Education and Information Technologies (2023) 28:1455–1489
https://doi.org/10.1007/s10639-022-11255-6

I am Alexa, your virtual tutor!: The effects of Amazon Alexa's text-to-speech voice enthusiasm in a multimedia learning environment

Tze Wei Liew[1] · **Su-Mae Tan**[2] · **Wei Ming Pang**[1,2] · **Mohammad Tariqul Islam Khan**[1] · **Si Na Kew**[3]

Received: 26 February 2022 / Accepted: 20 July 2022 / Published online: 30 July 2022

Abstract

Modern text-to-speech voices can convey social cues ideal for narrating multimedia learning materials. Amazon Alexa has a unique feature among modern text-to-speech vocalizers as she can infuse enthusiasm cues into her synthetic voice. In this first study examining modern text-to-speech voice enthusiasm effects in a multimedia learning environment, a between-subjects online experiment was conducted where learners from a large Asian university ($n = 244$) listened to either Alexa's: (1) neutral voice, (2) low-enthusiastic voice, (3) medium-enthusiastic voice, or (4) high-enthusiastic voice, narrating a multimedia lesson on distributed denial-of-service attack. While Alexa's enthusiastic voices did not enhance persona ratings compared to Alexa's neutral voice, learners could infer more enthusiasm expressed by Alexa's medium-and high-enthusiastic voices than Alexa's neutral voice. Regarding cognitive load, Alexa's low-and high-enthusiastic voices decreased intrinsic and extraneous cognitive load ratings compared to Alexa's neutral voice. While Alexa's enthusiastic voices did not impact affective-motivational ratings differently from Alexa's neutral voice, learners reported a significant increase of positive emotions from their baseline positive emotions after listening to Alexa's medium-enthusiastic voice. Finally, Alexa's enthusiastic voices did not enhance the learning performance on immediate retention and transfer tests compared to Alexa's neutral voice. This study demonstrates that a modern text-to-speech voice enthusiasm can positively affect learners' emotions and cognitive load during multimedia learning. Theoretical and practical implications are discussed through the lens of the Cognitive Affective Model of E-learning, Integrated-Cognitive Affective Model of Learning with Multimedia, and Cognitive Load Theory. We further outline this study's limitations and recommendations for extending and widening the text-to-speech voice emotions research.

Keywords Enthusiasm · Emotional design · Text-to-speech · Voice effect · Amazon Alexa · Multimedia learning

Extended author information available on the last page of the article

1 Introduction

Two iconic, albeit contrasting artificial intelligent voices were conceived and depicted in two acclaimed sci-fi movies, capturing the attention and enriching the imagination of audiences: *HAL 9000* in the film "2001: A Space Odyssey" by Stanley Kubrick and Arthur C. Clarke, and *Samantha* in the movie "Her." Although HAL9000's voice is soft, calm, and conversational, its mechanical and emotionless tone induces disquiet and distrust in people. On the other hand, Samantha's voice exudes emotional cues and expressiveness that mimic a natural human voice, such that the film's protagonist expresses amazement: "You seem like a person — but you're just a voice in a computer," and eventually falls in love with the artificial entity. Of course, artificial voices do not exist merely in science fiction but are pervasive in today's digital society, although this technology is evolving still. Nass and Brave (2005)'s influential book "Wired for speech: How voice activates and advances the human–computer relationship" provides compelling data and discourse asserting that artificial voices automatically evoke a wide array of social responses from listeners, thereby underscoring the importance of designing computer voices to optimize usability and engagement.

Within the educational context, text-to-speech vocalizers have generated artificial voices to narrate instructional content in multimedia learning environments. However, Mayer and his colleagues have cautioned that text-to-speech voices' mechanical and monotonous tone is detrimental to learning (Atkinson et al., 2005; Mayer & Dapra, 2012; Mayer et al., 2003). The researchers put forward the voice principle, advocating multimedia learning content to be narrated using a friendly human voice. The voice principle derives from the Social Agency Theory (Mayer, 2014), which asserts that a warm and familiar human voice can convey likable social cues that encourage learners to consider multimedia learning a genuine social interaction. As a result, learners are motivated to process the learning materials more deeply and achieve better learning outcomes. On the other hand, a mechanical-sounding text-to-speech voice can not effectively prime a sense of social presence. Hence, learners listening to this voice are less inspired to perceive the learning as a social engagement, impeding cognitive engagement and learning performance.

