Time-compressed speech: Emerging ideas for audio in computer-based learning

Winnie Gîllner
OTEN Open High School

and

John Harvey
School of Educational Psychology, Measurement & Technology
University of Sydney

The advent of multimedia - the coordinated use of text, graphics, animation, video and sound - in computer-based learning (CBL) has permitted instructional designers to use speech to augment or replace screen text. The use of audio in CBL generally is increasing with a rise in interest of non-speech sounds and terms such as 'auditory icons' and 'earcons' are coming into the literature (Blattner and Greenberg, 1992; Blattner, 1993). The focus of the present paper however is on speech.

Although text clearly remains dominant, designers have to cope with the limited space available on a screen. Although large colour monitors are available they remain expensive and require high-end computers. Most programs are used on monitors which have screen areas about two-thirds of an A4 page or less, generating crowded displays if text and graphics are displayed together, a problem often exacerbated in that parts of the screen need to be set aside for a header, navigation buttons and status reports. Procedures such as multiple screens, scrolling fields, windowing, abbreviated point-form 'bullets', and reduced font size are used to cope with this difficulty. If the text relates to a graphic, as is often the case, then work on split attention effects has shown that it is clearly not good design to have to switch or 'page' between graphics and text (Chandler and Sweller, 1992).

There is also a problem that it can seem, not infrequently, that users of CBL programs, in our experience, simply do not read on-screen text, or at least do not read with an adequate depth of processing. This issue is recognised in the literature (eg., Gillingham, 1988; Mills and Weldon, 1987), although the formal evidence does not paint as bleak a picture as that implied in the previous sentence. Whether it is a function of being on a screen per se, or the way the reading task is constructed remains a research issue. Reinking (1992) considers that it is clearly the latter and that "new formats for presenting texts electronically with the aid of a computer can increase comprehension and learning compared with printed texts" (p.14). He argues that the poorer performances found in some comparative studies stem from designs which ignore research recommendations. His paper discusses some of

these new techniques and he suggests that "in the future, a wide range of diverse input might be employed to adjust the difficulty or other factors of electronic texts" (p.17).

That optimistic note notwithstanding, it remains as many have noted, that the spoken word is often perceived as a more natural and richer communication channel, as well as having the advantage of freeing screen space. Indeed, Blattner (1993) writes that "audio is probably the interface designer's most useful, and at the same time, most under-used tool ... audio is one of the most powerful methods of engaging the mind and providing information (p. 77). Developments in authoring software for CBL have now made it relatively straightforward to record, edit and play audio routinely as a part of a CBL package.

Compression techniques permit sound files to be stored without the need for exceptional amounts of secondary memory. Speech synthesisers, although not approaching the quality of natural speech, are now readily available (eg., PlainTalk for the Macintosh).

Thus in a test project one of the authors (Winnie Gîllner) has been working on concerning the history of art for HSC students, the full screen could be given over to displaying a complex painting while a voice-over carried the explanation. If the graphic had to compete for screen space with the text, the detail of the painting would have been lost.

A disadvantage of speech however is that it has to be processed by the listener as it happens. Text, and still graphics, on the other hand, generally can be processed in the reader's own time. In particular, ordinarily there is no way to "speed hear" speech, or conversely, to linger on a particular section of the audio and to listen to it more slowly. Of course, the speech track can be readily replayed, but as Arons (1993) has commented:

It is faster to speak than it is to write or type, however, it is slower to listen than it is to read. Skimming or browsing are traditionally considered visual tasks, as we instinctively perform them when reading a document or while window shopping. However, there is no natural way for humans to skim speech information because of the transient character of audio - the ear cannot skim in the temporal domain the way the eyes can browse in the spatial domain (p.1).

Junor (1992) estimates that listening to a message may take perhaps up to three times as long as reading it. He reports that a skilled reader of print typically reads at a rate of some 275-300 words per minute (most likely slower if on screen, however (Reinker, 1992)), whereas spoken reading or conversation is around 125-175 words per minute.

In an effort to overcome this 'listening lag' experiments have been conducted with speeded speech. People can readily adapt to

comprehending speeded speech. The sight-handicapped are known users of recorded tapes and speech synthesisers. The maximum rate at which the untrained ear can still comprehend speech is thought to be about 200 words per minute (Junor, 1992), that is, about 175% normal speed, however with experience much higher rates can be tolerated.

If a normal tape is simply played more quickly then the pitch of the speech also rises leading to 'chipmunk' voices. Junor (1992) reports that while this can be acceptable up to moderate increases in speed, it also depends on several factors such as the quality of the original recording and the pitch and timbre of the speaker's voice. It can be overcome by using special recorders (available commercially) which permit increases in speed without the accompanying increase in pitch.

