Audio Cues: Can Sound Be Worth a Hundred Words?

Jatin Bajaj[1(✉)], Akash Harlalka[1(✉)], Ankit Kumar[1], Ravi Mokashi Punekar[1], Keyur Sorathia[1], Om Deshmukh[2], and Kuldeep Yadav[2]

[1] Indian Institute of Technology Guwahati, Guwahati, Assam, India
{jatinbajajl993, akash.harlalka, kumar.ankiit7}@gmail.com, {mokashi, keyur}@iitg.ernet.in
[2] Xerox Research Center India, Bengaluru, India
{omdeshmukh, r.kuldeep}@xerox.com

Abstract. Multimedia content is increasingly being used in the context of e-learning. In the absence of classroom-like active interventions by instructors, multimedia-based learning leads to disengagement and shorter attention spans. In this paper, we present a framework for using audio cues interspersed with the content to improve student engagement and learning outcomes. The proposed framework is based on insights from cognitive theory of multimedia learning, modeling of working memory and successful use of audio in the film industry. On a set of 20 freshmen engineering students, we demonstrate that the systematic use of audio cues led to 37.6 % relative improvement in learning outcome and 44 % relative improvement in long-term retention. Post-study interviews establish that the associated students improved recall and engagement to the presence of audio cues.

Keywords: Cognitive theory of multimedia learning · Working memory · Audio cues · E-learning performance · Student retention

1 Introduction

Increasing use of technology in the educational domain has given rise to new models of learning such as flipped classrooms [1], blended learning [2] and Massive Open Online Courses (MOOC) [12]. Using multimedia educational content is a central aspect in all of these models. This has in turn led to active research in the design of multimedia instructional content [9], efficient ways of content delivery [3] and the effect of such content on student engagement and performance [11]. The aspects of design research encompass issues such as video production style, duration of a typical module, and cognitive load of the content. Content delivery is being studied to understand how best to cater to the diverse set of devices and be robust to lossless and variable-speed delivery channels. Student engagement and performance is being studied by analyzing various behavior parameters such as the time spent on lectures, interactions with peers and instructor, performance on in-video quizzes and dropout rates.

P. Zaphiris and A. Ioannou (Eds.): LCT 2015, LNCS 9192, pp. 14–23, 2015.
DOI: 10.1007/978-3-319-20609-7_2

The use of multimedia in e-learning is however heavily skewed towards video content [10]. Speech and audio content is used only for conveying information whereas the potential use of audio for associative memory is largely ignored. Several studies have shown that the attention span for an instructional video is only about 6 min [11] further highlighting the non-engaging nature of current e-content. This is also reflected in the poor course completion rates of several leading MOOCs [12].

In this paper we are interested in evaluating the role audio cues can play in improving student engagement and learning outcomes in the context of multimedia e-learning. We propose a framework for synergistic combination of audio cues and the learning content. The theoretical underpinnings of the framework are based on cognitive theory of multimedia learning, models of working memory and successful use of audio in the film industry. While audio cues have been used in communicating emotions to computers [13], for conversation visualization [14] and to aid people with disabilities [15], ours is the first attempt to study the efficacy of audio cues in an educational setting.

The rest of the paper is organized as follows. In Sect. 2 we present the theoretical motivation for the proposed framework followed by detailed explanation of proposed audio-cue setup in Sect. 3. Experimental setup, results of user studies and discussions are explained in Sects. 4–6 respectively.

2 Theoretical Underpinnings for Audio Cues

Authors in [8] have done pioneering research work in formulating the Cognitive Theory of Multimedia Learning (CTML). CTML is based on three basic principles of learning: **C1** dual channels of processing for auditory and visual information, **C2** limited processing capacity of each channel, and **C3** active coordination of cognitive processes. The theory of working memory [7] also supports dual channel by the following postulation: **M1:** human brain system has two independent but interconnected subsystems phonological loop for capturing speech-based information and visuo-spatial sketch for capturing visuo-spatial imagery. The interconnection allows for visual information to be represented as speech-like information using sub vocalization.

CTML further specifies five cognitive processes where relevant information is selected from these two channels to form two independent models of the input message followed by integration with the long-term memory (i.e., prior knowledge).

