Journal of Pedagogical Research
Volume 6, Issue 3, 2022
https://doi.org/10.33902/JPR.202214591

Research Article

The voice effect in multimedia instruction revisited: Does it still exist?

Nazmi Dinçer

National Defence University, Turkey (ORCID: 0000-0002-2901-5367)

The voice effect or principle assumes that people learn better when they are exposed to multimedia instruction that includes a human voice rather than a machine voice. This investigation reconsiders the voice principle by investigating the relationship between learning outcomes and mental effort of the learners. Text-to-speech (TTS) engines have improved dramatically since the early 2000s, thanks to technological advancements. The researchers employed sophisticated TTS engines in a pretest-posttest design to analyze the various voice types (human voice, traditional machine voice, and modern machine voice). The results indicated that the progress in TTS technology enabled to generate real-life-like voices, and therefore no significant difference was observed between the modern and human voice. Furthermore, the participants' cognitive load was consistent with the findings of the learning outcomes.

Keywords: Voice effect; Voice principle; Synthesized voice; Text-to-speech; Multimedia instruction

Article History: Submitted 31 January 2022; Revised 12 May 2022; Published online 26 May 2022

1. Introduction

Multimedia instruction is critical for improving teaching and learning since it allows for deeper learning when the instructional contents are properly structured. During the pandemic lockdown, the shift from face-to-face education to online education has paved the way for designing multimedia instruction more often than usual. Educators, however, should follow evidence-based principles when designing educational resources that combine auditory and visual inputs, such as videos, animations, webpages, or apps. While defining the cognitive theory of multimedia learning, researchers established research-based multimedia principles to assist designers in the planning and implementation of instructional materials (Atkinson et al., 2005). These are primarily stemmed from the effort over how to create the best multimedia instruction.

It is widely believed that the principle mentioned above will serve as a guide for improving the quality of multimedia materials (Sweller, 2010). However, it is essential to note that some of these principles put forward in the early 2000s when the technology was not advanced enough to test each one of them, have not been investigated thoroughly in experimental settings despite the well-examined cognitive impacts (Craig & Schroeder, 2019). The voice principle or voice effect is an excellent example of this kind of situation. To briefly explain, the voice principle is based on the

Address of Corresponding Author

Nazmi Dinçer, National Defence University, 34149 Yeşilyurt, İstanbul, Turkey.

✉ ndincer@hho.msu.edu.tr

How to cite: Dinçer, N. (2022). The voice effect in multimedia instruction revisited: Does it still exist?. *Journal of Pedagogical Research, 6*(3), 17-26. https://doi.org/10.33902/JPR.202214591

premise that people learn better when the natural human voice is used rather than synthetic computer-generated voices (Mayer, 2014). The voice principle implies that the human voice is the natural and socially appealing one for people. It is widely claimed that the integration of synthetic voices into educational contexts impedes comprehension and raises the cognitive load of the learners. In line with this claim, five experiments were conducted by Mayer (2020) over different studies in which there were comparisons between machine voice and human voice (Atkinson et al., 2005; Mayer et al., 2003; Mayer & DaPra, 2012). The results of these studies show that human voice is much better than synthetic voice. In addition, it was found that human voice had positive effects on retention scores and near and far transfer scores. Mayer et al. (2003), for example, showed students a 140-second narrated video of lightning formation that included spoken words. A non-personalized Russian accent speaker (unappealing and uninteresting) and a standard accent voice type were also available. Students who were exposed to the standard voice type scored higher than those who were exposed to the Russian accented voice type in the following transfer test, leading to the conclusion that human voice with distractive characteristics could be harmful to people because it reduces the learner's social stimuli. This exploration created an opportunity and motivation for other research investigations to establish precise conclusions about multimedia instruction.

Table 1
The research studies supporting the voice principle.

