The SSF Model for Designing the Modality of Instruction

Bruce L. Mann¹, Henry Schulz² And Jianping Cui³

May 7, 2009

SUMMARY
Researchers keep saying that there is no advice for designing instruction with sound. Yet the origin of SSF model turns 60 this year. This paper tested the SSF model, this time with 4th and 5th graders (n = 133) to determine if they would learn grammar as well or better from convergent temporal sound cues developed from the SSF model, or from temporal text cues. A pedagogical agent delivered the cues from either a stationary position floating above the content in the window, or moving across the content in the window and gesturing. Results showed that children in the convergent temporal sound cueing group out-performed those in the convergent temporal text group on both immediate and delayed posttests. Agent movement had no effect. The implication for instructional designers and interface developers is that convergent temporal sound from the SSF model can provide the necessary reading context for young children learning from multimedia.

THE SSF MODEL

This research addressed the ongoing call in the literature for purposeful advice about how audio should replace or enhance onscreen text (Barron & Kysilka 1993; Koroghlanian & Klein, 2004), should be more than narrated screen text (Bishop, Amankwatia & Cates, 2008), beyond word lists (Gyselinck, Jamet & Dubois, 2008), with the right speech for a pedagogical agent (Woo, 2008), should show a strong cognitive foundation (Mayer, 2001) and sustained learning effects over time (Mann, 1994, 1995a, 1997a; Segers, Verhoeven & Hulstein-Hendrikse, 2008) - toward designer sound for the computer system (Buxton, 1989). In practice however, instructional designers and interface developers use little or no audio, or use their favourite heuristic, that may or may not have any cognitive foundation. Most of these heuristics aim to compliment the visual instruction in some way. Their choice of heuristic nevertheless, belies an underlying belief in how people learn from media (see table 1). The structured sound function (SSF) is different (Mann, 2009a). The structured sound function (SSF) is different (Mann, 2009a). The SSF aims to enhance a certain aspect of the visual for a particular effect. Enhancing visuals for a particular effect is particularly important in disciplines such as Intellectual Property Law where audio can enhance the quality of a disputant’s online presentation thereby helping to balance the bargaining conditions between the parties (Mann, 2009b).

1 Professor, Faculty of Education, Memorial University, Canada
2 Associate Professor, Faculty of Education, Memorial University, Canada
3 Learning Design Instructor, Red Deer College, Canada
Table 1. A summary of design heuristics and corresponding underlying assumptions about how people learn from multimedia *

<table>
<thead>
<tr>
<th>Design Heuristic</th>
<th>Underlying Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Whatever works</td>
<td>The developer's intuition is the best</td>
</tr>
<tr>
<td>2. Designed by type</td>
<td>A taxonomy is followed</td>
</tr>
<tr>
<td>3. Favorite feature</td>
<td>The right technology can be selected</td>
</tr>
<tr>
<td>4. Favorite method</td>
<td>The right instructional method can be devised</td>
</tr>
<tr>
<td>5. Balanced input</td>
<td>Logogens &amp; imagens will make the necessary referential connections</td>
</tr>
<tr>
<td>6. Maximum impact</td>
<td>Auditory &amp; visual episodes will eventually change long term memory</td>
</tr>
<tr>
<td>7. Cognitive load, first</td>
<td>Learning can be assisted by germane load</td>
</tr>
<tr>
<td>8. Structured sound function</td>
<td>Learners focus their attention by watching details &amp; listening to gist</td>
</tr>
</tbody>
</table>