While there is extensive research study on the role speech plays in providing verbal redundancy and facilitating dual coding in visual and auditory channels in the context of multimedia educational content [5] there is no study on using audio cues to facilitate coherent binding of information that is spread over time: an important aspect for active coordination of cognitive processes. Authors in [10] surveyed 12 award-winning instructional software products and came to the conclusion that sound is used largely for instructional purposes and not to drive any associative learning.

Authors in [6] propose a Structured Sound Function (SSF) model for use of sound in instructional content. Specifically, they mention that when sound is assigned to a visual event, it should serve one of the five purposes: **S1:** connect to a past/future event, **S2:** present a point of view, **S3:** establish a place, **S4:** set a mood, and **S5:** relate to a character.

One domain where sound is used very effectively is the film industry: be it to enhance the narrative, to immerse in an illusion, to accentuate a mood, or to add continuity across a number of temporal events. Authors in [4] make four recommendations from the film industry for using sound in e-learning: **F1:** consider sound's use from the start of the design process, **F2:** identify key storytelling elements to be amplified by sound, **F3:** capitalize on the way people listen to sounds, and **F4:** be systematic about how sounds are incorporated. For F3, the authors suggest that the four types of listening modes should be utilized: **L1:** reduced, where listener only pays attention to the main qualities of the sound and incurs minimal cognitive load, **L2:** causal, where the sound is associated with a descriptive category that could have likely cause the sound, **L3:** semantic, where the meaning behind the sound is decoded (e.g. Emotion in the spoken message), and **L4:** referential, where the sound evokes image(s) of familiar things. L4 is particularly useful in the context of dual-channels of information coding (C1).

Based on these findings, we build our framework as follows:

(The corresponding primary findings are mentioned in the parenthesis.)

(a) Identify important learning concepts to be retained based on applicability in exams and future course work [F2].
(b) Study the coursework to identify which concepts talk across a multitude of lectures and hence need a common binding [S1, F1].
(c) Identify concepts which, while not central, are good-to-know and can be reinforced through reduced listening [C1, C2, L1].
(d) Identify concepts which tend to occur together and/or can be most confusing and hence should use audio cues which evoke very different associations [C2, L3, L4].
(e) Identify events which can easily be connected with a visual image and identify corresponding sounds that can evoke these images [C2, L4, M1].

It is important to base the usage of audio cues in educational multimedia content on strong theoretical foundation as inappropriate usage may not only not enhance learning but may act as one of the detriments to learning [4].

3 Audio-Cue Framework Applied to a Narrative

We are in the process of formulating a systematic audio-cue framework for two data-intensive semester-long eleventh grade courses on Chemistry (which includes a list of chemical reagents along with their reactionary properties towards each other) and History (which includes major battles or events, the corresponding timelines, important parties involved and the outcome) working closely with a local high school administration. But to evaluate the efficacy of audio cues in learning outcomes and retention, we first built the framework in the context of an imaginary narrative that draws on basics of Mathematics, Physics and Chemistry. We also plan to use the insights generated from this study on the narrative to further influence the audio-cue framework formulation for the courses.

Here is a brief description of the narrative: After a long party, Marc, the main character, has difficulty falling asleep and is hallucinating. In each episode of hallucination, he goes on a new expedition with his friends: On one such expedition he experiences zero-gravity free fall; on another, he has to navigate through a castle; and in yet another, one of his friends is bit by a snake and the team has to create an antidote through a combination of chemicals (purely fabricated combination using the products commonly found in a kitchen whose combination can create effects akin to those created by chemical reactions. Hence creating situation similar to a chemical reaction (purely fabricated combination) while not creating any bias.

The narrative is divided into two different videos (part-1 and part-2) to simulate two lectures on the same topic. Each video is about 4 min 30 s long in accordance with the average attention span of video lectures [11]. Using the design principles mentioned in the previous section, we first identified the highlights of the narrative that we would like the students to remember and based on the kind of information or event, the sound cues were pre-cued/post-cued or played simultaneously.

- Pre-cued: To hint listener about a forthcoming event.
- Simultaneous: to add more emphasis and clarity to an ongoing event by establishing the place or the activity through the sound clip running in the background.
- Post-cued: To emphasize on the information that was just narrated and shown on the screen e.g. names of people, numbers, scientific term.