Research	*Topic*	*Effect size*
Social cues in multimedia learning: Role of speaker's voice (Mayer et al., 2003) (Expt. 1 and 2)	Lightning	0.90, 0.79
Fostering social agency in multimedia learning (Atkinson et al., 2005) (Expt. 1 and 2)	Mathematics	0.69, 0.78
An embodiment effect in computer-based learning with animated pedagogical agents (Mayer & DaPra, 2012)	Solar cells	0.63

The technology used to generate machine voice in these studies, however, was extremely different from the one currently available (Craig & Schroeder, 2017). The developing software engineering enables educators to make use of higher quality machine synthetic voices through text-to-speech engines. Recently, sophisticated speech engines such as Neospeech, Microsoft, and Google have been offering lifelike speech synthesis closed to natural human voice and predominantly cannot be distinguished. They are presently getting more common and often used while developing educational materials. Lately, the issue of voice effect in multimedia instruction was reconsidered through the incorporation of advanced technology (Craig & Schroeder, 2017; Craig et al., 2019; Davis et al., 2019). In these experiments, the investigators made comparisons among human voice, lower quality machine voice, and higher quality machine voice. Results of the study conducted by Craig and Schroeder (2017) demonstrated that the score of people exposed to modern computer voice was significantly better than those exposed to the human voice and traditional computer voice. The authors suggest that the voice effect or voice principle may not exist anymore as it was in the past. The reason for this claim is chiefly based on the improvement of text-to-speech technology that performs as well as the natural human voice. In a more recent study conducted by Craig et al. (2019), the trust ratings of people were the highest when the human voice was used, while the scores did not differ significantly among the groups. Furthermore, Davis et al. (2019) suggest that categorizing the voice type as human or computer may not be enough to understand the dynamics behind the voice effect despite the findings indicating that modern computer voice is as effective as a human voice. In this respect, while replicating the previous studies examining the different voice types, they also investigated the effects of strong and weak prosodic voices on people. The results showed that the human weak prosodic voice performed higher compared to the modern computer voice while the human strong prosodic voice scored lower against advanced machine voice. Therefore, the recent studies

contradict the voice effect phenomenon emerging in the early 2000s. But there is not enough evidence to support the idea that there is no more voice effect. In order to find out whether the modern machine voice performs as well as a human voice, there should be further studies conducted in experimental settings with a different group of participants and multimedia materials.

Two theoretical perspectives are prominent to provide explanations to the claim that machine voice might be less effective than a human voice. Researchers considering the cognitive load (Paas & Sweller, 2014) propose that synthetic voices bring about extraneous cognitive load and decrease the usable cognitive capacity to incorporate the new information with the available knowledge. Furthermore, theorists considering social agency (Atkinson et al., 2005) suggest that human voice could be identified quickly due to social interaction and familiarity, and in this way, active learning occurs. In this section. The above-mentioned theories have been discussed and examined briefly by reviewing relevant studies which have focused on the voice effect.

1.1. Cognitive Load Theory

The ultimate purpose of cognitive load theory is to improve learning higher cognitive tasks by transforming scientific knowledge into cognitive structures in an ordered manner (Paas et al., 2003; Sweller et al., 2011). According to cognitive load theorists, a cognitive architecture of a human being is divided into three parts: a limited working memory, a limitless long-term memory, and schemas that work to organize in long-term memory (Sweller, 2011). While receiving new information, three types of cognitive load are invested in applying mental effort. First, extraneous cognitive load is described as an additional burden on the brain since the design and application of materials may not be appropriate or pedagogically well planned (Paas et al., 2003). In this respect, it is believed that people need to devote more cognitive effort when learning with machine voice or voice in poor quality rather than the human voice and, thus, impeding the learning process (Sweller et al., 1998). Recent studies, however, have been contradicting this claim through the integration of advanced text-to-speech engines. Second, intrinsic cognitive load stems from the complexity of the instructional materials (Van Merriënboer & Sweller, 2005). If the learning process is well designed for students' age, needs and level, intrinsic cognitive load is expected to be low. A task with many unknowns, on the other hand, demands a higher intrinsic cognitive load. Third, germane cognitive load is defined as the use of cognitive resources for productive purposes. There is somehow complex interaction between cognitive load and mental processing as working memory has limited processing, but the additional cognitive load is required for intrinsic and extraneous types, which implies the dependent on particular variables (Leppink & Heuvel, 2015).

