Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 431–441
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Displaying prosodic text to enhance expressive oral reading

Rupal Patel *, Catherine McNab

Department of Speech Language Pathology and Audiology, Northeastern University, Boston, MA 02115, United States

Received 23 June 2010; received in revised form 17 November 2010; accepted 24 November 2010
Available online 3 December 2010

Abstract

This study assessed the effectiveness of software designed to facilitate expressive oral reading through text manipulations that convey prosody. The software presented stories in standard (S) and manipulated formats corresponding to variations in fundamental frequency (F), intensity (I), duration (D), and combined cues (C) indicating modulation of pitch, loudness and length, respectively. Ten early readers (mean age = 7.6 years) attended three sessions. During the first session, children read two stories in standard format to establish a baseline. The second session provided training and practice in the manipulated formats. In the third, post-training session, sections of each story were read in each condition (S, F, I, D, C in random order). Recordings were acoustically examined for changes in word duration, peak intensity and peak F0 from baseline to post-training. When provided with pitch cues (F), children increased utterance-wide peak F0 range (mean = 34.5 Hz) and absolute peak F0 for accented words. Pitch cues were more effective in isolation (F) than in combination (C). Although Condition I elicited increased intensity of salient words, Conditions S and D had minimal impact on prosodic variation. Findings suggest that textual manipulations conveying prosody can be readily learned by children to improve reading expressivity.

Keywords: Children; Oral reading; Prosody; Reading software; Expressive reading

1. Introduction

Reading aloud is a motor, cognitive, and linguistic feat that takes years to learn and master. Early readers must not only decode the grapheme (letter) to phoneme (sound unit) sequence, they must also integrate this decoded message with syntactic and semantic information and then coordinate their speech musculature to produce the correct sounds in sequence. Given the complexity of this task, it is not surprising that even when children have developed adequate decoding and sight word recognition abilities, some continue to have difficulty reading aloud in a fluent, expressive manner. In fact, 44% of 4th grade students in the United States cannot fluently read grade-level stories (NAEP, 1995).

Speed and accuracy have traditionally been the focus of teaching children to become fluent readers. However, recent reports by the National Reading Panel (National Institute of Child Health and Human Development, 2000) and National Assessment of Educational Progress (NAEP, 1995) have expanded the definition of fluency to include "proper expression" and "ease or naturalness of reading." Reading with expression is now more widely accepted as an integral component of reading fluency (Fuchs et al., 2001; Hudson et al., 2005; Pinnell et al., 1995; Samuels, 1988; Schreiber, 1980, 1991; Stanovich, 1991). The extent and direction of the relationship between reading with expression and reading comprehension, however, remains unclear. There are currently two primary theories, the first which suggests that reading with expression mediates (or at least partially facilitates) reading comprehension, while the second poses that reading comprehension predicts reading with expression (Kuhn and Stahl, 2003; Miller and

* Corresponding author. Address: Department of Speech-Language Pathology and Audiology, Northeastern University, 360 Huntington Ave., 102 Forsyth Building, Boston, MA 02115, United States. Tel.: +1 617 373 5842; fax: +1 617 373 2249.
E-mail address: r.patel@neu.edu (R. Patel).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.11.007

Schwanenflugel, 2006; Schwanenflugel et al., 2004; Young and Bowers, 1995). Whatever the exact nature of this relationship, many early readers have difficulty reading aloud with expression. To address this issue, we designed a set of visual cues to convey intended prosodic variation in written text. The purpose of the present study was to explicitly assess the acoustic impact of these text manipulations on prosodic oral reading. We hypothesized that increasing awareness of prosody when reading aloud would improve reading fluency and may ultimately impact comprehension.

Prosody refers to stress, rhythm, intonation, and pause structure in speech and serves a wide range of linguistic and affective functions (Bolinger, 1961, 1989; Brewster, 1989; Crystal, 1979; Cutler et al., 1997; Lehiste, 1970; Shattuck-Hufnagel and Turk, 1996; Williams and Stevens, 1972). Reading aloud with expression or appropriate prosody includes placing emphasis on salient words within an utterance, pausing at the meaningful junctures, and varying intonation to convey communicative intent, which can be achieved by modulating fundamental frequency (F0) (perceived as pitch), intensity (perceived as loudness) and duration (perceived as length). Despite young children's abilities to control and use prosody (Bolinger, 1961; Brewster, 1989; Furrow, 1984; Gilbert and Robb, 1996; Lind and Wermke, 2002; Locke, 1993; Protopapas and Eimas, 1997; Werker et al., 1994), mastery of these skills continues to develop over much of early childhood and into adolescence (Cruttenden, 1985; Cutler and Swinney, 1987; Grigos and Patel, 2007; Snow, 1994, 1998; Stathopoulos and Sapienza, 1997; Wells et al., 2004). Thus, early readers may be acquiring skills in reading aloud while simultaneously refining their ability to control prosodic cues.

It is well understood that prosody plays a pivotal role in segmenting oral language into meaningful syntactic phrases (Bolinger, 1989; Cutler et al., 1997; Cromer, 1970; Gibson and Levin, 1975; Morgan and Demuth, 1996). In fact, children are more reliant than adults on prosody for speech segmentation (Schrieber, 1987) and their oral language processing is highly influenced by inappropriate prosody (Read and Schrieber, 1982). Early readers are then presented with a significant challenge when introduced to written text, which contains only sparse cues as to the underlying prosodic variation. In English, punctuation does not reliably mark phrase units or appropriate pause structure (Schreiber, 1991; Chafe, 1988). For example, commas do not always necessitate a pause, and questions do not always necessitate rising intonation (Bolinger, 1989; Chafe, 1988). With these unreliable or absent textual cues, early readers must infer the appropriate prosodic mapping to accurately convey the author's intended meaning (Miller and Schwanenflugel, 2006; Whalley and Hansen, 2006). It has been assumed that once children develop proficient decoding and word recognition skills, they will automatically transfer the melodic aspects of speech to written text (Carver, 1993; Hoover and Gough, 1990; LaBerge and Samuels, 1974). Yet many children (and even some adults) sound deliberate and expressionless, reading aloud in a word-by-word manner with insufficient prosodic variation even when they may have adequate word decoding skills (National Institute of Child Health and Human Development, 2000; Kuhn and Stahl, 2003; Miller and Schwanenflugel, 2006; Schwanenflugel et al., 2004; Read and Schrieber, 1982; Schrieber, 1987; Cowie et al., 2002; Dowhower, 1987; Herman, 1985). Perhaps if written text provided richer information about prosodic targets, early readers would have the scaffolding to learn and generate more appropriate prosody.