Accordingly we formulated appropriate post-video questions. Two important characters in the narrative were assigned two distinct audio cues (high pitch vs. low pitch), each important event location in the hallucinations was assigned an appropriate audio effect (e.g., a long hallway had 'echoes', the free-flow event had 'high speed wind noise'), important concepts were preceded by a consistent audio cue (e.g., every instance of acidity was associated with a consistent 'burp' cue). Appropriate chemical reactions were peppered with corresponding likely audio effect (e.g., to evoke images of strong repulsive smell, the audio of an uncomfortable cough was used, adding water to a chemical compound had the sound of 'water flowing through a tap'). The quiz corresponding to the part-1 video had 10 questions: 5 had highlighting specific audio cues associated with the likely answers whereas the other 5 had no cues. Similarly, the quiz corresponding to the part-2 video had 10 questions, 6 of which had specific associated audio cues. The questions were not in the chronological order of the narrative to avoid likely temporal bias.

In the experiments described here we used the audio narration of the above narrative with synchronized scrolling of the corresponding text on the video (similar to closed-captioning but on full screen). We used such a setup for our current experiments for two main reasons: (a) make optimal and synergistic use of the dual channels [C1 and C2], and (b) in practical e-learning situations, we have very little scope of adding relevant visual content to a video whereas audio cues can be inserted with relative ease. Two versions of the videos were created: the Narration Only (NO) version had no audio cues whereas the Audio-Cues (AC) version had audio cues interspersed with the audio-visual narration. The visual message was exactly the same in both the version.

4 Experimental Setup

We recruited 20 college freshmen students (18−20 years old) for this study and split them into two equal groups: the Narration-Only group (NO) and the Audio-Cues (AC) group. The NO group was shown the NO version of the videos whereas the AC group was shown the AC version of the videos. The students were informed about the end-of-the-video quiz. They were also told that they have to answer more than eight questions correctly, failing which they will have to watch the video and give the quiz again. The part-2 video was shown only after the students passed the part-1 quiz. Each student was presented the video on a 15-in. laptop computer along with a pair of earphones in a quiet room. The students were asked to complete each attempt of the part-1 quiz in five minutes and each attempt of the part-2 quiz in 10 min.

The questions were a mix of multiple choice, one word or one sentence, explanatory types. In the sentence long questions, marks were solely awarded on the mention of one or two keywords. For easy recall, the questions were asked in the chronological order (the order in which the events occurred in the narration).

The students were called in for a surprise quiz five days after they watched the part-2 and were given the same quizzes as earlier. This was done to evaluate their long-term retention (as per the forgetting curve, humans tend to forget nearly 80 % of the content within 4−5 days in the absence of any revision). At the end of this surprise quiz, we had an informal exit-interview with the students to gather their feedback on the entire process.

5 Results

Figure 1 shows the average student performance across multiple attempts for the NO and the AC group for part-1 and part-2 of the video narratives. As expected the performance improves after every attempt. Students in the AC group show much better performance than those in the NO group in the first attempt itself. The average performance of AC group students combined across the two parts of the videos in the first attempt is 11.7 whereas that of the NO group students is 8.5, which shows that audio cues lead to a 37.6 % relative improvement in learning outcome. Moreover, students in the AC group reach the passing criterion much quicker than those in the NO group: For example, only 1 student in the AC group needed three attempts while there were 3 such students in the NO group for part-1 video. In the case of part-2 video, 5 students in the NO group needed 3 attempts to reach pass the quiz whereas all the students in the AC group passed the quiz in the second attempt. The significant jump in performance of AC students in the part-2 video as compared to that of the NO students could be because of the multiplicative effect where the students ‘learn’ to optimally utilize the audio cues. Further experiments are needed to validate the existence of such an effect.