The component of voice is a crucial attribute in theories investigating cognitive load. Based on the assumptions of cognitive load theory, human brain receive instruction in two different channels as verbal and visual before information processing initiates, and the capacity is relatively limited. Therefore, synthetic machine generated voice types might increase the extraneous cognitive load of those who were exposed to multimedia instruction or engaged in a multimedia material as it appeals uninterested and distracting without sufficient social cues (Wouters et al., 2008). This concern, however, has been believed to be eliminated in the last ten years with the advanced voice engines. The study by Mayer and DaPra (2012), finding no significant differences between modern voice types generated through new voice engines and a recorded human voice, constituted preliminary evidence that synthetic machine voices are no longer influential in the use of capacity for the extraneous cognitive load. Similarly, Another study by Craig and Schroeder (2017) denoting no significant difference between modern machine voice, traditional machine voice and a recorded human voice, showed that the voice types might not diminish or raise the cognitive capacity by forming extraneous load in the information processing. Davis et al. (2019), on the other hand, emphasizes that cognitive load should be investigated within three distinct components as intrinsic cognitive load, germane cognitive load, and extraneous cognitive load in order to see the full picture regarding the relationship between voice types and cognitive load.

They found no significant difference between intrinsic and extraneous cognitive load in human strong-prosodic voice, human weak prosodic voice and modern machine synthesized voice while germane cognitive load was higher than modern machine synthesized voice in human weak prosodic voice type.

1.2 Social Agency Theory

Learning is defined as a social activity (Bandura, 1969) and thus, integrating social elements into multimedia could promote deeper learning (Richards & Dignum, 2019). Pedagogical agents are usually utilized to augment the learning environment as a sign of social cues (Wang et al., 2018). Designing the learning task with agents, adding recorded human voice for narration in multimedia materials, bring about familiar experience as talking to a human (Mayer et al., 2003). Voice effect implies that when preparing animation or videos, a human voice should be preferred instead of a synthetic voice. For instance, Atkinson et al. (2005) set forth that employing videos designed with a virtual human using a human voice performs better compared to the agent with machine voice. In brief, social agency theory is based on the premise that using social cues in multimedia learning improves the quality of education and increases retention (Dinçer & Doğanay, 2017). Despite recent research indicating pedagogical agents not liked by learners impede learning (Domagk, 2010), it is widely admitted that the use of PAs facilitates authentic learning experience and supports real-life like education with the growth of artificial intelligence (Johnson & Lester, 2018; Kim & Baylor, 2019).

Social agency theory is a collection of ideas that explain how social factors influence multimedia learning (Linek et al., 2010). Firstly, cues including voice or image of presenters integrated into a multimedia lesson might act as a social stimuli. In particular, the degree to which cues convey social notions can vary; for example, it has been found that a machine-synthesized voice does not carry the same degree of social cues as the human voice (Mayer et al., 2003). In a recent study, Liew et al. (2020) found that voice enthusiasm is a factor to influence the amount of social cues rather than mere differentiation between human or non-human voice types. While enthusiastic voice has shown that the enthusiastic voice prompted more affective social ratings, the calm voice led to a higher germane load. Additionaly, the embedded social elements raises the impression that multimedia instruction involves social communication instead of one-way passive lecturing. This could prompt learners to exert the same effort as interacting with human. In this regard, Nass and Brave (2005) highlight that as the number of social cues increases in media, people are more liable to respond socially. Furthermore, the embedded social cues in media are believed to promote seamless and meaningful message to audience (Grice, 1975). Last, learners exert more effort to comprehend the content with the integrated social elements and accordingly lead to beter retention and transfer performance (Lawson & Mayer, 2021).