The structured sound function (SSF) model addresses a persistent problem in learning from multimedia for adults and children, namely that they will ignore or forget to read important instructions and feedback presented in text or other visual displays (Pettersson, 1990; Ragsdale, 1988; Reinking, 1986). "Many learners do not notice the option to read directions or will try to save time by skipping them" (Alessi & Trollip, 1991, p. 22). Feedback following an error is more interesting than feedback following a correct response (Ragsdale, 1988). “Although aesthetically pleasing, feedback provided in text will go unnoticed by the student” (Alessi & Trollip, 1991, p. 72). Student enjoyment of multimedia has been either uncorrelated or negatively correlated to learning outcome (Clark, 2001; Clark & Feldon, 2005).
Two more reasons why children ignore or forget to read instructions and feedback in multimedia are: insufficient mental articulation of multimedia instruction due to an under-developed phonological loop, and; their inability to adequately control their attention during multimedia learning. The attentional capacity is smaller in children than in adults. Piaget described this capacity as *centration*, where children are capable of focusing on only one aspect to the exclusion of all others (Musgrove & Musgrove, 2004). Children between 9 and 11 years old (4th and 5th graders) for example (the focus of this research), are in middle childhood, a point of transition between the concrete operations and formal operations stages of their development. During this transitional period, the executive control process that makes the child fully conscious of their problem solving abilities and allows him or her to relate past information to the present problem in a systematic way is still maturing (Case, 1985). Children at this stage of development tend to use a similar approach to solving tasks with the same underlying problem, known as *psychological structures of the whole* and have general mental entities that correspond to this structure (Case, 1985, p. 255). However differences in children age differences in this stage show a wider range of variability in task performance than age differences among adults (Cowan, Elliott, Saults, Morey, Mattox, Hismjatullina & Conway, 2005), which could account for mixed results in research on multimedia learning with different ages of school-aged children over the past 50 years, as described in the next section. Perhaps “Piaget underestimated the capabilities of elementary schools students”, with older elementary school children able to mentally separate and control variables, especially when they are given hints about the importance of controlling all variables except the one they are testing (Ormrod, 2003, p. 32). Further study is needed to determine how children best learn from the presence of a moving and gesturing agent.

Figure 1 shows the structured sound function (SSF) model for designing the modality of instruction (Mann, 2008a). The SSF model has five functions and three structures for placing sound with visual events (Mann, 2008). “Once design components like text, graphics and sound have been assigned functions, their roles can become fixed in the designer’s mind, regardless of advances in the technology” (Bishop, Amankwatia & Cates, 2008, p. 480). The *sound* can be an effect, some music or an utterance (human or computerized). It can be a prompt from a tape or digitized file, a talking coach or agent in a computer application or website, a voice-over IP, or pod cast. Given a purpose or function, a sound can alert, caution, warn, remind or direct the user to visual events displayed by a computer program or Internet site. The *visual event* can be a sketch, diagram, photo, static or moving image, animated gif or cartoon, a linked reference, a procedural facilitation, or a worked example. Visual events can be displayed on a computer screen of any size. A visual event can include formatted-, unformatted- or hypertext, handwritten text, typewritten text, and word-processed text, a static or animated photograph, pictograph, film or video. Other times the visual is a character in a video or graphic, such as a computerized tutor, coach or mentor (human likeness or cartoon), appearing as an avatar or agent (static or moving). The recommendation for instructional designers and interface developers is that only the gist is presented in sound while the details are left as visual events (Mann 2008a).
The SSF Model for Designing Modality

<table>
<thead>
<tr>
<th>&lt; Structuring the sound with a visual event &gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The goal:</strong> is convergent or divergent</td>
</tr>
<tr>
<td><strong>The constancy:</strong> is continuous or discontinuous</td>
</tr>
<tr>
<td><strong>The density:</strong> is massed, spaced or summarized</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>&lt; Giving the sound a function &gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A temporal prompt:</strong> that cues that counterpoints that dominates that undermines</td>
</tr>
<tr>
<td><strong>A point of view:</strong> objective, subjective, performer, political, socio-cultural</td>
</tr>
<tr>
<td><strong>A locale:</strong> real, imaginary</td>
</tr>
<tr>
<td><strong>An atmosphere,</strong> feeling, mood</td>
</tr>
<tr>
<td><strong>A character's:</strong> past, future, personality</td>
</tr>
</tbody>
</table>


The origin of the SSF model is Zuckerman’s (1949) instructional film sound categories, used with audiences to describe the effectiveness of narration in post-war instructional films. Mann (1988) added new functions and consolidated others, converting the descriptive classification of psychological effects into a model for designing the modality of instruction, in keeping with the view that the quality of an instructional design model is dependant on the model's *prescriptivity* (Martin & Briggs, 1986). The first computer application of the SSF model appeared in 1992. See Mann
(1992) for details. Applications of the SSF model have been successful; initially with undergraduate students learning to edit expressionistic video clips from raw footage (Mann, 1988), and later with graduate students learning instructional design from a computer tutorial (Mann, 1994). In both applications, continuous, convergent temporal sound was associated with important visual events to shift student attention. In the former study, pre-recorded instruction, navigational direction, hints, feedback and reminders were played from a videotape. In the latter study, pre-recorded instruction, navigational direction, hints, feedback and reminders were played from a digitized file in a spreadsheet program. More recent research with the SSF model showed good results with adults (Mann, 1995a, 1997a) and equivalent results with 7th graders (Adams, Mann & Schulz, 2006; Mann, Newhouse, Pagram, Campbell & Schulz, 2002).