Previous attempts to provide visual prosodic cues within text have been limited to manipulations of spacing, punctuation, font and case, due in part to constraints imposed by traditional typesetting practices. Some researchers have recommended formatting text to display intra-sentence phrasal boundaries to facilitate chunking of text into meaningful units (Cromer, 1970; Levasseur et al., 2006; O'Shea and Sindelar, 1983). Others have suggested manipulating punctuation (e.g. My friend? My friend! My friend.) and font case (e.g. I like SOME of my relatives. vs. I like some of MY relatives) to practice modulating intonation (Blevins, 2001). Advances in desktop publishing software (e.g. Adobe PageMaker, Corel Ventura, QuarkXPress, Serif PagePlus) now enable the design of novel text orientation schemes for conveying a broad range of prosodic variations. Further research is required to determine whether text manipulation to relative word length, vocal loudness and pitch modulation could be easily learned by young readers and whether these cues would result in more expressive oral reading.

Toward this end, we designed an oral reading software program (ReadN'Karaoke) which translates a fluent adult reader's F0, intensity and duration variation into explicit visual cues embedded within text.[1] The software design includes novel text manipulation algorithms and leverages current evidence-based fluency instruction methods and principles. For example, repeated reading was incorporated by allowing children to record multiple reading attempts of each sentence during the training session. Guided oral reading, shown to be more effective in improving oral reading fluency compared to silent reading (National Institute of Child Health and Human Development, 2000; Kuhn and Stahl, 2003; Dowhower, 1987), was also utilized, enabling readers to hear an adult model prior to reading each sentence during the training session. Although other computerized reading programs (Read Naturally; Readers Theater) provide exercises to improve reading fluency through guided oral reading, they have not addressed expressivity through visual text manipulations.

The ReadN'Karaoke software displays text in standard and several manipulated formats to allow for comparisons across text conditions. The manipulated formats indicate

[1] Note that in this implementation text manipulations were based on only one fluent adult reader. In theory, however, renderings could be made for an indefinite number of readers or over a set of fluent readers. The current approach provided a scalable, data-driven option for semi-automated renderings within a relatively theory independent framework.

variations in fundamental frequency (F0), intensity level, and duration, either independently or in combination. We hypothesized that providing readers with explicit visual cues of the target prosody would facilitate appropriate modulation of these cues when reading aloud. The present usability study assessed the effectiveness of displaying text manipulations that convey prosody on children's variation of the corresponding acoustic cue(s). Acoustic analyses of oral reading samples allowed for comparison between manipulated and standard formats. Utterance and word level changes in acoustic cues were examined to determine whether text manipulations impacted overall prosodic variation as well as more fine-grained differences in modulation within utterances.

2. Method

2.1. Participants and setting

Ten typically developing children ages 6–9 years old (2M, 8F; $M = 7.6$ years), all native speakers of American English, were recruited to participate. Each child served as his or her own control in this repeated measures design, allowing for examination of the impact of text manipulation for children across a range of ages and reading abilities. Given individual differences in reading ability and prosodic maturation, each child's baseline reading expressiveness was obtained through direct measurement of prosodic variation while reading the two experimental stories during the first (baseline) session. All children passed the Clinical Evaluation of Language Fundamentals-4 Screening Test (Wiig et al., 2004) and had hearing thresholds that fell at or below 25 dB for 500, 1000, 2000, and 4000 Hz tones. Additionally, parental report indicated that none of the children had reading, speech-language or hearing problems nor did they receive any extra help in school (e.g., special education services). Nine children were recorded in a quiet room in their home, while one child was recorded in a sound treated booth. To minimize the impact of familiarity or practice effects, prior to data collection the experimenter ensured that each child had not been previously exposed to the experimental stories.

2.2. Materials and apparatus

The experimental stimuli consisted of two age and grade appropriate stories each comprised of 50 sentences. Actual children's storybooks with illustrated images were used in lieu of standardized passages to emulate the natural reading experience. Careful attention was given to selecting storybooks without rhyming words or repetition. Both experimental stories portrayed a mouse as the main character, were approximately equal in the number of picture frames and sentences, and included visually appealing graphics. Story texts were modified from their original published versions to equate the number of sentences between stories and to ensure that sentence length did not exceed 10 words. These constraints were imposed to accommodate word and character spacing in the interface and to minimize cognitive and physiological complexity for young children. Additionally, to elicit expressive reading, texts were modified to include declarative sentences, quotatives, *wh* questions, yes–no questions, adjectival phrases, and phrase boundaries – linguistic features which are known to require distinct prosodic reading in adults (Chafe, 1988; Cooper and Paccia-Cooper, 1980). Ten sentences from an additional children's story were used for training and demonstration purposes. Copyright permission for using the images and story content were obtained from the respective publishers.

The ReadN'Karaoke software was implemented in Java and run on an Apple laptop. The software displayed standard and manipulated text formats and recorded oral reading samples. Manipulated text formats were generated semi-automatically[2] based on recordings of a fluent adult reader (second author). The graphical interface (Fig. 1) included three panels, which displayed the story image, corresponding text, and a control panel of functions (i.e., play the pre-recorded adult model used to render the visual cues, record and replay own production, and navigate backward and forward in the story). The text could be displayed in one of five formats: standard text (S), duration manipulated (D), intensity level manipulated (I), F0 manipulated (F), or a combination of duration, intensity level, and F0 manipulations (C) (see Fig. 2). Spacing between characters indicated word duration and spacing between words signaled pause duration. The acoustic signal was divided into three discrete intensity levels, which were mapped to three shades of font color: black (highest intensity level), grey, and light grey. F0 variations were indicated by fitting text to the F0 contour of the adult model's productions (similar to Bolinger, 1989; Ladd, 2008).

Audio recordings were made directly to the laptop computer using a unidirectional head-mounted microphone (Shure, SM10A) at a sampling rate of 44,100 Hz with 16 bit linear quantization. The microphone was placed 1 in. from the corner of each child's mouth. Microphone to mouth distance and recording settings were maintained throughout each recording session and every effort was made to replicate the settings for subsequent sessions.

[2] Manipulated text formats were generated through a multi-staged process. Briefly, recordings of a fluent reader were manually annotated to demarcate the beginning and end of each word within each sentence (inter-labeler agreement, $r = 0.989$). For each sentence, pitch and intensity tiers were extracted using Praat (Boersma et al., 2007) and stored along with the word boundaries and sound recordings as data files. Pitch tracking errors were corrected by changing the parameters of the automated tracker and/or by manual adjustment. The data files were then used by the ReadN'Karaoke software to dynamically generate the manipulated texts using a set of pre-programmed rules to convert the acoustic data into the pixel layout of the text.