Concerning gender issue in voice, Linek et al. (2010) demonstrated that the female voice was more successful at gathering learners' attention and retention socres compared to the male voice. The social ratings of the female voice, additionally, was found to be more assertive and appealing.

In terms of the voice accent, a study on Russian language by Ahn (2010) showed that learners' thoughts, values and experiences are influential in their favorable voice accents. This, also, impacted their preferences on what types of accents work socially better and capture their attention. In the same vein, a more recent study by Rey and Steib (2013) illustrated that learners are more engaged in a multimedia message when exposed to the dialects which they are familiar with the characteristics of. They conclude that the sense of familiarity emerging out of the dialect that resembles the listeners social attributes might carry social cues to learners.

In short, when the above-mentioned studies are taken into consideration, it could be said that the advancement in text-to-speech engines might refute the argument that the human voice is the best of all. The studies using the modern machine voice, however, is quite limited due to the paucity of research that explore voice effect in different multimedia settings (Liew et al., 2020). This study,

thus, attempts to reconsider the issue of voice effect by implementing both modern and traditional machine voices against human voices. In this study, we examine the following research questions:

RQ 1) To what extent does voice type integrated into multimedia instruction impact learning outcomes?

RQ 2) To what extent does the voice type integrated into the multimedia instruction impact cognitive load?

2. Method

In the current study, a pretest-posttest research design was employed with a convenience sample of 51 undergraduate students. Three different groups ($n = 15, n = 17, n = 19$ accordingly) participated in the study voluntarily. The students were the 3rd Grade male students studying engineering in a state university in Eastern Europe. The age range was 20 to 23. Participants viewed the video presentation using a traditional text to speech engine (n = 15), a modern text-to-speech engine (n = 17) and human voice (n = 19).

2.1. Learning Materials

The multimedia material adapted from YouTube is a 4-minute video about collision avoidance precautions in aviation. This content involves a 670-word passage. The video is redesigned with different voice types through text-to-speech engines as modern and traditional to present in classes. In this stage, it was important to keep the temporal contiguity of the video in order to organize the speech and the presentation at the same time in harmony. In addition, there was no inflection or wavy production during the narration. The modern text-to-speech software was Microsoft Azure, and the voice was *Guy* while Google Cloud Speech Synthesis was preferred for the classic type and the name of voice was *US-Standard-B*. Lastly, a male human voice was already in the video selected for the presentation. The original video was in standard American English, and thus the same accent was implemented for other voice conditions.

Figure 1
Screenshot of the video used in this study

2.2. Instruments

2.2.1. Pretest-posttest measure

The same test was used before and after the treatment. A general knowledge test regarding collision avoidance precautions was used as pre-and post-test measures. The test includes 10 items multiple-choice questionnaire. An example question is: "Which one is the first way when a threat is

discovered?". The answers were given as peripheral vision, contact with air traffic control, movement, using short. Participants received 10 points for each question and 100 points for the whole test.

2.2.2. Cognitive load measure

The subjective mental effort scale developed by Paas (1992) was employed as a measure of students' cognitive load. The scale contains only one item and 9-point answers ranging from very, very low effort (1) to very very high effort (9). The item is: "In studying the preceding video, I invested." It was applied at the end of the learning phase. The scale is one of the most frequently implemented scales in order to measure mental effort (Paas et al., 2003), and the coefficient reliability is found .90 (Paas, 1992).

2.2.3. Procedure

The flow of the procedure was the same for all groups. First, participants were provided a pre-test about the content of the video to understand to what extent they have knowledge about the target topic. They have then watched the video lasting for 4-minute. It was followed by the post-test. Then, the mental effort question was given as the final assessment.

3. Results

Prior to utilizing inferential statistics, a Levene's test was implemented to ensure that the data complies with the standards of homogeneity. The results showed that no heterogeneity was found with the data at $F(2,47) = 1,80$, $p = 0,177$. Therefore, the data was suitable for inferential statistics.