A number of other methods have been developed to permit a reduction in the time needed to listen to recorded speech (Arons, 1992, 1993, 1994). A sampling or Fairbanks procedure removes segments from the signal. If, say, a two-fold increase in speed is needed then every other 50 msec segment can be removed. Note that unlike the previous method this algorithm does not increase the rate of any given unit of the signal - the increase in input speed comes about because the ear receives fewer units. A related dichotic procedure does not remove segments but rather plays segments 1-3-5-7-9 ... to one ear and segments 2-4-6-8-10 ... to the other in such a way that the segments overlap - the degree of overlapping controlling the increase in speed. Another procedure, termed the synchronised overlap method (SOLA) works by matching the degree of similarity of the wave form at the end of one segment with

the beginning of the next and if a satisfactory match is found then those sections of the segments are overlapped and the average presented as the signal. Speech can also be speeded by selecting and manipulating pauses, described in more detail below. These techniques are termed time-compressed speech (TCS).

Two interesting findings from research using these techniques are that, first, up to about twice normal speeds, comprehension (understanding content) and intelligibility (ability to identify isolated words) may improve (Arons, 1992) once the user adapts to the task. Junor (1992) notes that in regard to comprehension it has been claimed that it is an outcome "regardless of age, intelligence, or sex" (p. 103), although his own work suggests that such a broad generalisation may be misplaced. The second finding has been that users frequently report that after a period of familiarisation they prefer TCS to normal speech (Arons, 1993). In trials by the authors using simple speeded speech (synthesised and speeded human voices) it was also found that several of the volunteers said that either normal recorded speech seemed slow and uninteresting after about 30 mins experience with speech presented at about 175% of the base rate, or that they thought the speeded presentations were in fact, at the base rate. There is also a

suggestion that attention or concentration on the task is improved under TCS presentation (Fulford, 1993).

While TCS would appear then to be a promising way to overcome the difference between reading speed and listening speed, it does not help with the problem of skimming and browsing raised by Arons in the extended quote above. Further, as Junor (1992) points out, even though a recording might be made by a skilled or professional reader, "they rarely seek personal comprehension of the text, and therefore may not see any need to adjust delivery rate in accord with the intricacy of the topic" (p. 100). While it should be possible to build into the CBL speech interface a control to allow the user to select a preferred speed (the audio recorders to generate speeded speech have such controls for example) it would still only permit something of the equivalent of the fast forward/fast rewind controls available for CBL video clips. A considerably more sophisticated interface however is being developed by Arons (1993, 1994) and this will now be discussed.

The SpeechSkimmer as described by Arons (1993, 1994) is a calculator-sized touchpad device used with a Macintosh computer, designed to give listeners the flexibility to browse, scan, and return to previously processed areas that they have with visually presented text. In Arons' words:
A continuum of time compression and skimming techniques have been designed, allowing the user to efficiently skim a speech recording to find portions of interest, then listen to it time-compressed to allow quick browsing of the recorded information, and then slowing down further to listen to detailed information (1993, p.4).

The controls on the SpeechSkimmer allow levels of browsing, and within each level the signal can be time-compressed to a greater or less extent depending on user preference. The lowest level uses the normal, i.e., unprocessed, recording which can be speeded using a combination of compression techniques. At level 2 the signal is processed by removing pauses of less than 500 msec and shortening remaining pauses to 500 msec. This technique is based on work that suggests that pauses of the order of 500-1000 msec are used by speakers to 'sign post' boundaries. These are termed "juncture" pauses (Arons, 1993). Pauses of less than 500 msec on the other hand ("hesitation" pauses), appear not to signal junctures and can be removed. Level 3 is also based on the premise that pauses carry significant information and in this case

pauses of 900 msec and more are assumed to signal major shifts, as happens when a speaker moves to a new topic or pauses for audience reaction. At this level the listener using SpeechSkimmer hears the next few seconds of the file following each pause of more than 900 msec. A unit of 600 msec of silence is inserted between each segment as a signal that the message is about to switch to a new segment. At this level then the listener can, in effect, jump along the recorded speech

message to hear a fragment of the start of (assumed) new sections. When a desired fragment is detected the user can then switch to a lower level and/or another compression level to listen in more detail. A fourth level is based on analyzing pitch, or intonation, as it has been found that new topics also tend to be signalled by a rise in pitch as well as by longer pauses (Arons, 1994).

The SpeechSkimmer has been designed to be used with the computer screen blank as Arons has in mind uses such as reviewing a lecture, listening to a backlog of voice mail, or locating a certain section of the discussion at a meeting (1993, p.1). His idea is that the device should not require any visual control so that users can use their eyes for other concerns. He notes however that the system could be redesigned with the controls on screen and operated by a pointing device such as a mouse, and suggests a possibility such as the SpeechSkimmer being "used to skim through the audio track while the related video images are synchronously displayed" (Arons, 1994, p.137).