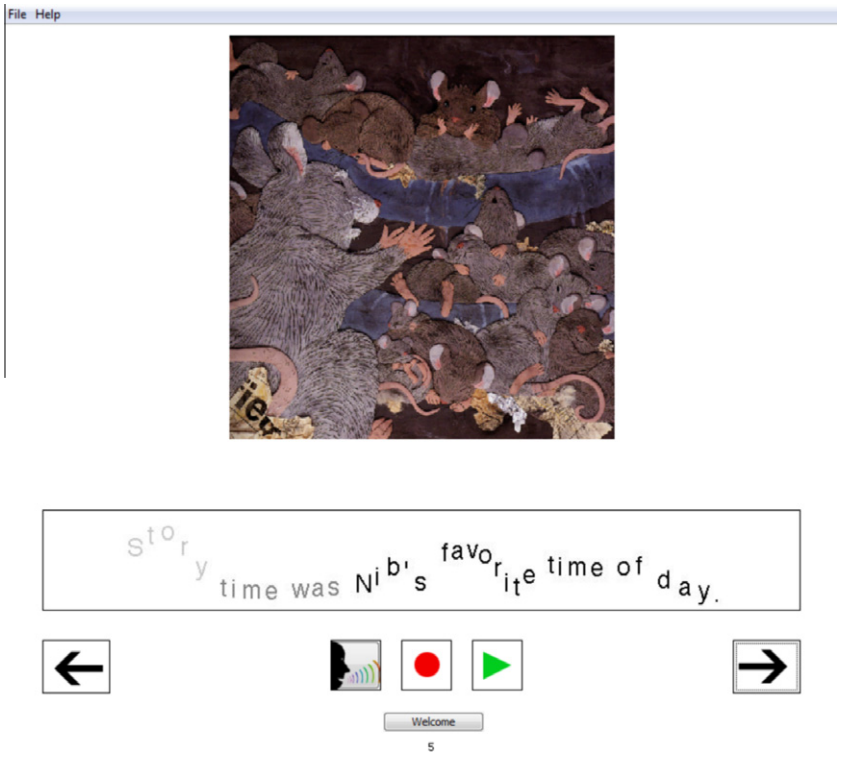

Fig. 1. Screenshot of the ReadN'Karaoke interface with text displayed in combination manipulated format.

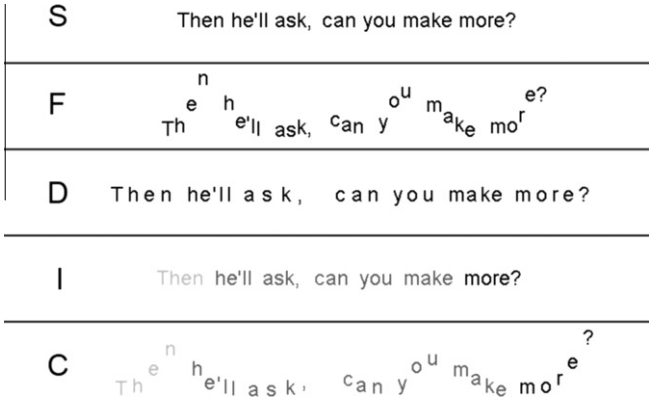

Fig. 2. Samples of the five text conditions: Standard un-manipulated format (S), frequency manipulated (F), intensity manipulated (I) and duration manipulated (D).

2.3. Procedures

Data collection took place over three sessions (baseline, training and post-training) each lasting approximately 45 min to an hour. Short breaks were offered regularly and provided upon request to minimize fatigue. During the baseline session, children were recorded reading the two experimental stories (A and B) presented in standard text format (S). Baseline recordings provided a direct measurement of participants' prosodic variation before training. The training session was conducted within 2 weeks of the baseline session (average intersession interval = 6.7 days).

During the training session, children were introduced to each manipulated text format and had opportunities to practice reading aloud in each text condition. A training story of 10 sentences was used to practice all five formats in the following order: S, D, I, F, and C. For each cue, the experimenter (second author) used a pre-scripted list of instructions to explain how to interpret the visual text. Children then listened to a pre-recorded adult model and were recorded reading aloud with that particular cue. They then listened to their own recording, received corrective feedback from the experimenter as needed, and were recorded reading the sentence again. At the end of the training session, a brief comprehension probe was verbally administered to ensure that children understood the cues in each manipulated condition. Children were required to answer all questions correctly; all passed the comprehension quiz on their first attempt with no re-explanation of cues needed. Each child returned within 2 weeks of completing the training session to complete the post-training session (average intersession interval = 7.7 days).

During the post-training session, the experimenter reviewed each of the five text conditions using two pages per text condition from the training story. This review followed the same instructions as the initial training session. The experimental session proceeded once the child demonstrated that they understood what each text manipulation was indicating. Each child served as his/her own control in that the he/she read the same experimental stories (A and B) as those used during the first session. Note that in the post-training session, as in the first session, playback

functions of the adult model and the child's own recording were disabled. Thus children had to rely on the visual cues alone. The percentage change in each acoustic variable (F0, intensity, and word duration) from baseline to post-training served as the dependant measures. In the post-training phase, the two experimental stories were divided into five sections corresponding to the five text conditions (10 sentences per section). Thus for each text condition, direct comparisons of the change in acoustic features from baseline to post-training could be made for a subset of 10 sentences from each experimental story for a total of 20 tokens per condition. The S condition was always presented first and the C condition was always presented last with the order of presentation of the remaining conditions (D, I, and F) randomized across children. This ensured that although different story sections had varying numbers of each linguistic sentence type and sentences length, across children, sections of each story were read in a variety of conditions. Additionally, the order of the stories (A-first vs. B-first) was counterbalanced across children. While children received encouraging verbal praise, they were not provided with any corrective feedback during the post-training phase. At the end of the session, a usability survey was verbally administered to help guide future modifications and improvements of the software and the overall experimental protocol.

2.4. Acoustic analysis

A total of 2000 recordings (50 sentences $*$ 2 stories $*$ 10 participants $*$ 2 sessions) were acoustically analyzed to examine the effect of text manipulations on children's oral reading samples. Note that direct comparison of the same sentences could be made since the experimental stories were read in standard format during baseline, and sections of each story were re-read during the post-training session in each of the five text conditions. Thus, acoustic changes associated with the standard condition captured practice effects associated with repeated reading, while changes in the other text conditions also included the effect associated with text manipulations.

The Praat speech analysis software package (Boersma et al., 2007) was used to calculate the change in peak word F0, peak word intensity level, and word duration from baseline to post-training for each condition. Sentences produced during training were not analyzed. For each utterance, Praat text grids were used to mark the beginning and end of each word using a combination of visual spectrographic analysis and auditory confirmation ($r = 0.968$ inter-labeler reliability for marking the beginning and end of words for 10% of the data sample). Words with nuclear and non-nuclear accents were also annotated within each text grid. Below are two examples of sample sentences in which the context of the story is used to define the nuclear (indicated in bold) and non-nuclear accented words (indicated in italics).

(1) I **do** *work*, said Frederick.
(2) The **mice** told *silly* stories.