ATTENTIONAL CONTROL THEORY

The theoretical foundation for the SSF model is the attentional control theory of multimedia learning (ACTML), a conceptual framework of working attention during learning that integrates the functions of Brainerd and Reyna's fuzzy-trace theory within the 4-component structure of Baddeley’s framework of working memory. Figure 2 illustrates the structure of working attention and learning process in the ACTML.

Learner's Executive Controller
accesses data from the senses, focuses attention to learn, sends for data from LTM, compensates for deficiencies and guides behaviour

Auditory System
Learner encodes gist directly to phonological store. Details indirectly through an articulatory loop

Visual System
Learner forms an interface between visual and spatial data in a visuo-spatial sketchpad

Learner’s Long-Term Memory
is gradually altered by the process

Learner reads on-screen text, a graphic or animation as a visual event

Learner listens to speech, music or sound effect that motivates, directs, instructs, hints, reminds or gives feedback

The theoretical foundation for the ACTML was adapted from two well-known streams of research that developed independently of one another: Baddeley’s revised model of working memory, and Brainerd and Reyna’s fuzzy-trace theory. Baddeley’s model is widely known in applied psychology (Gyselinck, Jamet & Dubois, 2008; Rende, Ramsberger & Miyake, 2002). The ACTML has modified the structure of Baddeley’s generic working attention framework to illustrate learning from media. Baddeley’s model (Baddeley, 2000, 2002; Baddeley & Andrade, 2000; Jefferies, Ralph & Baddeley, 2004) has also been called, more accurately perhaps, ‘working attention’ because performance seems dependent on the maintenance of attentional processes (Baddeley & Logie, 1999; Neath 2000). Brainerd and Reyna's fuzzy-trace theory (FTT) is also widely known in applied psychology. FTT provides a relevant approach to interpret memory performance (Brainerd & Reyna, 2002; Brainerd, Reyna, & Brandse, 1995; Reyna, 2000; Reyna & Brainerd, 1995; Reyna & Kiernan, 1994). During encoding of to-be-remembered instructional material, verbatim and gist traces are formed in parallel. A verbatim trace represents the item's surface form while a gist trace represents the semantic, relational, and elaborative properties of the instruction or feedback. Listeners attend to the gist or meaning (Hildyard & Olson, 1982; Penney, 1989; Brainerd & Reyna, 1990; Reyna, 1992), whereas readers attend to details or surface features of the text (Tannen, 1985), also called verbatim information (Martin & Briggs, 1986; Penney, 1989). Hildyard & Olson (1982) reported that performances were better for listening to highlights (gist) and reading the details (substance, verbatim). The terms “gist” and “verbatim” are used here in the psycholinguistic sense, in compliant with fuzzy-trace theory (FTT) (Brainerd & Reyna, 1993, 2001, 2004; Reyna & Brainerd, 1990, 1994; Wolfe, Reyna & Brainerd, 2005). Forgetting is assumed to be the gradual fragmentation of traces, which is higher for verbatim (as opposed to gist) representations. Therefore, people are more likely to rely on verbatim representations immediately after the to-be-remembered is presented, but shift to gist after a delay (Reyna & Kiernan, 1994). FTT posits that people retrieve verbatim and gist traces separately, with correct recollection being supported by both kinds of traces (Brainerd & Reyna, 2002). Figure 2 describes how learners process gist and verbatim instruction in different modalities differentially (Halliday, 1987; Penney, 1989; Higginbotham-Wheat, 1991).

Adults listening to a sound presented in multimedia will encode the gist directly into their phonological store, and encode the details indirectly through their articulatory loop. In young children however, unarticulated material is akin to extraneous cognitive load reported in adults (Kalyuga, Chandler & Sweller, 1999; Mann, 1997; Mayer, Heiser & Lonn, 2001; Sweller & Chandler, 1994). Unlike adults, young children are not fully capable of mentally articulating instructions and feedback presented in text. Their auditory memory consists of a phonological store without a phonological loop (Gathercole, Pickering, Ambridge & Wearing, 2004). Young students reading difficult or unfamiliar text from a web page for example, will articulate the sound of the words to “hear” themselves say the words, and may experience the common side-effect of a dry throat from sub-vocalizing the sound of the words or phrases by the inner voice to be heard by their inner ear. This reliance on reading context decreases as a function of reading development and ability (Goldsmith-Phillips, 1989; Swantes, 1991).