Table 1
Means and Standard Deviations of Pre-Test and Post-Test Scores

Voice Condition	*Pre-Test*			*Post-Test*		
	N	Mean	SD	N	Mean	SD
Traditional	15	26.00	9.103	15	42.67	10.998
Modern	17	23.53	11.147	17	54.12	12.776
Human	18	25.00	14.653	18	60.56	8.726

3.1. Learning Outcomes

Analysis of variance (ANOVA) tests were conducted to explore the participants' learning outcomes and to find out whether there were any differences between groups. Table 1 illustrates the means and standard deviations of pre- and post-test scores of learning outcomes. While there was not a significant difference in the pre-test, ANOVA tests revealed a statistically significant difference among groups in the post-test ($F(2,47) = 11.12$, $p = .00$). In addition, a Tukey post hoc test revealed that participants' scores were statistically significantly higher scores with a large effect size after viewing the video with the human voice (m = 60.56, $p = .00$, d = 0.90) and the modern machine voice (m = 54.12, $p = .013$, d = 0.75) compared to the traditional machine voice (m = 42.67). Last but not least, no statistically significant difference were found between the human voice and modern machine voice ($p = .200$)

3.2. Cognitive Load

ANOVAs were performed on the cognitive load to analyze how different voices influence the mental effort of the participants after the treatment. Table 2 shows the means and standard deviations of cognitive load scores. ANOVA results indicated that there was a significant difference among the groups ($F(2.47 = 3.71$, $p = .032$). A Tukey post hoc test revealed that participants learning through a video embedded with a traditional text-to-speech generated voice invested in significantly higher effort compared to human voice (m = 5.39, p=.40), but there was not a significant difference observed between traditional and modern machine voices ($p = 0.75$). In addition, the difference between human and modern voices did not differ remarkably ($p = 0.964$).

Table 2
Means and Standard Deviations of Cognitive Load Scores

Voice Condition	*N*	*Mean*	*SD*
Traditional	15	6.80	1.42
Modern	17	5.53	1.90
Human	18	5.39	1.42

4. Discussion

The current study examined the impact of the voice condition integrated into the multimedia instruction on the participants' learning outcomes and reported mental effort. The findings suggest that the effect of machine voice depends on the type of technology used to generate the voice. While statistically significant differences were observed on learning outcomes between traditional computer-synthesized voice and human voice, the results did not differ remarkably between modern computer-synthesized voice and a human voice. Those receiving the instruction with a human voice condition could not perform better than those in the modern voice condition.

The results are somewhat contradictory to the findings of Atkinson et al. (2005), Mayer et al. (2003), Mayer (2014), which claim that the human voice outperformed synthesized voices with regard to learning measures. These studies have also explained the advantage of the human voice by putting forward that the machine voice requires more effort to comprehend and thus decreases the amount of capacity for cognitive processing. While the conventional voice condition supports theis claim, the modern voice did not yield the same results. Moreover, the results are in accordance with previous research that suggests the voice effect or voice principle phenomena may no longer exist in the modern voice (Craig & Schroeder, 2017, 2019; Mayer & DaPra, 2012).

Virtual humans are used in multimedia learning to introduce social affordances (Mayer et al., 2003). As a result, one of the aspects determining social presence and retention achievement is the voice of a virtual human. The social agency theory and its effects on the voice effect are supported by this argument. Furthermore, productive learning performance is a strong predictor of socially designed material (Lawson & Mayer, 2021). Unlike the studies conducted in the early 2000s, the advancement of text-to-speech engines allowed to generate human-like voice which is hard to distinguish. The improved voice quality in text-to-speech engines prompted learners to view a multimediage message as socially more. The data is also consistent with the social influence model (Blascovich, 2002) in that the social elements could be increased when the advancement of technology rises and accordingly this influences students' retention and transfer performance. At this point, additional research might be needed to understand the attributes that are influential in a sense of social familiarty with the voice types. Furthermore, it is noteworthy to see that while previous studies involve visually represented virtual human (Craig & Schroeder, 2017; Mayer & DaPra, 2012), the current study provides consistent results in multimedia presentation without the physical appearance of a virtual human.