Once the recordings were annotated, Praat was then used to generate a time series of relative intensity values (dB) and frequency values (Hz) for each word in each sentence (14,360 words across speakers, sessions and stories). A customized software routine operated on the Praat-generated F0 and intensity values and the manually annotated text grids to calculate peak F0, peak intensity level, and duration of each word in each sentence. The highest nominal frequency or intensity point within a labeled segment was selected as the peak F0 or peak intensity level, respectively. Word duration was calculated as the length of the labeled segment. Irregularities in F0 estimates (e.g., octave jumps, pitch breaks, glottal fry) were flagged using an automatic routine and through additional visual inspection of each acoustic file (22.2% of the total 14,360 words). Values that could not be verified auditorily were manually re-calculated (approximately 20% of total number of flagged values). Words produced with glottal fry were not included in the analysis and removed from both the baseline and post-training dataset. Some acoustic samples also required further examination due to reading errors. These errors were classified as repetitions, omissions or substitutions. When repetitions (sound, word, or phrase) or self-corrections were observed (7.85% of total 2000 utterances), the child's second (correct) attempt was used in analysis. Of the second attempts, seven were excluded due to either unnatural exaggeration (as judged by the first and second author) or whispering. When word omissions were observed, that word was removed from the corresponding baseline/post training dataset (33 omissions in total dataset). Word substitutions read consistently across baseline and post-training were included in the analysis (8 of 66 total substitutions) while substitutions read inconsistently across sessions were removed from both datasets.

For the purpose of this study, expressivity was operationally defined in terms of utterance (sentence) level modulation and word level precision. Utterance level analysis provided a measure of overall increases in prosodic variation, or the range of each prosodic cue, due to each text manipulation. For each utterance, the change from baseline to post-training in word duration range, peak intensity level range, and peak F0 range, was calculated. Peak F0 and peak intensity range were calculated as the difference between the highest and lowest word-level peak F0/intensity across the utterance. For each text condition (S, F, I, D, C), the change in each acoustic variable from baseline to post-training was averaged across all 20 sentences (10 from each story).

Word level analyses were undertaken to determine whether text manipulation aided in accurate modulation of acoustic cues on linguistically salient words within an utterance. The relative height of a nuclear or non-nuclear pitch accent was taken as the percentage change in peak F0 on the accented word compared to the average of the

peak F0s across unaccented words in the utterance. Similarly, relative differences in intensity level of the nuclear and non-nuclear accented words were taken as the ratio of the peak pressure on the accented word to the mean of the peak pressures across unaccented words. Changes from baseline to post-training in these relative differences were then compared. Additionally, the change in word duration of nuclear and non-nuclear accented words from baseline to post-training was calculated. Changes in each acoustic parameter for both utterance and word level measures were calculated per participant, and then averaged across participants for each text condition.

3. Results

A repeated measures experimental design was used to examine acoustic changes at the utterance and word levels from baseline to post-training in each text condition (S, D, I, F, and C). Using this design, each child served as his or her own control, allowing for examination of the impact of the manipulated text across a variety of age groups. At the utterance and word levels (nuclear and non-nuclear accent), separate repeated measures analyses were performed for each of the three dependent variables (F0, intensity level and duration) with one within subjects factor of text condition (five levels: S, F, I, D, and C) and one between subjects factor of participant. The F statistic was used to test the null hypothesis at $\alpha = 0.05$. Interactions between main effects were further examined using post hoc T-tests at a Bonferroni adjusted alpha level to account for multiple comparisons.

3.1. Utterance level analyses

A statistically significant main effect of change in peak F0 range within an utterance from baseline to post-training was found for text condition ($F = 5.96$, df $= 4971$; $p < 0.0001$). The between subjects factor of participant was not significant ($p = 0.134$). Condition F resulted in the greatest change in peak F0 range (mean change = 34.5 Hz) from baseline to post-training. At post-training, productions in all other text conditions, including S, trended toward increased peak F0 range (see Fig. 3). Post hoc contrasts revealed significant differences between F vs. S conditions ($p = 0.004$). There were no significant differences between condition F vs. I, D or C.

For changes in peak intensity level range, the main effect of text condition almost reached statistical significance ($F = 2.36$, df $= 4971$; $p = 0.051$), however, none of the post hoc contrasts reached statistical significance at the adjusted alpha level. Moreover, intensity level range increased to a greater extent in condition F (3.857 dB) compared to all other conditions including condition I (see Fig. 4). In fact, when provided with intensity cues, children's utterance-wide intensity level decreased by 0.33 dB from the baseline level to post-training. Once again, there were no significant differences between participants ($p = 0.611$).

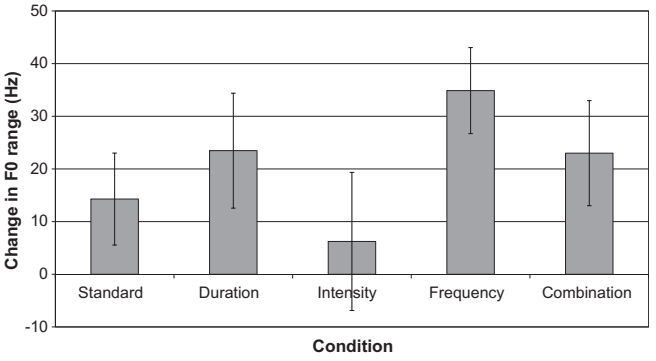

Fig. 3. Average change in peak word F0 range (Hz) from baseline to post-training sessions.

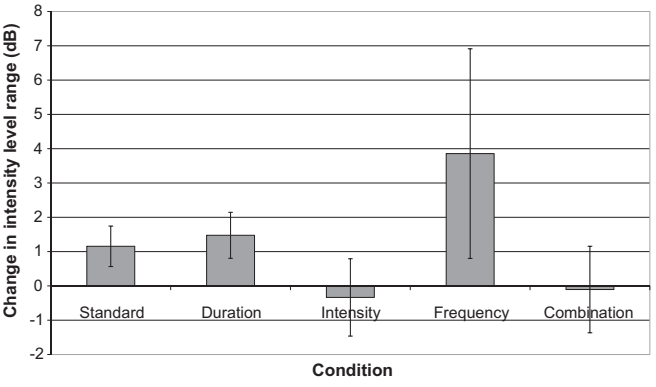

Fig. 4. Average change in peak word intensity level range (dB) from baseline to post-training.

A statistically significant main effect for change in word duration range was found for the within subjects factor of text condition ($F = 17.3$, df $= 4971$; $p < 0.0001$) but not for the between subjects factor of participant ($p = 0.247$). Once again, children increased word duration range to the greatest extent in condition F (0.146 s). Although word duration also increased in condition D (0.087 s), the extent of this increase was similar to condition C (0.086) (see Fig. 5). Post

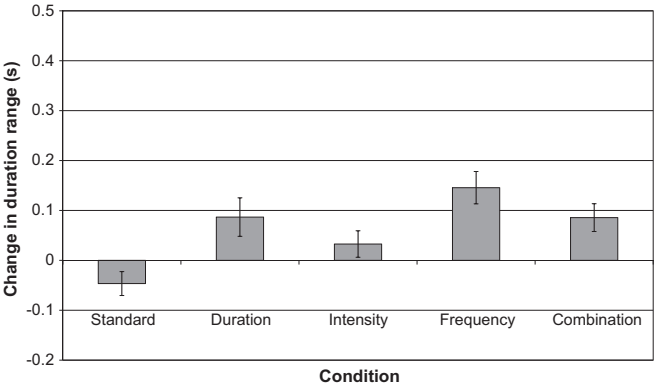

Fig. 5. Average change in word duration range(s) from baseline to post-training.

hoc contrasts revealed differences between D vs. S ($p < 0.001$) but not between other conditions. It should be noted, however, that observed duration changes in manipulated conditions were very small and may not represent perceptible or meaningful differences.