The inclusion of voice has been found to be a key element for agent-based learning (Baylor & Ryu, 2002). In separate trials, Atkinson (2002) using a human
digitized human voice with undergraduate students \((n = 50, n = 75)\), indicated better learning with the speaking agent, than the agent with text. Pedagogical agents are a type of animation (Baylor, 2000), cartoon-like characters that can be programmed to move around a computer screen gesturing, doing tricks, and talking to the participating student through a text-to-speech engine in either speech or printed bubble text, or both. Research on learning from pedagogical agents has explored several factors, including: the presence and movement of an agent, the agent persona, the modality accompanying the agent (speech and static, flashing or bubble text), and the number of agents on the screen. Agents are typically expected to maintain student attention to critical information on the computer interface (Dehn & van Mulken, 2000) and provide dynamically individualized scaffolding through educational websites (Callaway, Lester, Towns & Voerman, 1999; Moreno, Mayer & Lester, 2000). In this way, pedagogical agents are intended to enhance the learning experience.

At present however, most of the research on using animated pedagogical agents has been conducted with adult learners, mostly undergraduate psychology majors or pre-service teachers. Some which reported that the visual presence of a pedagogical agent neither enhanced nor impaired learning (Moreno, Mayer & Lester, 2000; Moreno, Mayer, Spires & Lester, 2001, experiment 4). A new study was needed with school-age students to focus on the prediction that they would learn better from animated pedagogical agents when critical information was presented as convergent temporal sound than temporal on-screen text.

The Present Study
The aim of the present study was to test convergent temporal sound from the SSF model. Unlike previous studies investigating whether sound can help focus attention on locations on a computer screen (Bonebright & Rees, 2008), our expectation was that 4th and 5th graders would learn grammar better with an agent speaking in convergent temporal speech cues, than those with an agent reading temporal text cues on the computer screen. Cueing can set the stage or serve as a signal for specific behaviours to take place (Burton, Moore & Magliaro, 2004). Convergent temporal sound is a audio cue that promotes convergent thinking by focusing the learner’s attention in a stepwise procedure toward a specific solution (Mann, 2006, 2008, 2009). During convergent goal setting the user is encouraged to use a variety of sources to bear on a problem (e.g., answer look-up) to produce an acceptable result. Selecting the goal and constancy for a sound cue for a visual event is the most important of the three structural components in the SSF model (Mann, 2006).

METHOD

Participants
Participants were 133 4th and 5th grade students (aged 9-12 years) enrolled at an urban public school that had a consent form that had been signed by a parent or guardian, from an initial group of 162. “Children at this level are eager to cooperate in experiments and are able to persist at them for long periods of time” (Case, 1985, p. 182). None of the participants had been introduced to singular and plural ownership or the proper usage of apostrophe prior to or during the experiment, and although they had been exposed to educational multimedia in the school setting, none of these children had seen an agent.
Materials

The multimedia treatments and tests used in this study were administered to participating students from the school’s WebCT course management system, and accessed from separate computer labs (on different floors of the building), unlike previous studies (Adams et al, 2006; Tabbers, 2002) where the treatments were administered consecutively in modified classrooms. The multimedia treatments in this study were administered by the random order generator in the Presentations Tool, and the tests by the random order generator in the Quiz Tool in WebCT. This helped to prevent participating students working together in the same lab from copying from one another. Furthermore, the sound-cued groups and text-cued groups accessed their respective treatments from separate computer labs within the school. This was a fairly normal situation for students in that school where classes of 4th and 5th grade student frequently shared the labs in this way to complete their assignments.

A Microsoft agent called “Genie the Magician” cued the visual events on a customized educational website, the files were uploaded to an independent server, to prevent accidental prior access to the content by the participating students. The agent character was implemented as an ‘advisor’ (Baylor & Ryu, 2003) flying-in from the top left hand corner of the screen to an initial position on the top left of the screen in physical proximity to the onscreen graphics and text boxes, in keeping with the spatial contiguity principle (Moreno & Mayer, 1999). Four distinct guided-tutorials were accessible from four separate Internet addresses.