Nonetheless, contemporary text-to-speech vocalizers have made artificial voices sound more natural and human-like (Cambre et al., 2020; Tan et al., 2021). Artificial voices in the form of Apple's Siri, Amazon's Alexa, and Google Assistant are increasingly ubiquitous in people's lives and are often treated as intelligent social beings with personality, social, and emotional traits (Cambre et al., 2020). The preceding conforms with the computers-are-social-actors (CASA) paradigm, positing that people respond socially to digital artifacts exhibiting human-like or anthropomorphic cues (Nass & Brave, 2005; Nass & Steuer, 1993). For multimedia learning, Craig and Schroeder (2017) demonstrated that a pedagogical agent's voice produced with a modern text-to-speech vocalizer (i.e., Neospeech voice engine) led to a higher learning transfer performance than the voices generated with a classic text-to-speech vocalizer (i.e., Microsoft speech engine) or a human voice. Further, the learners reported similar agent persona ratings regarding facilitating learning and credibility to the modern text-to-speech and human voices. Craig and Schroeder (2019) conducted a follow-up study featuring a multimedia learning environment without a visible pedagogical agent. They found that while a human voice enhanced higher agent persona ratings, particularly for humanness and engagement qualities than a modern text-to-speech voice and a classic text-to-speech voice, there were no differences in learning outcomes and cognitive efficiency between the voices. Collectively, these findings imply that modern text-to-speech voice qualities are effective for multimedia learning. Thus, educators and instructional designers can leverage modern text-to-speech vocalizers to conveniently, rapidly, and cost-effectively produce voice-overs, narrations, or dialogues for multimedia learning materials.

Amazon Alexa has a superior feature compared to other modern speech vocalizers insofar as she can imbue her voice with two types of emotional expression—enthusiasm and disappointment (Peters, 2019). Synthetic emotional expressions, including negative emotional tones, are of interest within the multimedia learning context as instructors can convey these emotions strategically to promote learning in the educational milieu. For instance, expressed disappointment or anger can serve as social and feedback cues for learners to acknowledge their performance or effort deficiency, prompting them to devote more effort to reduce the gap (Jeong et al., 2017; Johnson & Connelly, 2014; Kleef et al., 2011; Liew et al., 2022; Sullins et al., 2009; Tunstall & Gsipps, 1996). Notwithstanding, this research focuses on the effects of Amazon Alexa voice enthusiasm. Presently, Alexa can convey three levels of voice enthusiasm intensity (low-, medium-, and high-enthusiasm) apart from the neutral voice (no enthusiasm).

There is a rich literature on expressed enthusiasm in classroom settings (Keller et al., 2016), indicating positive effects on learners' emotional states (Kunter et al., 2011, 2013), interest and intrinsic motivation (Keller et al., 2014; Kim & Schallert, 2014; Moè, 2016), attention (Moè, 2016; Moe et al., 2021), and learning achievement (Frenzel et al., 2010; Kunter et al., 2011, 2013; Moe et al., 2021). Applied to multimedia learning environments, enthusiastic human voices can enhance learning outcomes, emotional states, and perceived speaker's social and persona ratings (Beege et al., 2020; Lawson & Mayer, 2021; Liew et al., 2017, 2020; Wang et al., 2022). While a human's voice enthusiasm has been explored for multimedia learning (Beege et al., 2020; Liew et al., 2017, 2020; Wang et al., 2022), no studies have examined the effects of a modern text-to-speech voice enthusiasm. Thus, this study aims to fill this research gap by investigating the social, affective-motivational, cognitive, and learning effects of Amazon Alexa's voice enthusiasm in a multimedia learning environment.

