Abstract

Multimedia sound is both durable and resistant to interference and forgetting. Yet sound alone is insufficient to learn from multimedia, hence the need for purposeful advice on how to enhance learning from technology with sound. The advice ranges from descriptions of the playback system to balancing the input to structuring the function of a sound. This paper describes five functions and three structures for multimedia sound that when combined can help students to focus their attention on important visual events in a multimedia learning environment.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Multimedia; Sound; Modality; Attention; Reading; Listening; Media; Audio; SSF model

1. Introduction

In practical terms sound can be used either within a design framework or as a job aid for the teacher or student, media professional or researcher, software or instructional designer. The end product can be a scripted procedure or a heuristic for helping end-users to control their attention during multimedia learning. Heuristic is taken here in its functional sense, rather than the computer modeling sense (Bregman, 1989). The value in giving a sound a structure and purpose in multimedia resides in the potential contributions to research and development in cognitive psychology, instructional technology and user-interface design. This paper begins with a rationale for including sound in models and theories of instructional design and is followed by an overview of the search for design guidelines. In Section 4 the two dimensions in the original SSF model are introduced. Section 5 describes changes to the original model and the factors affecting those changes. This account is followed in Section 6 by an explanation of the revised SSF model.

2. The problem

A persistent problem in learning from multimedia is that students ignore or forget to read important instructions and feedback presented in text or other visual displays (Ragsdale, 1988), the way illustrations and pictures are ignored, regardless of their intended function, and unless they were instructed to do so.
“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). Unlike entertainment multimedia, educational multimedia required reading and listening to instructions and feedback presented in the program or website.

Although there are well-documented methods of designing visual instruction to control attention (e.g., Carney & Levin, 2002; Rieber, 1994; Zahn, Barquero, & Schwan, 2004), this paper is only concerned with instructional design solutions that require sound. Sound can provide a context for learning from multimedia, especially for poor and beginning readers. Whereas good readers can use their context-free word recognition skills, poor and beginning readers use repetitive sentence context. Poor readers gain more from context than good readers, consistent with Stanovitch’s hypothesis (Goldsmith-Phillips, 1989; Nickerson, 1991; Swantes, 1991; Yeu & Goetz, 1994).

3. The rationale for sound

A strong rationale for using sound in the design of multimedia is that “memory for material presented in sound is more durable and resistant to interference from other modalities than visually-presented material, and more resilient to forgetting than visual traces” (Broadbent, Vines, & Broadbent, 1978 in Baddeley, 1986, p. 42). Neurological evidence shows that sound stimulation can evoke responses in visual areas of the brain, even in very young children (Goswami, 2004). O’Leary and Rhodes (1984) reported that when babies listened to an audio recording of one woman from a speaker located halfway between two videos of different women speaking simultaneously, the babies preferred to watch the face that belonged with the voice they were hearing. The babies shifted their attention until they associated the sound with the visual event. Young children will also use sound to monitor television programs for critical or comprehensive content to which they will then attend visually (Seels, Fullerton, Berry, & Horn, 2004).

Likewise multimedia learning experiments with adults (mostly undergraduate psychology majors and pre-service teachers) showed that they learned better from illustrations when the accompanying verbal information was heard, rather than read because the instructional material did not require them to split their attention between multiple sources of mutually referring information (Chandler & Sweller, 1991; Kalyuga, Chandler, & Sweller, 1999; Mayer & Moreno, 1998; Moreno & Mayer, 2000). When post-secondary students watched an animation, they learned more when verbal information was narrated rather than left as on-screen text (Mousavi, Low, & Sweller, 1995). Mayer and Moreno (1998) reported that college students learned better when information with animation was presented as speech rather than on-screen text both for concurrent and sequential presentations. These students generated 50% more creative solutions on problem-solving questions from animation with narration, than those with animations and text. Mann (1997a) also found that university students learned better from some text with speech and diagrams, than from either speech and diagrams, or text and diagrams. In protocol analysis research, Mann (1995a, 1995b) determined that university students working with speech and diagrams made more evaluative-level verbalizations than those with text and diagrams.

Reports like these about the durability of sound and its resistance to interference and forgetting provide support for including sound in multimedia. However, sound per se is not sufficient to consistently effect learning from multimedia. Multimedia learning is more than synaptic responses to sensory stimulation. Designers must decide if audio should replace or enhance the on-screen text (Barron & Kysilka, 1993). “A systemic model for designing sound is required, “designer sound for computer systems” (Buxton, 1989, p. 1), purposeful advice on the design of instruction for multimedia learning based on learning theory (Koroghlanian & Klein, 2004; Mayer, 2001). Hence the need for design guidelines.

4. Design guidelines and definitions of multimedia learning

A researcher’s opinion about how a student learns from multimedia can influence their choice of design guideline. Likewise the choice of design guideline can be seen to be a reflection of their working definition
of multimedia learning. Further, the researcher who says he or she never uses a guideline but keeps certain things in mind, is mentally applying a heuristic or guideline. Again heuristic is taken here in its functional sense. Table 1 shows the corresponding definitions of multimedia learning and design guidelines.