In the on-screen text cues with no agent movement (NMT) condition, Genie flew in and remained motionless. Genie spoke as silent text bubble appearing above its head. But there was no gesturing to the text, no changing position on the screen, and no magic tricks.

In the on-screen text cues with agent movement (MT) condition (shown in Figure 3), Genie flew in and gestured to the text on the screen with its outstretched arm. The prompts appeared as text bubble in a text box, with identical information to the speech condition. The Genie’s movements in the MT condition were identical to those in the MS condition.

In the speech cues with agent movement (MS) condition the Genie flew in, immediately gestured to the text on the screen with its outstretched arm, and spoke speech cues in a synthesized voice about a future event (direction, instruction hint) or a past event (feedback or reminder). Periodically, Genie would scratch his head, do a magic trick, and fly to a different position on the screen. The agent was programmed to speak only the gist (not explain or narrate), leaving the details in text and illustrations.

In the speech cues with no agent movement (NMS) condition, the Genie flew in and thereafter remained motionless. The Genie in the NMS condition spoke the same temporal speech prompts as the Genie in the MS condition. But there was no gesturing to the text, no changing position on the screen, and no magic tricks.
The SSF Model for Designing Modality

Scripting of the agent’s temporal speech cues was developed with the primary aim of directing attention to the details in the onscreen text, graphic or animation on the screen, as suggested in the structured sound function (SSF) model (Mann, 2008).

![Figure 3. A non-example showing the talking agent cueing the participating student with a verbal caution about singular ownership](image)

Participating students were asked to discriminate the correct format of the possessive expression from the incorrect ones. Further, students were asked to transfer the knowledge of the correct usage of apostrophe to make up sentences according to different graphics and pictures. The structure and content of the program were held constant across the treatment conditions as the animation and modality were varied. The language on the website was child-directed and designed to assist their learning, as suggested by Tomasello and Brooks (1999). A literacy expert examined the content and judged it to be appropriate and interesting for 4th and 5th grade students. Following a
quality review, the website was revised to reduce clutter and reduce extraneous cognitive load, compatible with the coherence principle (Moreno & Mayer, 2000) and cognitive load theory (Pollock, Chandler & Sweller, 2002). Students recruited for pilot testing were excluded from the study.

Two of the school’s best-equipped computer labs were made available for the experiment. One lab was set up for students in the MT and NMT conditions to watch a pedagogical agent called “Genie”, and read the on-screen instruction and feedback involving the proper usage of apostrophe to show singular and plural ownership. The other computer lab was configured for students in the MS and NMS conditions to watch agent “Genie” and to listen to equivalent instruction and feedback through headphones. In both labs, and prior to administration of the treatment, the investigator used a big screen television connected to a computer to demonstrate how students should navigate the program.

**Instruments**

A single 12-item test on correct apostrophe usage was developed and pilot-tested in the WebCT quiz tool with 4 students from the target grades levels. Some items were revised, based on their feedback. The items were administered in each testing occasion (i.e., pretest, immediate post-test, and delayed post-test) by the random order generator in the WebCT quiz tool. The test was graded out of total of 30 possible points and checked by investigators. There were 10 multiple choice questions each worth 2 points that asked the participant to choose which sentence was correct, and 2 short-answer questions each worth 5 points that asked the participant to type a sentence showing plural ownership based on a picture. The four experimental conditions in the study were deemed equivalent. The probability that this read-then-type requirement might disadvantage speech-cued participants, as suggested recently (Segers, Verhoeven & Hulstijn-Hendrikse, 2008) was very low, for two reasons. First, speech-cued participants had their cognitive load shared between auditory and visual memory systems. Second, speech-cued participants were required to process only the gist portion by listening to speech-cues, and the details portion by reading the text.

**Design**

The study used a 3-factor repeated measures (pre-test, immediate post-test, delayed post-test) design with three independent between-subjects factors: agent modality (speech cues vs. text cues), agent animation (movement or no movement), and student grade level (grade 4 or grade 5). We included the delayed post-test in this study to test the content of long-term memory over time, consistent with the ACTML, which links focusing attention in the executive controller with long-term memory, compatible with the view (Sweller, 2004) that the main purpose of instruction is to build knowledge in long-term memory.