2 Research rationale

This study is situated within the multicultural and multilingual Malaysian learners' profile. Malaysia has three main ethnic groups: Malays, Chinese, and Indians; thus, the Malaysian educational system affords lessons at the primary level in Malay, Mandarin, and Tamil languages. Complementing the Malaysian Language (Bahasa Malaysia) as the official language in Malaysia, the English language remains widely used in Malaysia, particularly in educational, professional, and business environments (Thirusanku & Yunus, 2014). English is a compulsory subject in primary and secondary school in the Malaysian educational system and is regarded as the primary language of instruction in most private colleges and universities (Ali, 2013). While most Malaysian students are conversant in English, their English proficiency standards vary greatly, partly shaped by diverse cultural and socio-economic backgrounds and schooling systems, e.g., Malay-medium national schools, English-language schools, or vernacular schools such as Chinese-and Tamil-language schools (Pillai & Ong, 2018). This, in turn, differently affects comprehension of spoken English imbued with varied and unfamiliar vocal prosodic cues (e.g., vowel length, pauses, and loudness) across diverse Malaysian learner profiles (Adnan et al., 2019; Rajadurai, 2006; Yap & Pillai, 2018). Enthusiastic voices exude strong and varied prosodic cues, potentially affecting non-native learners' comprehension and cognitive load within the multimedia learning context (Davis et al., 2019; Liew et al., 2020; Matthew, 2020). In light of the preceding, this study can deepen our understanding of how a modern text-to-speech synthesizer's enthusiasm can benefit learners from diverse socio-linguistic profiles, which aligns with the research agenda advocating inclusive multimedia learning design for non-native English speakers (Brom et al., 2017; Chan et al., 2020; Lee & Mayer, 2018; Liu et al., 2018; Mayer et al., 2014).

The present research's multimedia learning environment demonstrates how a denial-of-service attack occurs to business majors learners. A denial-of-service attack involves cyber-attackers attempting to disrupt or shut down a server's service by flooding the network devices with fake requests. This computer-related topic is significant to integrating Science, Technology, Engineering, and Mathematics (STEM) curriculum in business and other non-IT courses, fulfilling the ubiquitous demand and necessity for IT-savvy knowledge and skills in today's digital society. A prevalent issue motivates this study: there is a lack of learners' interest in STEM subjects globally and Malaysia (Ramli & Talib, 2017), with STEM subjects often perceived as challenging, uninspiring, and boring (Christensen et al., 2014; Yu, 2012). This can be especially apparent for business majors who generally hold lesser IT-related knowledge and experience, thereby compelling them to feel that such a topic is difficult, unfamiliar, and irrelevant to their business environment. Noteworthy, instructors' enthusiasm can positively impact learners' interest, motivation, and learning achievement in STEM subjects (Christensen et al., 2014; Jungert et al., 2020; Ramli & Talib, 2017). It is thus interesting to examine to what extent enthusiastic voices can affect the affective-motivational, cognitive, and learning outcomes of an IT topic among business majors in this study.

The findings of this work contribute theoretical insights into the scant but emerging literature on the confluence between artificial voice's synthetic emotional expressions and learning (Fountoukidou et al., 2021; Hillaire et al., 2019; Viegas & Alikhani, 2021). From an instructional design perspective, this study offers practical recommendations to educators and instructional designers regarding using modern text-to-speech with artificial enthusiasm to narrate multimedia learning materials.

The following section reviews text-to-speech technology for learning and the effects of voice enthusiasm through the lens of the Cognitive Affective Model of E-learning (Horovitz & Mayer, 2021; Lawson & Mayer, 2021; Mayer, 2020), Integrated Cognitive-Affective Model of Learning with Multimedia (Plass & Kalyuga, 2019) and Cognitive Load Theory (Mayer, 2014).