The choice to adopt one definition or another may depend on the requirements of the research, the knowledge of design guidelines and of course the researcher’s knowledge of definitions of multimedia learning. In any case when a design is adopted a definition has been adopted. When a definition is adopted, a design is forthcoming. In this context, this section will describe in some detail, six approaches to the design of multimedia for learning based on the premise that the approach is reflective of an underlying assumption about learning from multimedia: design by intuition, by technology selection, for balanced input, for cognitive load, for audio-visual sensations, or design by structured sound function.

4.1. Designing by intuition

Design by intuition is a stochastic approach to combining sound and visual events in multimedia that is costly and ineffective (Mann, 1997a). A design by intuition usually offers plenty of advice about how the machine should behave and what the student should be doing in the program. Stochastic methods employ a trial and error approach to interface design and describes human interactions with technology as a function of the playback system performance instead of whether or not student attention was sustained and their long-term memory store was changed. Sometimes this is because the state of the research at the time. In the days of teaching machines for example, programmed instruction research (Blyth, 1960; Crowder, 1960; Lumsdaine & Glaser, 1960; Skinner, 1961) used an intuitive approach to combining sound and visuals in educational technology. There were plenty of ideas about whether or not students were reading the screen or listening to directions,”The student may access auditory material and listen to a passage as often as necessary then transcribe it. The machine then reveals the correct text so that he may listen to the passage again to discover the sources of any error” (Skinner, 1958, p. 147). However, little was known of the effects of auditory manipulations on language learning (Porter, 1960). Consequently, early approaches to learning from audio-visual technology were found to be ad hoc, unproven and required frequent reexamination and clarification (Conway, 1968). Inconsistent outcomes from audio-visual presentations were due to the erroneous assumption that the sound would take care of itself, “experimentation showed that the sound does not take care of itself” (Hartman, 1961, p. 256).

Although more is known today about learning from technology, the results are still mixed (see a review in Koroghlanian & Klein, 2004). Sound, like illustrations can have a positive, neutral or a negative effect on learning (Levin, Anglin, & Carney, 1987; Rieber, 1994; Sims-Knight, 1992; Winn, 1993). For this reason design guidelines that rely on human intuition are too generic, too specific, irrelevant, or outdated for use in most research settings (Hammond, Gardiner, Christie, & Marshall, 1987; Lee & Boling, 1999).

4.2. Designing by technology selection

Researchers who compare one communication technology with another has a technology selection view of multimedia learning, or the delivery media view of multimedia learning (Clark, 2001; Mayer, 2001). This view is based mostly on media comparison research with adults and reports of production effects in television

### Table 1

<table>
<thead>
<tr>
<th>Corresponding definitions of multimedia learning and design guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definitions of multimedia learning</td>
</tr>
<tr>
<td>• Intuition definition –</td>
</tr>
<tr>
<td>• Technology selection (delivery media) definition –</td>
</tr>
<tr>
<td>• Balanced input (presentation modes: pictures-words) definition –</td>
</tr>
<tr>
<td>• Cognitive load definition –</td>
</tr>
<tr>
<td>• Audio-visual sensations (sensory modalities: audiovisual) definition –</td>
</tr>
<tr>
<td>• Attentional control definition –</td>
</tr>
</tbody>
</table>
programs on children’s attention (Schramm, 1972; Seels et al., 2004). Technology selection guidelines have helped to popularize the view that multimedia learning is a matter of selecting appropriate technology. A videotape or video clip, for example, would afford the researcher live action motion. Or a book or multimedia encyclopedia would afford the researcher indexing capability. The role for sound from a computer was to decorate, enhance, create fantasy, reward, or represent that which would have otherwise been least effectively communicated as text or numbers (Daiute, 1985; Malone, 1981).

Studies based on technology selection guidelines are widely published in the literature, even today, Technology selection research usually compares computer- or web-based material presented in a lab with an instructor-based presentation in a classroom (e.g., Jones, Morales, & Knezek, 2005; Steelman, 2005; Zhang, Perris, & Yeung, 2005), or distance education technologies with classroom instruction technologies (Bernard et al., 2004). Instructional design researchers, for example (Dale, 1946; Heinich, Molenda, Russell, & Smaldino, 1999; Reiser & Gagné, 1983), are well known for their technology selection guidelines. Selecting a technology usually means setting learning objectives and tests, and drawing the instructional analysis flowchart. Design and research for technology selection is likely to continue as long as evolving hardware and software permit designers and students more adaptive and nonlinear interactions and a greater differentiation of visual events and sound. Even haptic events may soon become part of multimedia learning, such as hands-on simulation, hand-sensing gloves and even simulated reach-in-and-grab technologies that can also be downloaded into content, computer applications for user interface navigation.