Voice appears to have a remarkable influence on the cognitive load of the participants (Akker & Cutler, 2003; Mayer, 2020) since the prosodic characteristics of the human voice could carry lexical, discourse and syntactic knowledge (Cutler et al., 1997). The cognitive load measures revealed significant differences between the traditional voice type and human voice. However, there was not a significant difference observed between the modern voice and human voice, which also verifies the explicit correlation between the cognitive load and learning outcomes. The premise that computer-synthesized voice could raise the cognitive load and leave little room for information processing is not promoted in this study with modern synthetic voice (Wouters et al., 2008). When participants invested more extraneous effort, there was less achievement. The modern voice condition, however, does not demand as much extraneous cognitive load as the traditional voice. This supports the argument that the quality of voice has a major effect on the cognitive load (Davis et al., 2019).

5. Limitation

The first potential concern of importance to the current study is the length of the multimedia video (4 minutes). Further studies might extend the time of videos to understand whether the time is an important factor. In addition, it could be beneficial to include a variety of videos on different subjects and accents. Another point is the level of the video. Although participants were capable of performing abilities in B1 level English Proficiency, the video was slightly above their capacity, which could be understood from the average post-test scores despite the significant improvement in each group. Another concern could be the sample size of the participants. The number of students in each group might not be sufficient enough to draw general implications for the field.

The other concern is the subjective scale used to determine the mental effort of the participants. Despite the satisfactory outcomes, there have always been questions whether subjective scales are precise enough to identify cognitive load compared to objective scales such as electroglottograph (EGG). Thus, it would be better for further studies to benefit from both subjective and objective scales.

6. Conclusion

The research sought to examine the voice principle with three different voice conditions and the relationship between the voice conditions and mental effort. The widely agreed assumption that the human voice provides much better narration in multimedia instruction than machine voice in terms of social influence, learning measures, and mental effort (Mayer, 2014) is rejected by the findings of the current study. In this respect, the study contributed to the literature by supporting recent studies (Craig & Schroeder, 2017; Davis et al., 2019), indicating that the so-called voice effect phenomenon might no longer occur due to the fact that technological improvement has aroused questions to the statements of early 2000s. The advanced text-to-speech engines and agent technologies have helped designers to generate real life-like materials for educational purposes.

Funding: No funding source is reported for this study.

Declaration of interest: No conflict of interest is declared by author.

References

Ahn, J. (2010). *The effect of accents on cognitive load and achievement: The relationship between students' accent perception and accented voice instructions in students' achievement* [Unpublished doctoral dissertation]. Ohio University, USA.

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, *6*(2), 81-96. https://doi.org/10.1017/S1366728903001056

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology, 30*(1), 117-139. https://doi.org/10.1016/j.cedpsych.2004.07.001

Bandura, A. (1969). Social-learning theory of identificatory processes. *Handbook of socialization theory and research*, *213*, 262.

Blascovich, J. (2002, September). A theoretical model of social influence for increasing the utility ofcollaborative virtual environments. In W. Broll (Ed.), *Proceedings of the 4th international conference on Collaborative virtual environments* (pp. 25-30). Association for Computing Machinery. https://doi.org/10.1145/571878.571883

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education, 114*, 193-205. https://doi.org/10.1016/j.compedu.2017.07.003

Craig, S. D., & Schroeder, N. L. (2019). Text-to-speech software and learning: Investigating the relevancy of the voice effect. *Journal of Educational Computing Research*, *57*(6), 1534-1548. https://doi.org/10.1177/0735633118802877

Craig, S. D., Chiou, E. K., & Schroeder, N. L. (2019, November). The impact of virtual human voice on learner trust. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 63, No. 1, pp. 2272-2276). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177/1071181319631517

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141-201. https://doi.org/10.1177/002383099704000203