3 Theoretical framework

3.1 Text-to-speech vocalizers

Text-to-speech vocalizers, also known as speech synthesizers, are computer systems that artificially produce human speech by converting texts into spoken words. This technology is integral in endowing voices and personalities to artificial agents, virtual assistants, and robots (Cambre et al., 2020; Alonso Martin et al., 2020; Poushneh, 2021). Several text-to-speech vocalizers avail development and commercial use, with popular ones, including Amazon's Ivona Software used in Amazon devices, e.g., Kindle electronic reader, Google Text-to-Speech used in Google Now virtual assistant, Microsoft Text-to-Speech used in Cortana virtual assistant, and Nuance Real Speak featured in Apple Siri.

Text-to-speech vocalizers have innovatively enriched the educational field. Zhang and Zou (2022)'s review showed that text-to-speech technology facilitates language learning, particularly for improving speaking and pronunciation (Liakin et al., 2017; Qian et al., 2018; Shadiev et al., 2018). The technology has also proliferated learning via listening to audiobooks narrated by text-to-speech engines, notably during the pandemic (Cambre et al., 2020). Text-to-speech vocalizers can aid learners with learning disabilities, emotional-behavioral disorders, mild intellectual disabilities, dyslexia, and attention-deficit/hyperactivity disorders who struggle with reading and writing (Bone & Bouck, 2017; Evmenova & Regan, 2019). Text-to-speech engines as assistive technology allow learners to adjust pace, pitch, and speech volume. This enables struggling learners to listen to their own words during the writing process, facilitating monitoring and revising their writing outcomes. Furthermore, text-to-speech applications support learners with vision disabilities by reading aloud digital texts. Adding spoken words to readable texts can mitigate struggles encountered by ADHD learners by lessening distractibility and stress, resulting in enhanced focus and reduced exhaustion. Last but not least, text-to-speech technology enables instructional designers to create spoken words to narrate multimedia learning materials, whether with visible pedagogical agents or disembodied voice-only agents (Atkinson et al., 2005; Craig & Schroeder, 2017, 2019; Fountoukidou et al., 2021; Mayer et al., 2003). This research is framed within this technological benefit, utilizing text-to-speech to generate spoken narrations to complement the visual learning information in a multimedia learning environment.

Learning is intrinsically a social process; therefore, learners infer social-emotional cues from the voice acting as the instructional medium. While infusing social-emotional expressions into text-to-speech voices is essential for learning with digital media, text-to-speech technology generally lags in producing natural-sounding voice emotions. A few text-to-speech engines for the English language lead this artificial voice emotions technology: Amazon Alexa with enthusiasm and disappointment tones, Typecast.AI with happy and angry expressions, and Microsoft Azure with excitement, cheerful, hopeful, angry, sad, and terrified voice tones. However, there is a lack of research assessing the effects of synthetic voice emotions on learning. This study extends recent works on artificial voice expressiveness for learning (Fountoukidou et al., 2021; Westlund et al., 2017) to Amazon Alexa's voice enthusiasm.

3.2 Cognitive affective model of e-learning

The Cognitive Affective Model of E-learning explains how expressed emotions in a multimedia learning environment can influence the instructor's persona ratings, cognitive effort, and learning performance (Lawson & Mayer, 2021; Lawson et al., 2021a, b, c. As Fig. 1 illustrates, the model posits five sequentially-connected events: (1) a pedagogical agent or a human instructor expresses emotions, (2) learners recognize the expressed emotional tones, (3) learners experience social connection with the pedagogical agent or human instructor based on the expressed emotional tones, (4) learners exert learning efforts based on the perceived social connection, and (5) learners' learning efforts affect learning performance.

Lawson and her colleagues further derived the Positivity Principle from the Cognitive Affective Model of E-learning; that is, learners can recognize a pedagogical agent's or human instructor's expressed positive and active emotion tones, which elevates the instructor's persona ratings and learning performance in a multimedia learning environment (Lawson & Mayer, 2021; Lawson et al., 2021c). Empirical studies have demonstrated that learners inferred the instructor's emotional tones comprising valence (positive/negative) and activity (active/passive) (Lawson & Mayer, 2021; Lawson et al., 2021a, b, c), and that positive and active expressed emotional tones tended to enhance the instructor's persona ratings (Lawson et al., 2021a, b), effort (Lawson et al., 2021a, b), and learning performance on a delayed posttest (Lawson et al., 2021b).