Some critics of the technology selection definition of multimedia learning argue that research about whether It teaches better than some other technology is a unproductive line of questioning (Mayer, 2001; Salomon & Gardner, 1986). Other critics of the technology selection definition argue that comparisons of technologies are necessary if only to show that an effect believed to be unique to a particular technology can actually be replicated with older technologies (Clark & Feldon, 2005), because all technology effects are, actually, due to instructional design (Clark, 1983, 1987).

It remains to be seen how the SSF model could be applied in research using multimedia as a model for technology selection. It seems likely that the prevalence of the technology selection definition in the literature is due to its timeliness and advice. To date, the original SSF model has not been tested as a method for selecting one technology over another.

4.3. Designing for balanced input

The researcher who aims to balance verbal and nonverbal representations in students’ working memory by weeding and off-loading information from the visual events into sound signals (Mayer, 2001, 2003) may be said to define multimedia learning as a balanced input of pictures and words. The balanced input definition, also known as the presentation modes view of multimedia learning (Mayer, 2001), assumes that communication technologies share certain characteristics (Lepper, 1985; Salomon & Gardner, 1986) that can produce similar verbal and nonverbal representations sufficient to produce learning from multimedia. The balanced input definition of multimedia learning is based on Paivio’s dual coding theory (Sadoski, Paivio, & Goetz, 1991). Dual coding theory makes a distinction between two distinct cognitive systems: language-like system of verbal codes called logogens, and nonverbal system of spatial codes called imagens (Paivio, 1986). The verbal system includes language, and the nonverbal system includes mental imagery. Referential connections between the systems account for the evocation of mental images by language (or language by images). Dual coding theory has had considerable influence on the development of important new theory, including the cognitive theory of multimedia learning (Mayer, 1997, 2001), and the dual processing theory of working memory (Mayer & Moreno, 1998).

There are many examples of research that rely on descriptions of the balanced input of pictures and words in students’ working memory and long-term memory store. A physics simulation by Rieber, Tzeng, and Tribble (2005) for example, was guided by dual coding theory. The researchers found more effective learning when information was encoded both visually and verbally. Notably in this definition, illustrations and guidelines for their effective use were believed to transcend the medium.

Initially the SSF model adopted the balanced input definition of multimedia learning to provide a partial explanation of how to more equitably balance the input of illustrations and words between the auditory and visual systems (Adams, Mann, & Schulz, 2006; Mann, Newhouse, Pagram, Campbell, & Schulz, 2002).
experiments with 7th graders, however, no significant effects were found. The findings suggested that some school-aged students using educational multimedia were unable to generate sufficient gist to solve problems; that their under-developed biological capacity to extract gist from speech limited their mental ability to generate sufficient referential connections between the speech prompts and the text, and the speech prompts and diagrams. A better definition of multimedia learning was needed to account for multimedia learning in children.

Learning for young children is different from learning for adults. Young children are not simply little adults, not capable of reasoning as an adult until they reach the age of 15 (Piaget, 2000). 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). The human brain is not fully developed until late adolescence and in some males not until early adulthood (Case, 1985). Moreover, age differences in children show a wider range of variability in task performance than age differences among adults (Cowan et al., 2005). Shilling’s (1991) study with 81 kindergartners reported statistically non-significant learning effects between students using conventional writing materials and/or computers with and without available synthesized speech feedback over an eight-month period. Similarly, Weiner (1989) investigated the differential effects of presentation conditions; visual only and visuals cued with speech, on sight-word learning with 55 handicapped third graders. The results indicated non-significant statistical differences between the presentation conditions. However, two years later Weiner (1991) reported different results with 24 junior-high students.

Multimedia learning depends on reading and listening, yet reading is not the same as listening. Readers and listeners extract different kinds of information from oral and written statements. Listeners recall more of the gist of a story and readers recall more of the surface or verbatim features of the story (Hildyard & Olson, 1982). Readers attend to details, surface features of the text, or gist (Tannen, 1985). Results of neurological testing (Brainerd, 1993; Brainerd & Reyna, 1990; Reyna, 1992) indicates that humans have distinct developmental paths for verbatim and gist memories of verbal information. Structured sound functions discussed later, can provide a reading context for learning from multimedia, especially for poor and beginning readers. Activities such as focusing on-screen information or deciphering warning signals require listening for a specific purpose (Mann, 1997a).

4.4. Designing for cognitive load

The researcher who aims to keep from over-loading (or under-loading) the student with sensations may be seen to apply the cognitive load definition of multimedia learning. Guidelines for sharing the sensory input has helped to define multimedia learning as cognitive load.

Cognitive load theory (Chandler & Sweller, 1991; Sweller, 1988, 1999; Sweller & Chandler, 1994) distinguishes between three types of cognitive load in working memory: Intrinsic, germane, and extraneous. Intrinsic load refers to the number of information elements the learner needs to hold in working memory simultaneously to comprehend the information. Element interactivity is intrinsic to the material being learned. Germane load is the utilization of cognitive resources left-over from intrinsic load that require extra effort invested in processes directly related to the simultaneous activation of elements in working memory to be integrated into existing schema. Extraneous load is that which is increased in a learning environment or instructional methods that requires students to engage in activities not required for new schema construction or automation. The cognitive processes of shifting mental effort during learning has been well defined elsewhere (Paas, Renkl, & Sweller, 2004; Sweller, van Merrienboer, & Paas, 1998).