**Procedure**

The experiment was conducted in three scheduled sessions over a 10-week period. From the initial group of 162 students that were given consent forms to have signed by a parent or guardian, 133 4th and 5th grade students returned with signed consent forms took the pre-test of apostrophe usage and interpretation of rules. For each experimental
session, the teachers brought the participating students into the labs. Their names were coded and then anonymously and randomly assigned to one of the treatment groups, either: speech cues with agent movement (M, S), speech cues with no agent movement (NM, S), on-screen text cues with agent movement (M, T), or on-screen text cues with no agent movement (NM, T). In a separate session, 133 students with signed consent forms took the pre-test of apostrophe usage and interpretation of rules. Administration of the pre-test took about 20 minutes. Four weeks later, these participating students took the treatment and immediate post-test. Administration of treatment and immediate post-test took about 50 minutes. Six weeks later, the participating students took the delayed post-test, which lasted about 20 minutes.

Results

Pre-test. A preliminary analysis of variance (ANOVA) on the pre-test means of the two grade levels crossed with the four treatment groups indicated that they did not differ significantly in their prior knowledge, $F(7,129) = 1.19$ ($p = .312$). This suggested that the eight groups were similar with low prior knowledge of apostrophe usage, and that that knowledge was uniformly low. Low prior knowledge was a necessary pre-condition in this study, as explained earlier.

Posttests. Statistical imputations were conducted by multiple regression with a stochastic component on cases that showed a participating student had taken the treatment but were absent for a test to retain as much of the data as possible, in accordance with recommended procedures (Little & Rubin, 1990; Smits, Mellenbergh & Vorst, 2002). There were 11 missing cases on the pre-test and 4 on the delayed post-test, evenly distributed across the treatment groups (from 2 to 4 imputations per group on the pre-test, and from 0 to 2 on the delayed post-test). Missing retention scores were then replaced with the value predicted from multiple regression of the retention test on the pre- and post-tests based on the original complete data. To this replacement was added a random normal component with mean = 0 and standard deviation = standard error of regression. Smits, Mellenbergh, and Vorst (2002) found these procedures to be effective for missing data, particularly where listwise deletion would be inappropriate. Listwise deletion methods usually discard important data available on other variables, and argued to be less appropriate than stochastic conditional imputation (Little & Rubin, 1987; Peugh & Enders, 2004; Raaijmakers, 1999; Raymond & Roberts, 1987; Roth, 1994). The procedures used permitted analysis of a complete data set that reflected the expected characteristics of the scores (Little & Rubin, 1987; Peugh & Enders, 2004; Raaijmakers, 1999; Raymond & Roberts, 1987; Roth, 1994). This resulted in complete data for 133 participating students (see Table 2 for numbers in each group).

Table 2. Means on apostrophe test for the eight groups over the three testing occasions

<table>
<thead>
<tr>
<th>Agent an.</th>
<th>Modality Grade</th>
<th>n</th>
<th>Pretest $M$</th>
<th>$s$</th>
<th>Posttest $M$</th>
<th>$s$</th>
<th>Delayed test $M$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move.</td>
<td>Speech</td>
<td>4</td>
<td>11.94</td>
<td>5.27</td>
<td>15.80</td>
<td>5.39</td>
<td>13.29</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>11.92</td>
<td>6.71</td>
<td>16.90</td>
<td>8.74</td>
<td>18.00</td>
<td>6.99</td>
</tr>
<tr>
<td>Text</td>
<td></td>
<td>4</td>
<td>8.53</td>
<td>3.67</td>
<td>10.00</td>
<td>3.94</td>
<td>8.22</td>
<td>4.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>9.72</td>
<td>6.91</td>
<td>11.13</td>
<td>7.61</td>
<td>12.28</td>
<td>6.58</td>
</tr>
</tbody>
</table>
The design then, consisted of a within-subjects factor (the three repeated measurements: pre-, post-, and delayed test) crossed with the three between-subjects factors (agent animation, modality, and grade level). Treatment effects of animation and modality would be evidenced by interactions with the repeated measures. There was only one statistically significant interaction for the within-subjects analysis: the interaction of the repeated measures with modality, $F(2, 250) = 5.02 (p = .007)$. While significant, it should be noted that this interaction effect accounted for approximately 4% of the variance attributed to the within-subjects part of the analysis, and should therefore be described as a small to medium effect. There was a significant within-subjects main effect due to the repeated measures, $F(2, 250) = 22.29 (p = .000)$, and a significant between-subjects effect due to modality $F(1, 125) = 21.54 (p = .000)$. While these main effects were significant, they were not pursued directly as the interaction noted above was significant, and it included both of these factors.