Horovitz and Mayer (2021) proposed a slightly different version of the Cognitive Affective Model of E-learning, postulating the following links: (1) learners recognize the instructor's affective states based on their expressed emotional tones, (2) learners experience the same affective states following the instructor's affective states, (3) the adopted affective states of learners influence their motivational states, and (4) learners' motivational states affect learning performance (see Fig. 2). Horovitz and Mayer (2021) revealed that learners could recognize displayed emotional

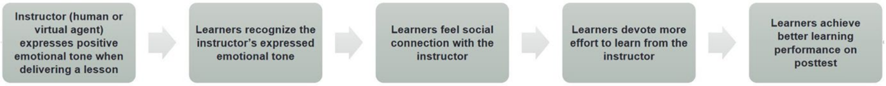

Instructor (human or virtual agent) expresses positive emotional tone when delivering a lesson
Learners recognize the instructor's expressed emotional tone
Learners feel social connection with the instructor
Learners devote more effort to learn from the instructor
Learners achieve better learning performance on posttest

Fig. 1 Cognitive Affective Model of E-learning emphasizing augmented social connection with a positive instructor

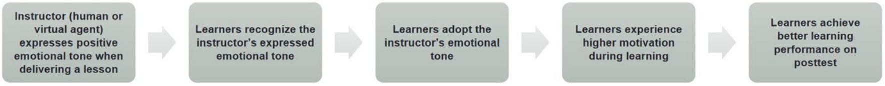

Instructor (human or virtual agent) expresses positive emotional tone when delivering a lesson
Learners recognize the instructor's expressed emotional tone
Learners adopt the instructor's emotional tone
Learners experience higher motivation during learning
Learners achieve better learning performance on posttest

Fig. 2 Cognitive Affective Model of E-learning emphasizing augmented affective-motivational state with a positive instructor

tones and that the instructor's expressed happiness enhanced learners' affective-motivational states, thereby endorsing the Positivity Principle.

This study regards Alexa's text-to-speech voice enthusiasm as a positive emotional cue, following Keller et al. (2016)'s conceptualization characterizing expressed enthusiasm as an affective construct associated with positive and highly activating arousal emotions such as joy and excitement, pleasure, and intrinsic motivation concerning a learning topic. Thus, Alexa's text-to-speech voice enthusiasm aligns with the Positivity Principle and is compatible with recent works indicating the benefits of positive emotional tones on persona ratings and learning performance. Liew et al. (2017) reported that a pedagogical agent who expressed enthusiasm through facial expressions, head movements, and vocal parameters, enhanced the agent's persona ratings and learning performance than a pedagogical agent who expressed a neutral emotional tone. Across two experiments, an enthusiastic human voice elevated social ratings of the speaker (human-like and engaging) and learning performance than a calm human voice (Liew et al., 2020). Relatedly, Fountoukidou et al. (2021) showed that a text-to-speech voice with expressive vocal qualities than a neutral text-to-speech voice accentuated social and persona ratings, that is, immediacy and affective perceptions toward the agent and the learning materials.

3.3 Integrated-cognitive affective model of learning with multimedia

The Integrated-Cognitive Affective Model of Learning with Multimedia model (Plass & Kalyuga, 2019; Plass & Kaplan, 2015) posits that multimedia learning's Selecting, Organizing, and Integrating cognitive processes are entwined with affective processes. Further, the cognitive-affective processes entail affective processes that require cognitive resources and vice-versa. The model emphasizes that multimedia learning environments can evoke affective responses through visual/verbal information — and the affective reactions influence the Selecting, Organizing via affect in the form of interest and motivation, and Integration of visual/verbal information and affect as emotional schemas stored in long-term memory. Based on the Integrated-Cognitive Affective Model of Learning with Multimedia model, Plass and Kaplan (2015) put forth the Emotional Design thesis that visual and verbal features in a multimedia learning environment can be designed to evoke learners' affective-motivational responses for promoting learning.

