noise in children with dyslexia. *Journal of Speech, Language, and Hearing Research*, 56(1):44 – 62.

Joanisse, M. F., Manis, F. R., Keating, P., and Seidenberg, M. S. (2000). Language deficits in dyslexic children: Speech perception, phonology and morphology. *Journal of Experimental Child Psychology*, 77:30–60.

Johnson, C. J., Kubalanza, M. C., Allen, J., Scheidiger, C., Biller, M., and Buie, J. (2015). Speech perception and reading disabilities in individual children. In *Seminar Presented at the American Speech-Language-Hearing Association*. Denver, Colorado.

Li, F. (2009). *Perceptual Cues of Consonant Sounds and Impact of Sensorineural Hearing Loss on Speech Perception*. PhD thesis, University of Illinois at Urbana-Champaign.

Li, F. and Allen, J. B. (2011). Manipulation of consonants in natural speech. *IEEE Transactions on Audio, Speech and Language Processing*, 19(3):496–504.

Li, F., Menon, A., and Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America*, 127(4):2599–2610.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358.

Lu, J. (Jul, 2018). *Speech perception in children with reading disabilities*. PhD thesis, University of Illinois at Urbana-Champaign.

Manis, F. R., McBride-Chang, C., Seidenberg, M. S., Keating, P., Doi, L. M., Munson, B., and Petersen, A. (1997). Are speech perception deficits associated with developmental dyslexia. *Journal of Experimental Child Psychology*, 66:211–235.

Messaoud-Galusi, S., Hazan, V., and Rosen, S. (2011). Investigating speech perception in children with dyslexia: Is there evidence of a consistent deficit in individuals?. *Journal of Speech, Language, and Hearing Research*, 54(6):1682 – 1701.

Miller, G., Heise, G., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41(5):329–335.

Miller, G. A. and Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. *Journal of the Acoustical Society of America*, 27(2):338–352.

NICHD/NRP (2000a). Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Technical Report Tech. Rep. No. 00-4769, NIH, Washington DC.

NICHD/NRP (2000b). Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Report of the subgroups. Technical Report Tech. Rep. No. 00-4754, NIH, Washington DC.

Phatak, S. and Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *Journal of the Acoustical Society of America*, 121(4):2312–26.

Phatak, S., Lovitt, A., and Allen, J. B. (2008). Consonant confusions in white noise. *J. Acoust. Soc. Am.*, 124(2):1220–33.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4):523.

Singh, R. and Allen, J. B. (2012). The influence of stop consonants' perceptual features on the Articulation Index model. *Journal of the Acoustical Society of America*, 131(4):3051–3068.

Tallal, P. (2000). The science of literacy: From the laboratory to the classroom. *Proc. Nat. Acad. Sci.*, 97(6):2402–2406.

Torgesen, J. K. (2004). Learning disabilities: An historical and conceptual overview. In *Chapter 1: Learning about Learning Disabilities*, pages 3–40. Elsevier.

Toscano, J. and Allen, J. B. (2014). Across and within consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research*, 57:2293–2307.

Toscano, J. C. and McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, and Psychophysics*, 74(6):1284–1301.

Van Hirtum, T., Moncada-Torres, A., Ghesquière, P., and Wouters, J. (2019). Speech envelope enhancement instantaneously effaces atypical speech perception in dyslexia. *Ear and hearing*, 40(5):1242–1252.

White-Schwoch, T., Woodruff Carr, K., Thompson, E. C., Anderson, S., Nicol, T., Bradlow, A. R., Zecker, S. G., and Kraus, N. (2015). Auditory processing in noise: A preschool biomarker for literacy. *PLOS Biology*, 13(7):1–17.

Williams, C. J. and Jacobs, A. M. (2009). The impact of otitis media on cognitive and educational outcomes. *Medical Journal of Australia*, 191(S9):S69–S72.

Wong, B. (2011). *Learning about Learning Disabilities.* Elsevier.

Ziegler, J. C., Pech-Georgel, C., George, F., and Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Developmental Science*, 12:732–745.