3.2. Word level analyses

3.2.1. Peak F0

A statistically significant main effect of the relative change in peak F0 of nuclear-accented words was found for text condition ($F = 3.24$, df = 4880; $p = 0.012$) but not for participant ($p = 0.139$). Conditions F (8.42%) and C (8.52%) resulted in the greatest increase from baseline to post-training. Increases in relative change in peak F0 were also observed in S, D, and I conditions, however, to a lesser extent than the C and F conditions (see Fig. 5). Post hoc contrasts revealed significant differences between F vs. S ($p < 0.0001$) and C vs. S ($p < 0.0001$) conditions.

Main effects of condition ($F = 0.682$, df = 4860; $p = 0.605$) and participant ($p = 0.415$) did not reach statistical significance across non-nuclear accented words. Relative increases in peak F0 for the non-nuclear accented words were greatest in condition F (4.22%) (see Fig. 6).

3.2.2. Intensity level

For relative changes in intensity level of nuclear-accented words, the main effect of condition ($F = 4.20$, df = 4905; $p = 0.002$) reached statistical significant while the between subject effect of participant ($p = 0.752$) did not. The intensity level of nuclear-accented words compared to the average of unaccented words increased to the greatest extent (2.02 dB) when text was presented in condition I (see Fig. 7). Post hoc contrasts revealed significant differences between I vs. S ($p < 0.0001$). Other contrasts, I vs. F ($p = 0.043$), I vs. C ($p = 0.032$) and I vs. D ($p = 0.023$) did not reach statistical significance at the adjusted alpha level.

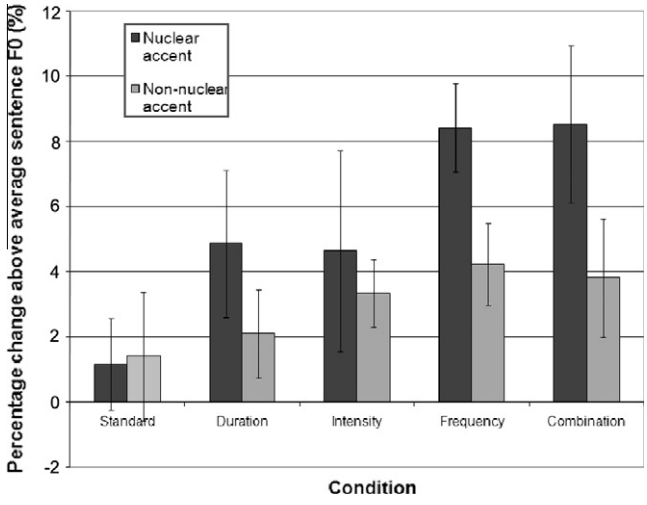

Fig. 6. Change in F0 above average sentence F0 from baseline to post-training.

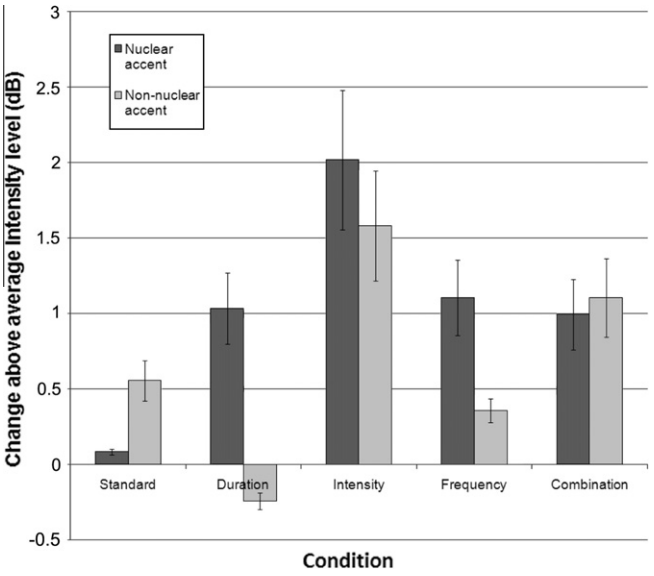

Fig. 7. Intensity level change above average sentence intensity level from baseline to post-training.

Although the main effect of condition reached statistical significance for non-nuclear accented words ($F = 4.16$, df = 4901; $p = 0.002$), post hoc contrasts did not reach statistical significance at the adjusted alpha level. Condition I resulted in the greatest increase above sentence average intensity level (1.58 dB) from baseline to post-training (see Fig. 7). Conditions S, F, and C also resulted in increased intensity level, while condition D resulted in a decrease. Additionally, there were no statistically significant differences between participants ($p = 0.313$).

3.2.3. Duration

A statistically significant main effect of change in word duration from baseline to post-training was found for text condition for both the nuclear-accented ($F = 28.3$, df = 4902; $p \leqslant 0.0001$) and non-nuclear accented ($F = 20.3$, df = 4893; $p \leqslant 0.0001$) words. The extent of increase in word duration from baseline to post-training of nuclear and non-nuclear accented words was similar in text conditions D, F and C (see Fig. 8). Across nuclear-accented words, post hoc contrasts revealed significant differences between D vs. S conditions ($p = 0.001$), F vs. S ($p < 0.001$), and C vs. S ($p < 0.001$). Similarly, for non-nuclear accented words, post hoc contrasts revealed significant differences between D vs. S conditions ($p = 0.007$), F vs. S ($p = 0.001$), and C vs. S ($p = 0.001$) but not between any other conditions. Once again there were no statistical differences between subjects (nuclear, $p = 0.471$; non-nuclear, $p = 0.293$).

3.3. Usability survey

Following the conclusion of the experimental protocol, an interview was conducted with each child to obtain feedback regarding the software and training methods.