It thus would have an apt application in the example briefly noted above of the CBL project involving presentation and discussion a series of art works, as the design in that project assumes that students will often return to a given screen to review aspects not fully comprehended. Their task at that point could be to reinforce key points made in the voice-over (say, a level 3 SpeechSkimmer transit), to review the commentary (say, at level 2, moderate compression and hesitation pauses deleted), or to quickly locate a particular point (level 3) and then listen in detail (level 1, little or no compression, all pauses included).

The concept of the artist or a critic talking about the work, using the richness of speech, is clearly more attractive than asking students to read, on screen, a text version of such a discussion. It opens the possibility that students could use the audio by itself, in conjuction with SpeechSkimmer, once they are familiar with the graphics and do not need to repeat that aspect of the program - the screen then being used as a word processor for writing tasks associated with the program.

So far we have discussed mostly only the technical aspects of audio and CBL. Many questions can be asked about the effectiveness of using speech rather than text, or, combining text and speech. Does the use of a variety of input modalities enhance the effectiveness of CBL? Some evidence has already been cited that TCS comprehension is comparable to that of reading. Clark and Craik (1992) in a substantial review note that few studies have compared the possible multiplicative effects of media in combination versus any of the individual modes used alone to teach the same material. They also note that after some two decades of work on dual coding theory (Pavio, 1969) that many elements remain controversial.

It also seems that there are considerable individual differences in

preferred mode and designers with few storage concerns (as is the case with CD-ROM or videodisk) sometimes provide both, leaving the choice to the user. On the latter point it is noted that while some programs provide speech tracks which mirror the text, others use different scripts for text and speech.

Of course, it is most probable that it is not so much the mode per se as much as how a given mode can shape and influence the quality of interactions and learning transactions asked of the learner using the program. It will be recalled that this was Reinker's conclusion from his review of speech and text and it is a point made by many writers on CBL (eg., Laurillard, 1993).

If the SpeechSkimmer is successively developed to a point where it is suitable for CBL and becomes commercially available it opens possibilities for more engaging interactions than is presently the case with orthodox audio, for example, in assisting students to create their own summaries, and could lead to useful advances in multimedia CBL designs.

References

Arons, B. (1992) Techniques, perception, and applications of time-compressed speech. In Proceedings of the 1992 Conference, American Voice I/O Society, September, 133-146.

Arons, B. (1993) SpeechSkimmer: Interactively skimming recorded speech. Paper presented at UIST '93: Symposium on User Interface Software and Technology. Nov 3-5, Atlanta, Georgia, USA.

Arons, B. M. (1994) Interactively skimming recorded speech. Ph.D dissertation, Massachusetts Institute of Technology.

Blattner, M.M. (1993) Sound in the multimedia interface. In H.Mauer (ed) Proceedings of ED-MEDIA 93 -World Conference on Educational Multimedia and Hypermedia. Orlando, Florida, USA, June 23-26.

Blattner, M.M. & Greenberg, R.M. (1992) Communicating and learning through non-speech audio. In A.D.N. Edwards & S.Holland (eds) Multimedia interface design in education. Berlin: Springer-Verlag. (NATO ASI Series, Vol. F76)

Chandler P. & Sweller, J. (1992) The split-attention effect as a factor in the design of instruction. British Journal of Educational Psychology, 62, 233-246.

Clark, E.C. & Craig, T.G. (1992) Research and theory on multimedia

learning effects. Berlin: Springer-Verlag.

Fulford, C.P. (1993) Can learning be more efficient? Using compressed speech audio tapes to enhance systematically designed text. Educational Technology, Feb., 51-59.

Gillingham, M.G. (1988) Text in computer-based instruction: What the research says. Journal of Computer-Based Instruction. 15(1), 1-6.

Junor, L. (1992) Teaching by tape: Some benefits, problems, and solutions. Distance Education, 13(1), 93-107.

Mills, C.B. & Weldon, L.J. (1987) Reading text from computer screens. ACM Computing Surveys, 19(4), 329-358

Laurillard, D. (1993) Rethinking university teaching: A framework for the effective use of educational technology. London: Routledge.

Pavio, A. (1969) mental imagery in associative learning and memory.

Psychological Review, 76, 241-263.

Reinking, D. (1992) Differences between electronic and printed texts: An agenda for research. Journal of Multimedia and Hypermedia. 1, 11-24.

Paper presented at 1994 Australian Association for Research in Education (AARE) Conference, Newcastle, NSW, Nov 27 - Dec 1

A sight-handicapped employee known to the authors checks his word processing files by setting his synthesiser to read back at 4 times the normal rate, that is, at over 500 words per minute.
Requires 25mHz 68030 processor. Optimised versions need a 33mHz 68040 machine. Arons (1993) notes that the algorithms run in real-time on the main processor and do not need require special signal processing hardware.