In practice, auditory explanations in multimedia are often used simultaneously with the same visually presented text (Kalyuga, 2000). This can increase the risk of overloading working memory capacity and might have a negative effect on learning. Corresponding idea elements in visual and auditory form can consume cognitive resources. In such a situation, the elimination of the a redundant visual information can be beneficial for learning. Similarly, auditory explanations (e.g., narration, explanation) may also become redundant in experienced learners. If an instructional presentation forces these learners to attend to the auditory explanations continuously without the possibility of skipping or ignoring them, learning might be inhibited (Kalyuga, 2000). These observations have been considered in the revision of the SSF model.
4.5. Designing for audio-visual sensations

The researcher who aims to classify visual and sound sensations separately may be adhering to an audio-visual sensations definition of multimedia learning, also known as a sensory modalities view of multimedia learning (Mayer, 2001). Some researchers organize their audio-visual sensations using expressionistic film-style or semiological guidelines (Bordwell & Thompson, 1979; Doane, 1980; Gorbman, 1983; Prendergast, 1977; Seidman, 1986; Spottiswoode, 1950). The rational for the audio-visual sensations definition is that students who are regularly bombarded by ever-deepening visual sensations during multimedia learning (e.g., hi-res graphics, video captures) needs heightened sound in their instructional messages if only to perceive them at all (Burkman, 1987; Paine, 1981; Ragsdale, 1988).

Baddeley and Hitch (1974) postulated a tripartite model of working memory that classified sensations into visuals and sounds – some that need articulation. The three components in this descriptive model are a phonological loop for acoustic and verbal information, a visuo-spatial sketchpad, and a limited-capacity attentional controller called the central executive. The central executive directs attention to relevant information, suppresses irrelevant information and inappropriate actions, and coordinates cognitive processes when more than one task must be done at the same time (Baddeley, 1992; Norman & Shallice, 1980). The phonological loop is comprised of two components, a phonological store and an articulatory rehearsal system. Whereas auditory presentations can gain direct access to the phonological store, visually-presented material is encoded in the phonological loop via subvocal articulation (Baddeley, 1992). The visuo-spatial sketchpad forms an interface between visual and spatial information accessed either through the senses or from long-term memory, and allows for a range of channels of visual information to be bound together with similar information of a motor, tactile or haptic nature. A newly conceptualized component, the episodic buffer that is recently separated from the central executive (Baddeley, 2002; Baddeley & Andrade, 2000), is critical to learning from multimedia. Although the structural elements in the sensory modalities definition were generally supportive of structured sound functions in adult learning, the explanations did not target learning from multimedia directly, nor learning in young children. Nevertheless these ideas were considered in the revised SSF model.

4.6. Designing by structured sound function

The researcher who designs by structured sound function aims to solve the persistent problem discussed earlier in this paper, namely that students ignore or forget to read important instructions and feedback presented in text or other visual displays (Mann, 1988). The rationale for giving sound a function and then structuring it with a visual event was that captivating images would not hold the students’ attention for long if the aural sense was not suitably stimulated (Alkin, 1973; Alten, 1981; Rosenbaum, 1978). The emphasis on suitably stimulated implied that certain kinds of stimuli were better than others. Unlike stochastic approaches however, the structured sound function heuristic had been adapted from a set of psychological categories used with audiences to describe the effectiveness of the narration in post-war instructional films (Zuckerman, 1949).

4.6.1. Attentional control

The attentional control view of multimedia learning is presented in this paper as a support for the structured sound function design model (Mann, 2006). Fig. 1 shows the structure and process of learning from multimedia according to the attentional control definition of multimedia learning.

The process of learning from multimedia for adults and children begins when the learner reads some text or watches a graphic or animation, and listens to speech, music or sound effect that motivates, directs, instructs, hints, reminds or gives feedback. The learner’s executive controller accesses these data from the senses, focuses attention to learn, sends data to and from long-term memory as needed, compensates for deficiencies, and guides behaviour. In the attentional controller the learner activates and combines gist from auditory memory system and details from their visual memory system into a coherent episode in the episode buffer. From the visual event, the learner forms an interface between the spatial and visual data in their visual-spatial memory. From the sound, the learner encodes the gist directly into a phonological store, and the details indirectly through an articulatory loop. As a consequence, the learner’s long-term memory is gradually altered in the process.
Taken together, the psychological description (i.e., the attentional control definition of multimedia learning) and educational design (i.e., the SSF design model) can be said to be a two-way street (Mayer, 2003). Whereas the attentional control view of multimedia learning can help us to understand what is going in the learner when the SSF model is used, investigations with the SSF model inform the functionality of the structures within the attentional control definition of multimedia learning. Most importantly, the attentional control definition of multimedia learning assumes that the under-developed phonological loop described in young children can be associated with insufficient mental articulation during reading and listening in multimedia. The SSF model can provide the needed context for reading instructions and feedback in multimedia.