Table 3. Apostrophe analysis of variance: repeated measures.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>$F$</th>
<th>$\eta^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No move. Speech</td>
<td>4</td>
<td>10.33</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.60</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10.44</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>10.33</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.60</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10.44</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>10.33</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.60</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10.44</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>10.33</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.60</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10.44</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>10.33</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.60</td>
<td>0.07</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10.44</td>
<td>0.07</td>
<td>0.007</td>
</tr>
</tbody>
</table>

The SSF Model for Designing Modality
The results show that the pattern of change from pre- to post- and delayed posttest was different for two modality conditions, Speech and Text. Since there were no more complex interactions involving modality, the eight groups were collapsed into two modality groups. The means of the two groups are depicted in Figure 3. The pattern for the Speech group appears distinctly different from that for the Text group: the Speech group means changed from 10.95 on the pretest to 15.73 on the posttest and 13.65 on the delayed posttest, whereas the corresponding Text group means changed from 8.80 to 10.60 and 9.56.
Simple main effects were tested for the three testing times within the Speech group and within the Text group to determine which of these differences were significant (using procedures outlined for repeated-measures analyses by Keselman and Keselman, 1993). There was a significant change over time for the Speech group accounting for over 27% of the variance associated with this group, $F(2, 130) = 24.51$ ($p = .000$). Further post hoc tests using the Bonferroni critical value of .017 for the three pair wise tests indicated that all three pairs of means differed significantly. The pretest and posttest means ($p = .000$) had a medium to large effect size of .76, the pretest and delayed posttest means ($p = .001$) had a medium effect size of .43, and the posttest and delayed posttest means ($p = .001$) had a small effect size of .31. Means for the speech group were highest on the posttest, followed by the delayed posttest, and then the pretest (see Figure 4). Thus, significant learning occurred for this group, and this learning maintained to the delayed posttest but with some loss.

The simple main effect results were also significant for the Text group but not as strong, accounting for only approximately 5% of the variance associated with this group on the three tests, $F(2, 132) = 3.27$ ($p = .041$). Post hoc tests using the critical value of .017 could not detect where these differences occurred, pretest and posttest means ($p = .030$), pretest and delayed posttest means ($p = .030$), posttest and delayed posttest means ($p = .268$), posttest and delayed posttest means ($p = .116$). The means for the Text group were more similar to one another than the means in the Speech group (see Figure 4). It is not as clear whether there actually was learning with the Text group, although some change did occur from pre- to post- and delayed post-testing.

Figure 4. Means on apostrophe test for the Speech and Text groups
Discussion

The aim of this research was to test a particular structured sound function called convergent temporal speech, with an animated agent character. Convergent temporal speech cueing has been studied more than any of the other structured sound functions from the structured sound functions (SSF) model of instructional design. Our expectation was confirmed, that cueing young children with convergent temporal sound from a talking agent would help them learn grammar. In general, the results for cueing with speech are in line with those in Segers, Verhoeven & Hulstein-Hendrikse (2008) who found a modality effect with 5th graders, and akin to results in previous modality research with adults (Chandler & Sweller, 1992; Mann et al, 2002; Mann, 1995a; Mann, 1997a; Mayer, 1997; Moreno & Mayer, 2000). Results on agent motion however, were not supported. Despite the apparent fascination with the agent or "little Dude", as he was called by some of the children, our expectations about them learning from agent movement was not supported.