As mentioned in Sect. 3, answers to only about 50 % of the questions had corresponding specific audio cues. To test whether the improvement in the AC group performance was only due to these ‘audio-cued questions’, we analyzed the question-wise performance of the students in both the groups and across the two parts of the video. Figures 2 and 3 shows the average number of correct answers by the students of

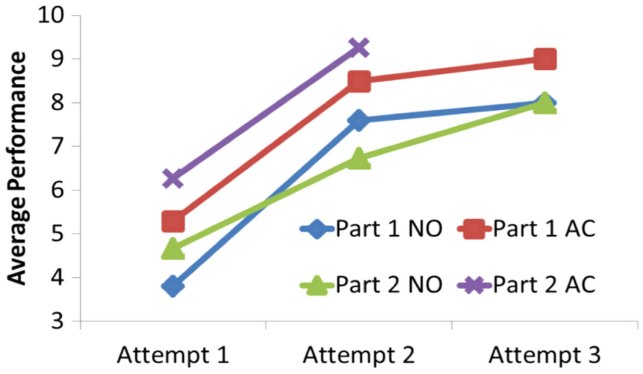

Fig. 1. Performance of students in the NO and the AC group on part-1 and part-2 video across multiple attempts.

NO and AC groups on audio-cued and non-cued questions for part-1 and part-2 videos. Notice that while the AC group performs better on the audio-cued questions, their performance on the non-cued questions is also better than that of the NO group. This leads us to believe that the audio-cuing phenomenon has a positive effect on the overall learning experience, which we term as the spread effect.

Figure 4 compares the average long term retention performance of students in NO and AC groups in terms of number of correct answers to the two quizzes. While the NO group answers about 12 questions correctly, the AC group answers about 17.3 questions correctly demonstrating a relative improvement of about 44 % in long term retention due to the use of audio-cues.

Figure 5 compares the performance of the NO and the AC group students for each of the 10 questions in the part-2 video quiz across multiple attempts. Each question has two horizontal bars. The bottom bar shows the performance for the NO group and the top bar show the performance for the AC group. The questions with

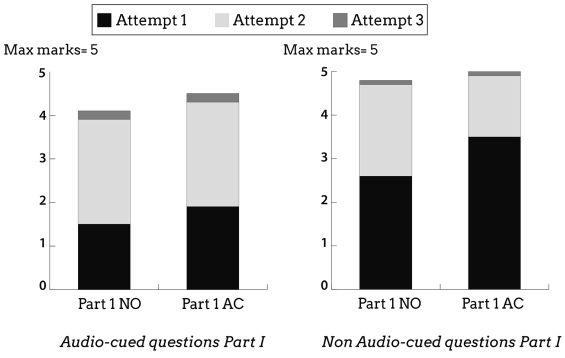

Fig. 2. Average number of correct answers by the students of NO and AC groups across multiple attempts to (i) audio-cued questions in part-1 video (5 questions), (ii) non-audio cued questions in part-1 video (5 questions).

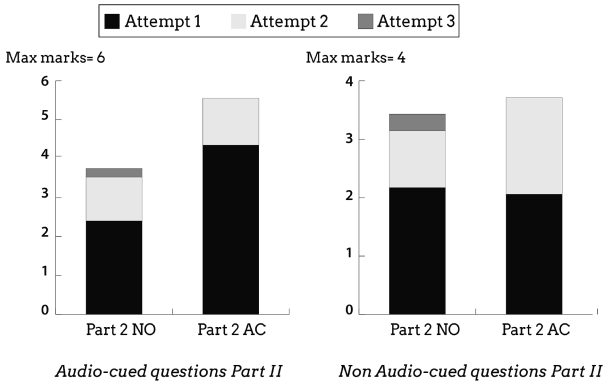

Fig. 3. (iii) audio-cued questions in part-2 video (6 questions), and (iv) non-audio-cued questions in part-2 video (4 questions).

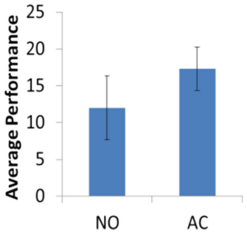

Fig. 4. Average long-term retention performance of students in NO and AC groups in terms of number of correct answers to the quiz questions.

specific audio cues are indicated separately on the y-axis. Performance on all the questions is improved by the use of audio-cues except for one question (question 3 from the top). This question asks the subjects to mention all the four ingredients used to create the antidote for snake bite. This can likely be attributed to the high cognitive load of recollecting multiple names. The other question with interesting behavior is the top most question which has poor performance by the NO group but substantial improvement by the AC group. This question asks the user to 'name the person who sobbed when Suzanne was bit by a snake'. The NO video only mentions the name 'Johanna' whereas the AC video has a special audio cue assigned to this character as well as plays 'girl sobbing' sound. These multiple cues helped the AC group to easily recollect the correct answer.