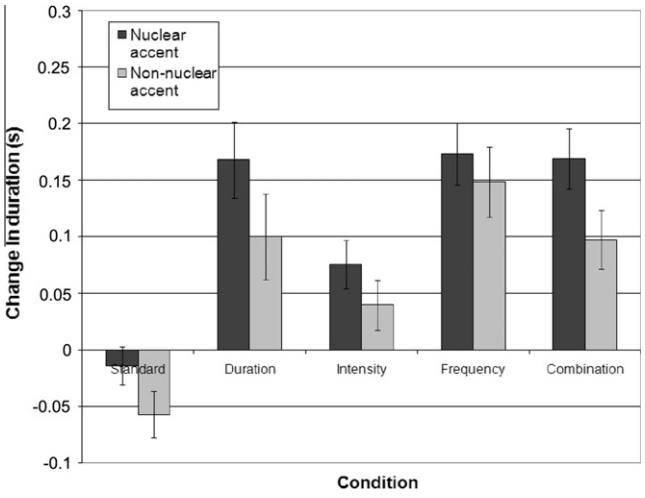

Fig. 8. Average change in word duration(s) from baseline to post-training.

Closed-ended questions allowed for comparisons across children, while open-ended questions allowed children to express individual opinions and offer suggestions.

In terms of reading ease, eight of the 10 children indicated that standard text was the easiest to read, while the combination condition was the hardest. Additionally, eight children indicated that reading the manipulated text on-screen was harder than reading a storybook. Specifically, several children noted instances in which it was difficult to determine word boundaries in condition F given that letters were placed along the intended F0 contour.

With regards to self-perception of their productions, nine of the 10 children felt they sounded "different" when reading the manipulated text compared to standard text. Six children felt that reading manipulated text helped them to understand what they were reading. One child noted, "It was easier to know what the animals' expressions were. Its hard to tell their expressions in a normal picture book." Another child observed that "you can see the emotion when they (the words) go up and down." A third child stated that, "...you don't know when to put your voice up and down (pointing to text in a regular book), loud or soft. I think it helped me learn how to read better."

4. Discussion

It is widely known that prosody is an essential component of reading fluency; however, written text lacks sufficient visual cues to guide the reader in how to read aloud (Hudson et al., 2005; Pinnell et al., 1995; Samuels, 1988; Schreiber, 1980; Stanovich, 1991). Thus, early readers must infer the prosodic structure to accurately convey meaning. ReadN'Karaoke was designed to address this challenge by explicitly displaying visual cues embedded within text to indicate prosodic variation. It was hypothesized that these text manipulations would help children modulate prosody when reading aloud. In the present usability study, acoustic analyses allowed for comparisons between manipulated (D, I, F, C) and standard (S) text formats on children's ability to modulate peak F0, peak intensity level, and word duration. Utterance level analyses provided an index of overall changes in prosodic variation resulting from each textual cue, while word level analyses indicated whether children could precisely apply these cues to linguistically salient words.

In this repeated measures design, the standard, unmanipulated text format controlled for practice effects and provided a baseline against which to compare the effects of manipulated conditions. Condition S did not result in significant differences from baseline to post-training along any acoustic parameter, thus familiarity with the text alone did not significantly improve prosodic variation. Additionally, at the utterance level, post-training samples read in condition S showed the least change along any acoustic feature compared to manipulated conditions with the exception of condition I. Furthermore, children marked nuclear and non-nuclear accented words with the least change in acoustic contrast in condition S compared to manipulated conditions. While repeated reading is a commonly used practice to improve reading fluency (Schreiber, 1991; Kuhn and Stahl, 2003; Morris and Slavin, 2002; Rasinski, 1990, 2003; Young et al., 1996), the present findings suggest that repeated reading alone may not be as beneficial as repeated reading with prosodic text manipulations.

In contrast to previous efforts aimed at providing visual prosodic cues (Cromer, 1970; Levasseur et al., 2006; O'Shea and Sindelar, 1983; Blevins, 2001), ReadN'Karaoke uses dynamic variations in spoken prosody to render text. These novel F0-, intensity-, and duration-based text manipulations were examined in isolation and in combination to identify the most effective subset of visual cues. Within each acoustic parameter, there were no statistically significant differences between participants. Results of both utterance and word level analyses suggest that presenting this sample of typically developing readers with prosodic text increased the range and precision of prosodic variation. Specifically, F0 manipulated text (F, indicated by text contour) had the greatest and most consistent effect, despite the fact that some children reported that it was the most challenging to read. Additionally, frequency cues in isolation (F) were generally more effective than presenting cues in combination (C), perhaps due to the increased cognitive burden of processing multiple cues simultaneously. Duration manipulations (D, indicated by spacing between characters and words) resulted in statistically significant yet relatively small changes in word length. Moreover, similar changes in word duration were noted when children were presented with F and C manipulated text. Although intensity manipulated text (I, indicated by text shading) did not influence overall utterance level intensity variation, these cues elicited increases in intensity level for nuclear and non-nuclear accented words within utterances. It should be noted, that while these word-level increases in intensity level were statistically significant, differences between conditions ranged from 1 to 2 dB indicating only subtle perceptual differences.

In the manipulated conditions, children tended to vary more than just the acoustic cue signaled by the text manipulation. For example, in condition F, children not only increased peak F0, but also increased peak intensity and duration. This redundancy may be explained in part by physiological interdependencies in modulating F0 and intensity level (Kent, 1997; Titze, 1994). Additionally, previous work suggests that young children use multiple cues to mark prosodic contrasts (Snow, 1998; Patel and Brayton, 2009; Patel and Grigos, 2006), like adults (Fry, 1955, 1958; Cooper et al., 1985; Traunmuller and Eriksson, 2000; Aylett and Turk, 2006).

Literature on the acquisition and development of prosodic control (Bolinger, 1961; Brewster, 1989; Lind and Wermke, 2002; Snow, 1994, 1998) may further account for the observed findings. Recall that full understanding and use of the various prosodic features continues to develop throughout childhood and into early adolescence (Young and Bowers, 1995; Cutler and Swinney, 1987; Wells et al., 2004; Bates, 1976). Specifically, modulating intensity has been shown to be a later developing ability, becoming adult-like at approximately 12–14 years of age (Stathopoulos and Sapienza, 1997). Thus, increased variability and reduced range of intensity level observed in the current dataset may be related to children's continued refinement of this motor skill beyond the period when children learn to read aloud. In contrast, acquisition of F0 control tends to be mastered between 5 and 8 years of age (Snow, 1994, 1998; Wells et al., 2004; Patel and Grigos, 2006), which may explain why this sample of 6–9 year olds were most proficient at applying F0 cues. While there were no significant between subject effects in the present dataset, it is possible that performance may differ across age groups in a larger sample. Although the present dataset did not allow for examination of differences among age groups, informal analyses revealed that older children increased F0 range to a greater extent than younger children (8–9 years; average increase of 50.9 Hz vs. 6–7 years; average increase of 23.6 Hz). This difference between age groups, however, was not evident for change in intensity level range (8–9 years; −1.28 dB vs. 6–7 years; 0.313 dB).