Besides visual aesthetics such as anthropomorphism and warm colors (Brom et al., 2018; Wong & Adesope, 2021), social cue qualities like attractiveness, cuteness, and emotional expressions of virtual or human characters in a multimedia learning environment can also influence learners' affective-motivational factors (Domagk, 2010; Domagk et al., 2010; Horovitz & Mayer, 2021; Plass et al., 2020;

Schneider et al., 2021). Along this line of reasoning, manipulating voice socio-emotive cues in a multimedia learning environment falls under the Emotional Design model (Beege et al., 2020; Liew et al., 2017; Rodero & Lucas, 2021; Wang et al., 2022). The literature has extolled the benefits of instructors' displayed enthusiasm in promoting learners' affective-motivational behaviors such as willingness to learn, learning enjoyment, and learning performance in classrooms (Keller et al., 2016; Kim & Schallert, 2014; Kunter et al., 2011, 2013; Moè, 2016; Moe et al., 2021). Recent evidence shows that these affective-motivational benefits of enthusiasm cues extend to multimedia learning environments. Liew et al. (2017) found that an energetic virtual agent expressing an enthusiastic human voice led to higher positive affect and intrinsic motivation than an agent conveying a neutral human voice and gestures. Beege et al. (2020) showed that an enthusiastic than a neutral human voice attached to a virtual agent increased learning performance for learners under low load conditions but decreased learning performance for learners under high load conditions. Horovitz and Mayer (2021) demonstrated that instructors' positive emotional cues, such as voice expressing happiness, evoked learners to experience higher positive affect (happiness), motivation, and interest than instructors' negative emotional cues, including voice expressing boredom. Similarly, Ba et al. (2021) showed that a pedagogical agent's voice expressing a positive than a neutral emotional tone enhanced learners' positive affect (pride) and transfer performance. Across three experiments, Wang et al. (2022) found that a pedagogical's agent positive emotional tone exhibiting a smiling facial expression and an enthusiastic voice robustly elevated learners' positive emotions and motivation than a neutral emotional tone depicting a neutral facial expression and a calm voice.

3.4 Cognitive load theory

Learning with multimedia materials imposes mental processing demands on learners' cognitive architecture. The Cognitive Load Theory posits that multimedia learning occurs with a limited working memory resource and distinguishes three types of load: Intrinsic, Extraneous, and Germane (Kalyuga, 2011; Sweller et al., 1998). Intrinsic cognitive load is utilized to process information related to a learning subject and is influenced by the subject's complexity and learners' prior knowledge of the topic. Extraneous cognitive load is the mental demand imposed by the design of learning materials that affect how information is presented to the learners. This load does not contribute to learning gains and should be minimized through sound instructional design (Mayer, 2014). Germane cognitive load is the mental resources used for learning-relevant cognitive functions, such as acquiring and constructing schema in the long-term memory, employing learning techniques like pattern-probing, reconfiguring problem representation to facilitate solving, and metacognitive tracking cognition and learning (Debue et al., 2014).

This study considers the cognitive load effects of Alexa's voice enthusiasm through the lens of two theoretical perspectives. The first is guided by Davis et al. (2019)'s analysis indicating that voice prosody can influence cognitive load, specifically germane cognitive load among non-native English speakers. Voice prosody refers to the vocal qualities such as pitch, tempo, stress, intonation, melody, loudness, accent, and pause. They convey critical nonverbal communication beyond a sentence's literal word meaning. The meaning of a sentence can be interpreted or understood differently based on voice prosodic variations. However, for non-native English speakers, a voice imbued with strong prosodic cues (a strong-prosodic voice) can impede fluent understanding of the intended content. This is because non-native English speakers are unfamiliar with the prosodic connotations associated with varied intonation, nuances, pronunciation, and speech rate (Davis et al., 2019).