Multimedia learners must be able to consciously focus their attention to bind the separate features of a stimulus such as colour, shape and words, into a unitary object (Matlin, 1994) or coherent episode (Baddeley, 2002). They must engage the appropriate attentive state by self-initiating the appropriate system of information processing (Borich & Tombari, 1995). This poses no problem with easy tasks or familiar items. The children implement automatic (Schneider & Shiffrin, 1977) or pre-attentive (Treisman, 1986) processing. Pre-attentive processing of easy tasks or familiar items can occur in parallel; that is they can handle two or more items at the same time. On difficult tasks or with unfamiliar items however, children must be able to consciously control their own mental processing (Schneider & Shiffrin, 1977) to focus attention (Treisman, 1986) and listen as they scan the graphics and relegate any text to a "need to know only" priority. In this study, the children were required to read onscreen text and graphics and listen to grammatical examples and non-examples. Under these conditions, attention focusing becomes serial; only one task is processed at a time. Reading and understanding the difficult tasks or unfamiliar items becomes a priority. The child begins a process of articulating the words in their mind's ear, the aim to create a cohesive episode (Baddeley, 2004) or integrated element (Merrienboer & Ayres, 2005) that gradually changes the structure of their long-term memory. If the temporal sound cues are brief and memorable, gist can evoke prior knowledge and skill, and elaborate and strengthen their existing schema.

These results are consistent with research with adults in which the expectation was that agent motion and gesturing would capture attention and offer feedback without breaking the learner's train of thought (Craig, Gholson & Driscoll, 2002; Johnson, Rickel & Lester, 2000; Veronikas & Shaughnessy, 2005). However the moving agent didn’t get any better results from that of the agent that did not move. Adults that learned better with the voice of a pedagogical agent, rated the lesson more favourably, recalled more, and were better able to use what they had learned to solve problems, than students who learned the same verbal materials presented as on-screen text (Moreno, Mayer & Lester, 2000). Moreno, Mayer, Spires and Lester (2001) also reported better problem solving transfer with a speaking agent, with both undergraduate psychology students (n = 44) and with seventh graders (n = 48), but did not find better results for their delayed post-test of facts in either group. Craig, Gholson and Driscoll (2002) also found better delayed post-
test of facts and problem solving transfer in undergraduate students (n = 135) using a speaking agent. Baylor and Ryu (2002) have reported that the inclusion of speech is a key element to successful agent-based learning environments.

**Limitations**

There were several limitations to this research. First, this research had a particular purpose and therefore did not investigate whether or not these children attended to screen locations, nor how they attended to cues, only whether they learned grammar better from convergent temporal sound or text cued by an animated agent character. If we had wanted to look at the psychological processes associated with attending to a cue during a grammar lesson (temporal speech vs. temporal text), we would have had to conduct an analysis of the cognitive episodes in their verbal protocols, as done previously with pre-service teachers (Mann, 1995).

Second, the learning effects reported in these children may be quite different from learning in adolescents and adults. Young children are not simply little adults, not capable of reasoning as an adult until they reach the age of 15 (Piaget, 2000), yet we sometimes expect children to think like adults when they are not incapable of doing so. Perhaps this is because “from the time they enter school, young children use language that seems adult-like in many respects” (Ormrod, 2003, p. 48). The human brain is not fully developed until late adolescence and in some males not until early adulthood (Case, 1985). Whereas adults and adolescents can be expected to examine auditory-visual events systematically and completely, learning for children is a different experience (Pressley & McCormick, 1995). Only with increasing age and experience, can a child's processing become efficient (Mussen, Conger & Kagan, 1974). Over time and with instruction, their capabilities develop and mature in receptive language (i.e., the ability to understand what is heard and read) and expressive language (i.e., the ability to communicate effectively through speaking and writing) (Ormrod, 2003), the focus of our research, specifically, the rules of syntax. Caution should therefore be exercised in generalizing the results of studies with children to adults, and vice versa.

A third limitation was that the treatments were all self-paced, not system-paced, again a normal situation for the students in that school. Like their adult counterparts, these school-age students were able extract details from reading, and acoustically encode the gist from listening, contrary to the view (e.g., Ginns, 2005) that that lack of time pressure in the self-paced treatments would not increase cognitive load sufficiently to show a modality effect.

Fourth, motivation was left unplanned because the focus of this study was solely informational. Only simple learner controls on the interface were used instead of more sophisticated methods evident in children’s educational software, such as learner-control with advisement.

A final limitation was that although they were randomly assigned to treatment groups, the children in both experiments were nevertheless, a population of convenience wherein whole (intact) classes of students were used.

**Summary**

The SSF and ACTML offers and practitioners the advice they’ve been seeking, and modality researchers some relief from “narrating the screen text” and other stochastic
roles for sound (Mann 1997b). Continued research is needed on listening to speech events in multimedia, and how school age and even pre-school children might focus their attention with a speaking agent.

REFERENCES


Mayer, R.E. (2003b) The promise of multimedia learning: using the same instructional design methods across different media