Alternatively, less pronounced changes in peak word intensity level and duration may be attributed to design decisions regarding the visual representation of these features. Although intensity varies along a continuous scale, it was discretized in the current software iteration to allow for simultaneous presentation of F0, intensity, and duration cues in condition C. Perhaps shading levels indicating intensity were not sufficiently distinct to capture children's attention. Similarly, within-word character spacing to indicate word duration may not have been exaggerated enough to prompt children to prolong salient words. Although word duration was the focus of the current analysis, the software also manipulated pause duration, which may have been more obvious to children. Hesitations and reading errors made it challenging to measure pause duration in this group of young readers.

It is encouraging that children in the present study were able to apply visual prosodic cues after a single, one hour training session. It is likely that with increased instruction and practice, children could further improve prosodic variation and precision when reading aloud using the ReadN'Karaoke software. This initial proof-of-concept evaluation provides promising evidence that visual text manipulations that convey pitch variation can be readily learned and applied by early readers. Determining whether this increased expressivity can enhance or bootstrap reading comprehension requires further inquiry.

5. Limitations and future directions

The present study is limited by the sample size ($N = 10$ participants) and number of tokens produced by each child in each text condition (20 sentences). Increasing the sample size would help generalize the findings and may allow for a stratified sample of children within narrow bands of age ranges and/or reading levels in order to avoid confounding effects of individual differences with developmental changes.

While the use of existing storybooks provided naturalistic contexts for reading aloud, it was difficult to control for sentence type and length. Further attention on developing novel stimuli that control for linguistic content and length may be beneficial (similar to Miller and Schwanenflugel, 2006). Such controls would enable analyses aimed not only at measuring improvements in expressiveness resulting from prosodic text but also to ultimately extend this work toward assessing the impact of prosodic text on reading comprehension.

Although significant changes in prosodic variation were observed in the text manipulated conditions, increasing practice time and standardizing the time between sessions may result in even more pronounced and longer lasting changes in expressivity. Implementing the study within a school or afterschool program may allow for greater standardization of the training session. Furthermore, a longitudinal design would enable assessment of changes in reading skill resulting from prosodic text manipulations, which were not possible to measure with the present cross-sectional design.

Future extensions of this work include exploration of alternative cue mappings to address readability of F0 manipulated text and to account for limited salience of the current intensity and duration cues. In addition to acoustic analyses, perceptual analyses of expressiveness and naturalness of recorded samples may provide qualitative impressions of the effectiveness of reading aloud with prosodic text. Exploring the application of the software to improve oral reading in individuals with reading disabilities, speech impairments and non-native speakers of English would also be warranted.

Acknowledgements

This study was conducted in the Communication Analysis and Design Laboratory in the Department of Speech

Language Pathology and Audiology at Northeastern University. This work was supported in part by funding from the American Speech and Hearing Association SPARC award (Students Preparing for Academic and Research Careers), the Northeastern University Provost Award, and the National Science Foundation (Grant IIS-0915527). The authors are grateful to Kevin Reilly, Michael Epstein and Timothy Mills for their time, effort, and suggestions and to Ghadeer Rahhal, who was instrumental in implementing the ReadN'Karaoke program and supplemental acoustic analysis software. Last, the authors thank the children and families who participated for their time, effort, and enthusiasm.

References

Aylett, M., Turk, A., 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. J. Acoust. Soc. Amer. 119, 3048–3058.

Bates, E., 1976. The acquisition of pragmatic competence. J. Child Lang. 1, 227–281.

Blevins, W., 2001. Building Fluency: Lessons and Strategies for Reading Success. Scholastic Inc., New York.

Boersma, P., Weenink, D., Praat, D., 2007. A system for doing phonetics by computer (Version 5.0.20) [Computer software]. Institute of Phonetic Sciences, Amsterdam.

Bolinger, D., 1961. Contrastive accent and contrastive stress. Language 37, 83–96.

Bolinger, D., 1989. Intonation and its Uses: Melody in Grammar and Discourse. Stanford University Press, Stanford, CA.

Brewster, K., 1989. The assessment of prosody. In: Grundy, K. (Ed.), Linguistics in Clinical Practice. Taylor & Francis, London, pp. 186–202.

Carver, R.P., 1993. Merging the simple view of reading with rauding theory. J. Reading Behavior 25, 439–455.

Chafe, W., 1988. Punctuation and the prosody of written language. Written Comm. 5, 396–426.

Cooper, W.E., Paccia-Cooper, J., 1980. Syntax and Speech. Harvard University Press, Cambridge, MA.

Cooper, W.E., Eady, S.J., Mueller, P.R., 1985. Acoustical aspects of contrastive stress in question–answer contexts. J. Acoust. Soc. Amer. 77, 2142–2156.

Cowie, R., Douglas-Cowie, E., Wichmann, A., 2002. Prosodic characteristics of skilled reading: fluency and expressiveness in 8–10 year old readers. Lang. Speech 45, 47–82.

Cromer, W., 1970. The difference model: a new explanation for some reading difficulties. J. Educat. Psychol. 61, 471–488.

Cruttenden, A., 1985. Intonation comprehension in 10-year-olds. J. Child Lang. 12, 643–661.

Crystal, D., 1979. Prosodic development. In: Fletcher, P., Garman, M. (Eds.), Language Acquisition. Cambridge University Press, Cambridge, pp. 174–197.

Cutler, A., Swinney, D.A., 1987. Prosody and the development of comprehension. J. Child Lang. 14, 145–147.

Cutler, A., Dahan, D., van Donselaar, W., 1997. Prosody in the comprehension of spoken language: a literature review. Lang. Speech 40, 141–201.

Dowhower, S.L., 1987. Effects of repeated reading on second-grade transitional reader's fluency and comprehension. Reading Res. Quart., 390–405.

Fry, D.B., 1955. Duration and intensity as physical correlates of linguistic stress. J. Acoust. Soc. Amer. 27, 765–768.

Fry, D.B., 1958. Experiments in the perception of stress. Lang. Speech 1, 126–152.

Fuchs, L.S., Fuchs, D., Hosp, M.K., Jenkins, J.R., 2001. Oral reading fluency as an indicator of reading competence: a theoretical empirical and historical analysis. Sci. Stud. Reading 5, 239–256.

Furrow, D., 1984. Young children's use of prosody. J. Child Lang. 11, 203–211.

Gibson, E.J., Levin, H., 1975. The Psychology of Reading. Massachusetts Institute of Technology Press, Cambridge, MA.

Gilbert, H., Robb, M., 1996. Vocal fundamental frequency characteristics of infant hunger cries: birth to 12 months. Internat. J. Pediatric Otorhinolaryngol. 34, 237–243.

Grigos, M., Patel, R., 2007. Articulator movement associated with the development of prosodic control in children. J. Speech Lang. Hear. Res. 50, 119–130.

Herman, P.A., 1985. The effect of repeated readings on reading rate speech pauses and word recognition accuracy. Reading Res. Quart. 20, 535–555.

Hoover, W.A., Gough, P.B., 1990. The simple view of reading. Reading Writing: Interdiscip. J. 2, 127–160.