Initial applications of the original SSF model have been successful; initially with undergraduate Communications students learning to edit expressionistic video clips from raw footage (Mann, 1988), and again with college students learning Calculus concepts from a computer tutorial (Mann, 1990). In both applications, continuous, convergent temporal sound was associated with important visual events to shift student attention. In the former case, pre-recorded instruction, navigational direction, hints, feedback and reminders were played from a videotape. In the latter, pre-recorded instruction, navigational direction, hints, feedback and reminders were played from a digitized file in a spreadsheet program.

4.6.2. Revising the structured sound functions
Revisions that were made to the original structured sound functions were a consequence, or at least a partial consequence of four factors: the growing set of definitions of multimedia learning, discussed previously, the perceived growth of user-oriented design, user-friendly software, and pilot testing the model. These factors are discussed next.

4.6.2.1. Factor 1: More definitions of multimedia learning. The first factor affecting the revisions to the original structured sound functions was the ever-growing set of definitions of multimedia learning. Multimedia means different things to different people (Clark, 2001; Mayer, 2001). There are at least six definitions of multimedia learning in use at the present time. When there are more definitions to consider, there will also be more design guidelines to learn to apply. As we have seen, a researcher’s choice of design guideline is a reflection of their working definition of multimedia learning. Why should the number of definitions of multimedia learning affect the use or outcomes of the original SSF model? The answer is that the SSF model should be flexible enough to accommodate any definition of multimedia learning. Revisions were needed to the SSF model to gain some flexibility.

4.6.2.2. Factor 2: The perceived growth in user-oriented design. The second factor affecting revisions to the SSF model has been the perceived growth in user-oriented design, “perceived” because course design appears to
have retained the teacher as the designer of the course, symptomatic of instructor-centred or objectivist instructional design. Instructor-centred instruction is the traditional definition of instructional design (Reigeluth, 1983), in which a teacher or instructional designer plans and revises the media and methods best suited to bring about changes in student knowledge and skills. Learners are typically excluded from the design process itself. Objectivist instructional design was the standard practice during the 1990s with researchers hiring instructional designers to design the instructional materials. However, instructional design models and theories were hardly ever used with contemporary multimedia authoring, and when they were, were misapplied (Gros, Elen, Kerres, Merrienboer, & Spector, 1997), found to be static (Boshier et al., 1997), or inert (Yang, Moore, & Burton, 1995) or simply unusable (Wild & Quinn, 1998) for prescribing the conditions of interactive learning for all learners in all settings. Some (Tergan, 1998) believed that instructional design was based on psychological and educational theories that were simply too broad in scope and too rigid for contemporary multimedia instruction.

Learner-centred or constructivist instructional design has a different emphasis. Hannafin and Hill (2006) introduced the term constructional design to mean a learning environment that enables and supports an individual by engaging them in design and invention tasks where knowledge-building tools are provided but concepts are not explicitly taught. Unlike the snapshot-of-the-learner used in instructional design, constructional design assumes the learner to be an active, changing entity. The shift in emphasis from teacher-centred to learner-centred design (Reigeluth, 1999) is consistent with the principles of user-designer (Banathy, 1991) and user-oriented instructional design (Burkman, 1987), prompted by dissatisfaction in trying to apply familiar models and theories of instructional design (e.g., Dick & Carey, 1996; Gagné, Briggs, & Wager, 1992).

Why should the perceived growth in user-oriented design affect the use or outcomes of the original SSF model? Again, the answer is that the SSF model must be flexible enough for use in either definition of instructional design, that is, instructor-centred (objectivist) or learner-oriented (constructionist), as required.

4.6.2.3. Factor 3: The prevalence of free or low-cost, user-friendly software. The third factor affecting revisions to the SSF model has been the prevalence of free or low-cost, user-friendly software. During the 1990s authoring systems such as HyperCard (Apple Computer, 1991), ToolBook (Asymetrix Corporation, 1990) and PC PILOT were purchased by colleges and schools for use by educators and students alike (Andrews & Collins, 1993; Brader, 1990; Christie, 1992). Colleges and schools working in the MS-DOS environment for example, learned to use ToolBook (Asymetrix Corporation, 1990) to develop applications for on-the-job computer-based training. Those working on the Apple platform learned HyperCard (Apple Computer, 1991) and its powerful programming language HyperTalk to produce pleasing visuals and sound. The rapid prototyping capabilities of HyperCard, HyperSound and MacRecorder made them appropriate developmental tools for structuring sound into learning environments (Mann, 1994, 1995a; Mann et al., 2002).