Davis, R. O., Vincent, J., & Park, T. (2019). Reconsidering the Voice Principle with Non-native Language Speakers. *Computers & Education*, *140*, 103605. https://doi.org/10.1016/j.compedu.2019.103605

Dinçer, S., & Doğanay, A. (2017). The effects of multiple-pedagogical agents on learners' academic success, motivation, and cognitive load. *Computers & Education*, *111*, 74-100. https://doi.org/10.1016/j.compedu.2017.04.005

Domagk, S. (2010). Do pedagogical agents facilitate learner motivation and learning outcomes?. *Journal of media Psychology*. https://doi.org/10.1027/1864-1105/a000011

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds), *Syntax and semantics* (pp. 41-58). Academic Press.

Johnson, W. L., & Lester, J. C. (2018). Pedagogical Agents: Back to the Future. *AI Magazine,* 39(2). https://doi.org/10.1609/aimag.v39i2.2793

Kim, Y., & Baylor, A. L. (2016). based design of pedagogical agent roles: A review, progress, and recommendations. *International Journal of Artificial Intelligence in Education,* *26*(1), 160-169. https://doi.org/10.1007/s40593-015-0055-y

Lawson, A. P., & Mayer, R. E. (2021). The Power of Voice to Convey Emotion in Multimedia Instructional Messages. *International Journal of Artificial Intelligence in Education*, 1-20. https://doi.org/10.1007/s40593-021-00282-y

Leppink, J., & van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on medical education*, *4*(3), 119-127. https://doi.org/10.1007/s40037-015-0192-x

Liew, T.W., Tan, S.-M., Tan, T. M. & Kew, S. N. (2020). Does speaker's voice enthusiasm affect social cue, cognitive load and transfer in multimedia learning?. *Information and Learning Sciences*, *121*(3/4), 117-135. https://doi.org/10.1108/ILS-11-2019-0124

Linek, S. B., Gerjets, P., & Scheiter, K. (2010). The speaker/gender effect: does the speaker's gender matter when presenting auditory text in multimedia messages?. *Instructional Science*, *38*(5), 503-521. https://doi.org/10.1007/s11251-009-9115-8

Mayer, R. (2020). *Multimedia learning*. Cambridge University Press. https://doi:10.1017/9781316941355

Mayer, R. E. (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied,* *18*(3), 239. https://doi.org/10.1037/a0028616

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of educational Psychology,* *95*(2), 419. https://doi.org/10.1037/0022-0663.95.2.419

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship* (p. 9). Cambridge: MIT press.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of educational psychology,* *84*(4), 429. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, *38*(1), 1-4. https://doi.org/10.1207/S15326985EP3801_1

Rey, G. D., & Steib, N. (2013). The personalization effect in multimedia learning: The influence of dialect. *Computers in Human Behavior*, *29*(5), 2022-2028. https://doi.org/10.1016/j.chb.2013.04.003

Richards, D., & Dignum, V. (2019). Supporting and challenging learners through pedagogical agents: Addressing ethical issues through designing for values. *British Journal of Educational Technology,* *50*(6), 2885-2901. https://doi.org/10.1111/bjet.12863

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, *22*(2), 123-138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J. (2011). Cognitive load theory. *Psychology of Learning and Motivation,* *55,* 37-76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In J. Sweller, P. Ayres, & S. Kalyuga *Cognitive load theory* (pp. 71-85). Springer. https://doi.org/10.1007/978-1-4419-8126-4_6

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, 10(3), 251-296. https://doi.org/10.1023/A:1022193728205

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review, 17*(2), 147-177. https://doi.org/10.1007/s10648-005-3951-0

Wang, F., Wenjing, L., Richard, E. M., & Huashan, L. (2018). Animated pedagogical agents as aids inmultimedia learning: Effects on eye-fixations during learning and learning outcomes. *Journal of Educational Psychology, 110*(2), 250-268. https://doi.org/10.1037/edu0000221

Wouters, P., Paas, F., & van Merriënboer, J. J. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research*, *78*(3), 645-675. https://doi.org/10.3102/0034654308320320