Similarly, there was one question in the quiz of part-1 video to test if students would latch on to secondary details that are provided only through audio-cues. The question was 'what is the castle door made of?'. The narration had no mention of the type of the door and the audio cue had a creaking sound of a rusty iron door. None of the NO group students could answer the question correctly, but 60 % of the AC group students caught on to the audio cue in the first attempt.

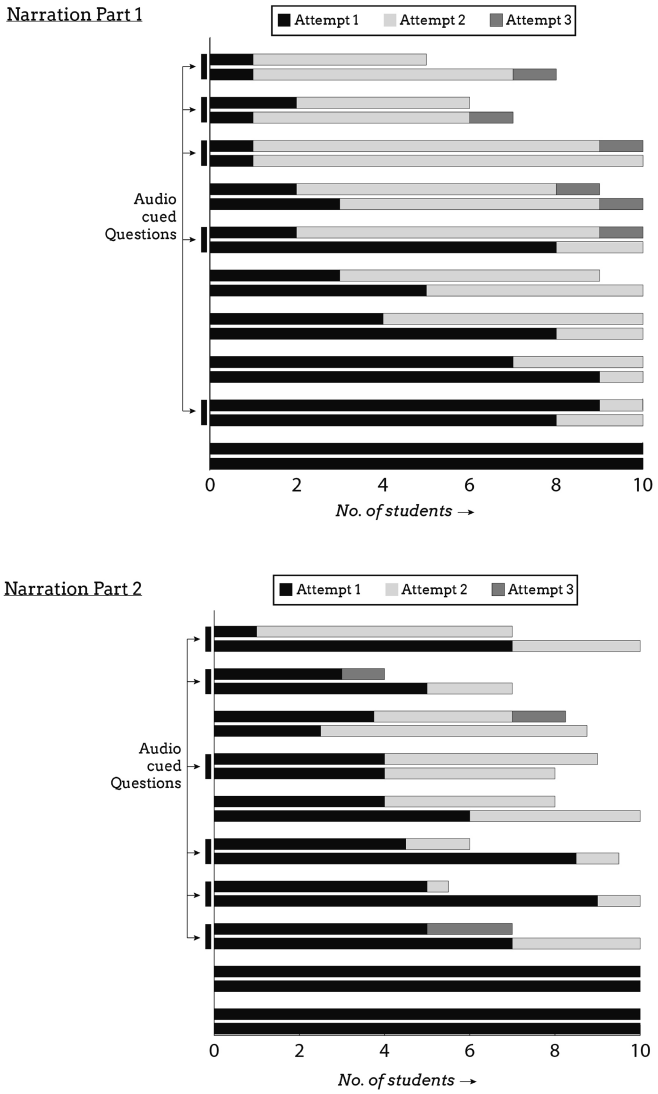

Fig. 5. Performance of the NO and the AC group students for each of the 10 questions in the part-1 and part-2 videos across multiple attempts. For each question, the bottom horizontal bar represents the performance of the NO group and the top bar represents the performance of the AC group.

This has significant ramifications for instructional multimedia content in that subtle associations (such as reminder of where a historic battle was fought) can be effectively reinforced through appropriate audio-cues rather than having to spell those out in the main lecture. Finally, as we conducted post-study interviews with the students, almost all the AC students attributed their high performance to the audio cues. Several of them

mentioned that it took them a while to make the connection that the audio cues are reinforcing the important aspects but once they understood it, the part-2 video was much easier to understand and remember.

Subjects with effects said that they could visualize the story by identifying the sound cues and the Left or Right sound channel activation. For example sound of crickets and sound coming from the Left channel helped them in creating an imagery of the path took by Marc and answer the question- "Which direction did Marc turn in to switch on the light?". A few subjects also said that at times it was difficult for them to concentrate on both narration and sound cues running simultaneously. All subjects needed help to understand the meaning of the word 'hallucination'. Students with effects could better relate and understand the situation of hallucination in the narrative. Subjects with effects could very well comprehend the situation of closed well and concept of echo in their first attempt. When asked to identify the character sobbing, subjects replied that the answer (Johanna - a girl) was obvious; however, they could not tell why! The gender of the character was implicitly embedded in the listener's memory by the effect of audio cues.