Today these and other authoring programs have been replaced by comparable software at low or no cost. For example, educators and students who have purchased a copy of the Windows XP operating system can access the free authoring tools Photo Story (MicroSoft Corporation, 2006a; MicroSoft Corporation, 2006b), or Movie Maker bundled with the program. Other users may prefer Flash Basic 8 to design simple motion graphics and interactivity for delivery in a Flash player, or Macromedia Director MX 2004 for K-12 or Flash MX 2004 for 2D animation.

Teachers using any of these programs can learn to design learning materials or let their students take an active role in the design of their own materials. Combined with intuitive software then, the SSF model can be used as a job aid for learning through multimedia, useful in situations when it is not feasible or worthwhile to commit a procedure to memory. Job aids are often used instead of instruction to save time and money (Rossett & Schafer, 1991), when an individual or group must remember how to complete a task that is infrequently performed, or when the task must be accomplished exactly the same way every time (Boyd, 2005). The individual may understand the task, but the specific sequence of steps in completing the task may be esoteric or difficult to remember (Brown & Green, 2006).

How does the prevalence of free or low-cost, user-friendly software affect the use or outcomes of the original SSF model? The answer is that when combined with structured sound functions, easy to use software can indirectly affect student learning. Since the software is less expensive, more educators are likely to use it. Since more can be done with the new software, more will be expected of the new software, and more will be expected
of the SSF model for use with the new software. Instruction and job aids must therefore be accessible and flexible enough in the SSF model to help the novice designer to work with their new software.

4.6.2.4. Factor 4: Effectiveness of the SSF model with different groups of learners. The fourth factor affecting revisions to the SSF model was the effectiveness of the model with different groups of learners. What did testing reveal about the original SSF model? Independent tests showed good results with adults (Mann, 1988, 1994, 1995a, 1997a) and in 4th and 5th graders (Mann, Cui, & Adams, 2002), but only equivalent results with 7th graders (Adams et al., 2006; Mann et al., 2002). A research design was needed that would account for long-term learning effects, and for developmental factors in children. A new model was developed called the attentional control model of multimedia learning, discussed earlier.

5. Revised structured sound functions

Revisions were made to the original SSF model. Some were substantive, such as the addition of political and social-cultural points of view. Other revisions were made for convenience. In the original model, functions and structures for sound were represented separately (Mann, 1992, p. 54, 57). In the updated version of the model, the functions and structures of sound are represented together. The classification scheme for choosing a function for the sound was also revised. In its original form, the old classification scheme had 15 strategies (i.e., informational/emotional, pacing, rhythm, review and delivery), so many that the functions and structures of associating sounds with visual events had to be represented in separate diagrams (Mann, 1992, 1995a, 1995b, 1997a, 1997b), and too complex for use by the part-time or student designer. Consequently, the total categories in the revised model were reduced to 7 and are represented together in a single diagram (see Fig. 2). Temporal and atmosphere functions are unchanged. Character sound has been consolidated into a single category. Locale sound now has real and imaginary categories. Two new points of view have been added (i.e., political and social-cultural). Fig. 2 shows the five functions and three structures of sound in the revised SSF model to help learners to focus their attention on visual events in multimedia.

<table>
<thead>
<tr>
<th>&lt; Structuring the sound with a visual event &gt;</th>
<th>The goal: is convergent or divergent</th>
<th>The constancy: is continuous or discontinuous</th>
<th>The density: is massed, spaced or summarized</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; Giving the sound a function &gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A temporal prompt:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that cues</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that counterpoints</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that dominates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that undermines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A point of view:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>objective, subjective, performer, political, socio-cultural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A locale:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>real, imaginary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An atmosphere, feeling, mood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A character’s:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>past, future, personality</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. The five functions and three structures of sound in the revised SSF model.
5.1. Giving sound a function

Fig. 2 shows the structured sound function (SSF) model and the five possible functions conceptualized for sound, to describe the context of a character, a place, a time, or subject matter with a visual event presented in multimedia. 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, voice-over IP, or a podcast. Given a purpose or function, a sound can alert, caution, warn, remind or direct the learner 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).

Selecting a sound for a specific purpose or function according to this model is only half the job done, however. The other half is setting the function into a suitable structure that determines how when and how often the sound will play that function. Put another way, any visual event that is assigned a sound for a specific purpose, must also have a structure for that sound.

5.1.1. Temporal sound

A temporal sound is an alert, caution, warning or direction about a future event, or a reminder about a past event that is displayed as a visual event in a computer program or at an Internet site. Some examples of temporal sound cueing include: instruction, navigational direction, hinting, feedback and reminders.

5.1.2. POV sound

A point of view (POV) sound describes a particular perspective toward adding sound to help learning from multimedia. When an objective, subjective or performer point of view is presented in sound, it can imply another point of view or more information about the point of view than what is stated or implied in the visual event. Alternatively, a POV sound can be prescribed within a character’s personality, showing internal conflict between objective, subjective and performer points of view. POV sound can also present opposing opinions about political performers, or deeply felt moral, cultural or religious beliefs.