Hudson, R.F., Lane, H.B., Pullen, P.C., 2005. Reading fluency assessment and instruction: what why and how?. Reading Teacher 58 702–714.

Kent, R.D., 1997. The Speech Sciences. Singular Publishing Group, San Diego, CA.

Kuhn, M., Stahl, S., 2003. Fluency: a review of developmental and remedial practices. J. Educat. Psychol. 95, 3–21.

LaBerge, D., Samuels, S.J., 1974. Toward a theory of automatic information processing in reading. Cognitive Psychol. 6, 293–323.

Ladd, D.R., 2008. Intonational Phonology, second ed. Cambridge University Press, Cambridge, UK.

Lehiste, I., 1970. Suprasegmentals. MIT Press, Cambridge, MA.

Levasseur, V.M., Macaruso, P., Palumbo, L.C., Shankweiler, D., 2006. Syntactically cued text facilitates oral reading fluency in developing readers. Appl. Psycholinguist. 27, 423–445.

Lind, K., Wermke, K., 2002. Development of the vocal fundamental frequency of spontaneous cries during the first 3 months. Internat. J. Pediatric Otorhinolaryngol. 64, 97–104.

Locke, J., 1993. The Child's Path to Spoken Language. Harvard University Press, Cambridge, MA.

Miller, J., Schwanenflugel, P.J., 2006. Prosody of syntactically complex sentences in the oral reading of young children. J. Educat. Psychol. 98, 839–843.

Morgan, J., Demuth, K., 1996. Signal to syntax: an overview. In: Morgan, J., Demuth, K. (Eds.), Signal to Syntax: Bootstrapping from Speech to Grammar to Early Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 1–24.

Morris, D., Slavin, R.E., 2002. Every Child Reading. Allyn & Bacon, Boston, MA.

National Assessment of Educational Progress (NAEP), 1995. Listening to Children Read Aloud: Oral Fluency. <http://www.nces.ed.gov/pubs95/web/95762.asp> (retrieved 21.07.08).

National Institute of Child Health and Human Development, 2000. 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 (NIH Publication No. 00-4769). US Government Printing Office, Washington, DC.

O'Shea, L.J., Sindelar, P.T., 1983. The effects of segmenting written discourse on the reading comprehension of low- and high-performance readers. Reading Res. Quart. 18, 458–465.

Patel, R., Brayton, J., 2009. Identifying prosodic contrasts in utterances produced by 4 7 and 11-year old children. J. Speech Lang. Hear. Res. 52, 790–802.

Patel, R., Grigos, M., 2006. Acoustic characterization of the question-statement contrast in 4 7 and 11-year old children. Speech Comm. 48, 1308–1318.

Pinnell, G.S., Pikulski, J.J., Wixson, K.K., Campbell, J.R., Gough, P.B., Beatty, A.S., 1995. Listening to children read aloud. Office of Educational Research and Improvement, US Department of Education, Washington, DC.

Protopapas, A., Eimas, P.D., 1997. Perceptual differences in infant cries revealed by modifications of acoustic features. J. Acoust. Soc. Amer. 102, 3723–3734.

Rasinski, T.V., 1990. Effects of repeated reading and listening while reading on reading fluency. J. Educat. Res. 83, 147–150.

Rasinski, T.V., 2003. The Fluent Reader: Oral Reading Strategies for Building Word Recognition Fluency and Comprehension. Scholastic, New York.

Read, C., Schrieber, P.A., 1982. Why short subjects are harder to find than long ones? In: Wanner, E., Gleitman, L. (Eds.), Language Acquisition: The State of the Art. Cambridge University Press, Cambridge, UK, pp. 78–101.

Read Naturally, The Read Naturally Strategy. <http://www.readnaturally.com/approach/default.htm> (retrieved 15.07.08).

Readers Theater, Readers Theater Scripts and Plays. <http://www.teachingheart.net/readerstheater.htm> (retrieved 14.12.09).

Samuels, S.J., 1988. Decoding and automaticity: helping poor readers become automatic at word recognition. Reading Teacher 41, 756–760.

Schreiber, P.A., 1980. On the acquisition of reading fluency. J. Reading Behavior 7, 177–186.

Schreiber, P.A., 1991. Understanding prosody's role in reading acquisition. Theory Pract. 30, 158–164.

Schrieber, P.A., 1987. Prosody and structure in children's syntactic processing. In: Horowitz, R., Samuels, S.J. (Eds.), Comprehending Oral and Written Language. Academic Press, New York, pp. 243–270.

Schwanenflugel, P.J., Hamilton, A.M., Kuhn, M.R., Wisenbaker, J.M., Stahl, S.A., 2004. Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. J. Educat. Psychol. 96, 119–129.

Shattuck-Hufnagel, S., Turk, A.E., 1996. A prosody tutorial for investigators of auditory sentence processing. J. Psycholinguist. Res. 25, 193–247.

Snow, D., 1994. Phrase-final syllable lengthening and intonation in early child speech. J. Speech Lang. Hear. Res. 37, 831–840.

Snow, D., 1998. Children's imitation of intonation contours: are rising contours more difficult than falling ones? J. Speech Lang. Hear. Res. 41, 576–587.

Stanovich, K.E., 1991 (Word recognition: changing perspectives). In: Barr, R., Kamil, M.L., Mosenthal, P., Pearson, P.D. (Eds.), Handbook of Reading Research. Longman, New York, pp. 418–452.

Stathopoulos, E.T., Sapienza, C.M., 1997. Developmental changes in laryngeal and respiratory function with variations in sound pressure level. J. Speech Lang. Hear. Res. 40, 595–614.

Titze, I.R., 1994. The Principles of Voice Production. Prentice-Hall, Englewood Cliffs, NJ.

Traunmuller, H., Eriksson, A., 2000. Acoustic effects of variation in vocal effort by men women and children. J. Acoust. Soc. Amer. 107, 3438–3451.

Wells, B., Peppe, S., Goulandris, N., 2004. Intonation development from five to thirteen. J. Child Lang. 31, 749–778.

Werker, J., Pegg, J., McLeod, P., 1994. A cross-language investigation of infant preference for infant directed communication. Infant Behavior Develop. 17, 323–333.

Whalley, K., Hansen, J., 2006. The role of prosodic sensitivity in children's reading development. J. Res. Reading 29, 288–303.

Wiig, E.H., Secord, W.A., Semel, E., 2004. Clinical evaluation of language fundamentals-4 screening test. Pearson Assess. Inform..

Williams, C., Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. J. Acoust. Soc. Amer. 52, 1238–1250.

Young, A., Bowers, P.G., 1995. Individual difference and text difficulty determinants of reading fluency and expressiveness. J. Exp. Child Psychol. 60, 428–454.

Young, A., Bowers, P.G., MacKinnon, G.E., 1996. Effects of prosodic modeling and repeated reading on poor reader's comprehension. Appl. Psycholinguist. 17, 59–84.