6 Discussions

In this work, we presented our initial results on a framework for the use of audio-cues in educational multimedia content. The framework is based on insights from cognitive theory of multimedia learning, modeling of working memory and successful use of audio in the film industry. We demonstrated that a systematic use of audio cues can indeed improve student performance and engagement. We are currently formulating the use of audio cues for two semester-long courses: Chemistry and History. We are conducting in-field studies to collate sounds that naturally occur in classrooms and to identify their associations (e.g., sounds corresponding to tapping of the chalk on the board or slapping the duster on the table to capture students' attention, teacher's footsteps for close monitoring, student murmur for classroom discussion, etc.).

Further research on the retention capabilities of students shows that the capability to recall content is a complex phenomenon. Most students may not be able to recall the content visually presented to them but audio cues may serve as a brilliant medium to provide hints, we can call them audio anchors. Audio anchors may serve as a great help to students who are not able to answer the question presented to them in the right away but are able to recall the answer once given a small hint or a head start in words. The audio anchors will serve as an effortless memory anchor and will help in recalling the content.

Our initial findings on use of audio cues for improving learning outcomes and student retention have been encouraging. To conclude, our study shows that the answer to the question raised in the title of this paper is in the affirmative!

Acknowledgments. We gratefully acknowledge Safinah A. Ali for all her help and assistance in building our audio cue framework.

References

1. Tucker, B.: The flipped classroom. Educ. Next **12**(1), 82–83 (2012)
2. Horn, M.B., Staker, H.: The rise of K-12 blended learning. Innosight Institute (2011)
3. Zhao, X., Okamoto, T.: Adaptive multimedia content delivery for context-aware u-learning. Int. J. Mob. Learn. Organ. **5**(1), 46–63 (2011)
4. Bishop, M.J., Sonnenschein, D.: Designing with sound to enhance learning: four recommendations from the film industry. J. Appl. Instr. Des. **2**(1), 5–15 (2012)
5. Mayer, R.E., Moreno, R.: Aids to computer-based multimedia learning. Learn. Instr. **12**(1), 107–119 (2002)
6. Mann, B.L.: The evolution of multimedia sound. Comput. Educ. **50**(4), 1157–1173 (2008)
7. Baddeley, A.: Working memory. Science **255**(5044), 556–559 (1992)
8. Mayer, R.E.: Cognitive theory of multimedia learning. In: The Cambridge Handbook of Multimedia Learning, pp. 31–48 (2005)
9. Clark, R.C., Mayer, R.E.: E-learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning. Wiley, San Francisco (2011)
10. Bishop, M.J., Amankwatia, T.B., Cates, W.M.: Sound's use in instructional software to enhance learning: a theory-to-practice content analysis. Educ. Tech. Res. Dev. **56**(4), 467–486 (2008)
11. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: an empirical study of mooc videos. In: Proceedings of the First ACM Conference on Learning@ scale Conference, pp. 41–50. ACM (2014)
12. http://www.insidehighered.com/news/2013/03/08/researchers-explore-who-taking-moocs-and-why-so-many-drop-out
13. Sebe, N., et al.: Emotion recognition based on joint visual and audio cues. In: 18th International Conference on Pattern Recognition, ICPR 2006, Vol. 1. IEEE (2006)
14. Bergstrom, T., Karahalios, K.: Seeing more: visualizing audio cues. In: Baranauskas, C., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4663, pp. 29–42. Springer, Heidelberg (2007)
15. Velasco-Álvarez, F., Ron-Angevin, R., da Silva-Sauer, L., Sancha-Ros, S., Blanca-Mena, M.J.: Audio-cued SMR brain-computer interface to drive a virtual wheelchair. In: Cabestany, J., Rojas, I., Joya, G. (eds.) IWANN 2011, Part I. LNCS, vol. 6691, pp. 337–344. Springer, Heidelberg (2011)