5.1.3. Locale sound

A locale sound can fill an informational role when it is associated with a visual event presented in a video clip, graphic, or a paragraph of formatted text. Most often familiar sounds are added to establish a place, real or imaginary. Locale sound in public spaces can transform a passive experience into an active one through an earpiece. For example, cell phone technology can offer users in a city or historic site, or at a public event, a more comprehensive education of their surroundings. Using an electronic tour guide for example, participants take the role of information receiver, learning from curators about art objects. *Hear & There* (Rozier, 2000) is a more sophisticated application of locale sound. Hear & There is a computer system that uses sound overlays to permit users to leave their personal signature in a particular location in physical space. Their sound is a customizable and can be placed at a physical location. Using the authoring toolkit in Hear & There, an author can create some augmented content for a particular physical space by walking (or driving) to the space, creating a sound and embedding it into a portable or fixed computer at the location. A person traversing the designated location will then hear the author’s personal signature, and may wish to augment their imprint with their own.

5.1.4. Atmosphere sound

Atmosphere sound can provide the context for an event for the listener/viewer in the absence of visual information. For example a broadcast journalist using streaming technology over the Internet can verbally provide an eyewitness account of a scene after-the-fact. Alternatively, a voice-over plus a map or photo of the speaker may suffice in lieu of a visual evidence of the event. Atmosphere sound can also provide a feeling about a human condition. Still other times, atmosphere sound can set the mood, as in a celebration or a political meeting. Atmosphere sound can easily be misused however, by manipulating the structure of the sound—
visual relationship. In a major CD project, for example, where a range of sounds were integrated as the learners navigated the information, a learner made a choice to see more detail about Thailand and a short musical track was played. The learner commented that the sound was not Thai music and decided that the remaining information would not be useful” (Sims, 2006, p. 5). Juxtaposing inappropriate music or laughter over a visual event therefore, can undermine the original intent.

5.1.5. Character sound

A character sound refers either to a character’s past, future, or personality, either real or virtual. Personality sound refers to the subtext, story spine or tragic flaw in a character. Like personality sound, the character’s past or future sound contains auditory references to a character's personal, private or public event or idea. Unlike personality sound, however, this function does not plumb the depths of the character’s psyche.

5.2. Structuring with a visual event

Fig. 2 illustrates that to design the sound for a multimedia program or website sound function should have three structurtes to prescribe how, when and how often a sound should occur with a visual event. Put another way, any visual event that plays a sound to provide a context must have a function for that sound, according to this model.

Structuring the goal, constancy and density of a sound with a visual event during multimedia learning is a method of associating sound with that visual event. The process is analogous to describing a scientific process or telling a story.

The goal of a sound can be either convergent or divergent with a visual event, just as learning to solve an algebraic equation for example may be either convergent or divergent. A convergent sound goal for a visual event means that the student’s attention is directed toward a solution in a single, stepwise procedure. A divergent goal for the sound means that the student’s attention is directed toward visual authoring tools, libraries of resources, or advice that permits easy access and experimentation in an open environment.

The constancy of a sound describes its duration and is either continuous or discontinuous with the visual event. A continuous sound structure has uninterrupted sound at regular intervals throughout the sequence or entire program. Continuous sound requires a specific purpose or function for the visual event – temporal, POV, locale, atmosphere, character. Without a purpose the sound has only a continuous reporting role. Oral reporting role for sound is insufficient to help users to pay equal attention to several tasks by distributing student attention between more than one competing message or source. See Mann (1997b) for details and examples of oral reporting. Discontinuous sound structure has interrupted sound at regular or irregular intervals with a visual event throughout the sequence or entire program. Discontinuous sound requires a temporal, POV, locale, atmosphere or character function.

The density of a sound describes the periodicity for each chosen sound function in a script or sound with a visual event. When fit with a temporal, POV, locale, atmosphere or character function, the density of a sound can be massed, spaced or summarized with the visual event. A spaced, massed or summarized sound describes when and how often an auditory warning, music or speech is reviewed with a visual event. Problem solving skills can be effectively presented with a spaced sound density. Personal goal areas on video clips or still graphics can be presented all at once and even out of program context using a massed sound density. Corporate applications tend to use a review sound density to reinforce role-modeling techniques. Internet news overdubs use summarized sound density to recap main stories.

Selecting the goal and the constancy for a sound are most important of the three structural components. Prescribing sound for a novice is best served with a convergent goal and continuous constancy. Continuous sound may be especially needed for a difficult task or with an unfamiliar item. On difficult tasks or with unfamiliar items the student must consciously control their mental processing (Schneider & Shiffrin, 1977) to focus their attention (Treisman, 1986). Under these conditions, attention focusing becomes serial; only one task is processed at a time. Learners must consciously focus their attention “to bind the separate features of a stimulus such as the colour, shape, words, into a unitary object” (Matlin, 1989, p. 57). To engage the appropriate attentive state, the student must self-initiate the appropriate system of information processing (Borich & Tombari, 1995). Sometimes working memory simply is incapable of highly complex interactions using novel
elements. Discontinuous sound may be sufficient for an easy task or an unfamiliar item. On easy tasks or with familiar items, a student will implement automatic processing (Schneider & Shiffrin, 1977) also known as pre-attentive processing (Treisman, 1986). Pre-attentive processing of easy tasks or familiar items occur in parallel; that is they can handle two or more items at the same time. Under these conditions discontinuous sound may suffice. An example of a continuous temporal reminder would be the repetitive squealing of monkeys in Milie’s Math House (Edmark, 1995) when there is no student input.

5.2.1. Structuring for temporal sound

A convergent goal for a temporal sound can shift attention to a visual event. One example of a convergent goal is a spoken direction about how or where to look to find out how to create a personal objective or learning outcome to answer a question. An example of a divergent goal for temporal sound would be a procedural question spoken during a brainstorming task, such as the spoken reminder to ‘click and fully explore’ the multimedia environment or website. A divergent goal for sound would deviate, elaborate, or even contradict a visual event. In structuring for temporal sound for novices (i.e., for difficult or unfamiliar task, or poor or beginning readers) use a convergent goal and continuous constancy. That is, using a hint or reminder can cue the learner frequently to take action with the information presented in the visual event. The cue might request that the student write or draw something, or discuss an issue with their peers on site or online. When the temporal sound is massed it means that the auditory alert, caution, warning or direction is concentrated within one part of the multimedia program or website. A massed temporal sound occurs during an early interaction with the program or website, similar to a news pre-cap, headline or advance organizer. A spaced temporal sound alerts, cautions or reminds the user, such as incoming chat or email. A summarized temporal sound repeats the substance of a longer discussion, like the recap in a television news story.

5.2.2. Structuring for POV sound

Structuring for convergent goal for POV sound aims to resolve a conflict between objective, subjective and performer points of view. A divergent goal for a POV sound is an unresolved difference of opinions among political performers, or among those with deeply felt moral, cultural or religious beliefs. An example of a divergent goal for a POV sound would that generated in the Decisions, Decisions: The Environment (Tom Snyder Productions Inc., 1997) where students learn and apply the lessons of history through role-playing simulation software. Structuring POV sound for novices requires a convergent goal and continuous constancy. Continuous POV sound can be a diversity of spoken comments about a visual event. An example would be a lively debate. Discontinuous POV sound can be a distinctive sound that may be a laugh or interruption that reflects a particular opinion. An example would be an infrequent remark that affects the interaction within the multimedia experience. A massed POV sound is typically an auditory interlude or introduction to an event according to an objective, subjective or the performer’s perspective. A summarized POV sound is usually an auditory re-cap or wrap-up of an event according to an objective, subjective or the performer’s perspective. A spaced POV sound is typically an auditory reminder through an event.

5.2.3. Structuring for locale sound

Structuring for convergent goal for a locale sound would immediately reveal the environment or learning context of the multimedia environment. Structuring for continuous locale sound for novice end-users would require frequent reminders of the learning context. Frequent reminders would preempt the need to explore the context thereby allowing closer engagement with the visual events.

5.2.4. Structuring for atmosphere sound

Structuring for convergent goal for an atmosphere sound aims to provide a desired feeling in the novice user. Structuring for continuous atmosphere sound for a novice would mean making distinctive recurring messages, warning signals or a musical leitmotif.

5.2.5. Structuring for character sound

Structuring for convergent goal for a character sound aims to restrict the character’s role to something in his or her past or future, or a notable aspect of the character’s personality. Selecting a continuous character sound
for a novice would require an ongoing auditory effect, music or utterance (human or computerized) such as a tutor or mentor’s voice that identifies, signifies or personalizes a visual event.

6. Conclusion

Multimedia sound has evolved from an add-on to a learn-from technology. Sound can provide a reading context for learning from multimedia, especially for young and beginning readers that rely on context to read. In this paper a heuristic for designing instruction for multimedia learning called the structured sound function (SSF) model was described for helping students to control their attention to visual events in educational multimedia. Sometimes a student really wants to hear the teacher’s spoken feedback to reassure them, clarify a goal, or direct their learning. These situations warrant a structure-before-function approach to adding sound to the multimedia program (Mann, 1997a, 1995b). In other situations the designer notices that a visual event could be enhanced with a reminder about a previous event in a chemical process, hint in a mathematical proof, or point of view in a story segment (e.g. Mann, 2000). These situations require a function-before-structure approach to adding sound. For best results in educational settings the SSF model should be used with a strong curricular focus, and delayed post-tests to assess long-term learning effects. As job aid or method of designing multimedia instruction, research with the SSF model should also consider the environmental, content and learner factors that will always to some degree, affect student learning and performance.

References


Asymetrix Corporation (1990). ToolBook 1.0 [Computer software].


MicroSoft Corporation (2006b). *Photo Story 3.0 for Windows* [Computer software].