Because non-native English speakers tend to adopt bottom-up processing that focuses on words than semantic structures (Osada, 2001), a strong-prosodic narration can impose more cognitive demands during learning than a weak-prosodic narration expressing less variability in pitch, intonation, and speech rate (Davis et al., 2019). The experiment found that non-native English speakers reported higher germane cognitive load ratings when learning from a weak-prosodic narration than a strong-prosodic narration. In this study's context, an enthusiastic voice is more likely to exhibit strong-prosodic cues comprising varied tones, tempo, stress, intonation, and loudness than a neutral voice. Thus, voice enthusiasm can influence germane cognitive load, particularly among non-native English speakers. This proposition is supported by Liew et al. (2020)'s study demonstrating that an enthusiastic voice led to lower germane cognitive load ratings than a calm voice among non-native English speakers.

Secondly, the cognitive load effects of Alexa's voice enthusiasm can be theorized through the Integrated-Cognitive Affective Model of Learning with Multimedia and the Emotional Design hypothesis (Plass & Kalyuga, 2019; Plass & Kaplan, 2015). The Emotions as Suppressor view suggests that processing emotions can impose extraneous cognitive load, which competes for working memory resources during multimedia learning (Plass & Kalyuga, 2019). On the other hand, the Emotions as Facilitator view associates positive activating emotions (e.g., enjoyment of learning) with high motivation and mental effort, building on the Control-Value Theory of Achievement Emotions (Pekrun, 2006; Plass & Kalyuga, 2019). Emotional Design meta-analyses have indicated that learners tended to report lower perceived difficulty with multimedia learning materials augmented with aesthetically-pleasing features such as warm colors and anthropomorphic images than the control versions (Brom et al., 2018; Wong & Adesope, 2021). Considering voice enthusiasm as an emotional design feature, Beege et al. (2020) found that mental load conditions moderated the effects of an enthusiastic human voice on cognitive load. When the learners were subjected to the low mental load conditions, the enthusiastic than the neutral human voice decreased extraneous cognitive load (subsumes processing of learning-irrelevant details like the design of learning environment) and increased germane cognitive load (subsumes generative processing contributing to schemata construction and understanding of the learning materials). Whereas, for the learners subjected to the high mental load conditions, the enthusiastic human voice increased extraneous cognitive load while decreasing germane cognitive load. Other studies found that while human voices with positive emotional tones such as enthusiasm and happiness in a multimedia learning environment enhanced learning performance, perceived difficulty was not impacted by the voices (Lawson et al., 2021b; Liew et al., 2017, 2020; Wang et al., 2022).

4 Hypotheses

4.1 Alexa's voice enthusiasm effect on perceived enthusiasm

Findings endorsing the Cognitive Affective Model of E-learning model have indicated that learners could recognize a virtual agent or an instructor's expressed emotions through nonverbal cues (facial expressions, body gestures, and vocal characteristics) in a multimedia learning environment (Lawson et al., 2021a, b, c). Further, Lawson and Mayer (2021) demonstrated that learners could infer the intended emotional tones based on a human voice alone in a multimedia learning environment. This study extends the premise to synthetic emotional tones with text-to-speech vocalizers by testing the hypothesis that:

> *H1: Learners listening to Alexa's enthusiastic voices will infer significantly more voice enthusiasm than learners listening to Alexa's neutral voice.*

4.2 Alexa's voice enthusiasm effect on persona ratings

Supporting the Positivity Principle derived from the Cognitive Affective Model of E-learning model, studies have demonstrated that expressed positive and active emotional tones were associated with elevated instructor's persona ratings (Lawson et al., 2021a, b). Similarly, Liew et al. (2017) and Liew et al. (2020) showed that a virtual agent's and a speaker's expressed enthusiasm rather than neutral emotional tone enhanced social qualities and persona ratings, whereas Fountoukidou et al. (2021) reported that a synthetic voice expressiveness positively impacted perceived immediacy and affective perception toward the instructor. Hence, this study predicts that:

> *H2: Learners listening to Alexa's enthusiastic voices will assign significantly more positive persona ratings to Alexa than learners listening to Alexa's neutral voice.*

4.3 Alexa's voice enthusiasm effect on affective-motivational ratings

This study regards an enthusiastic voice as an emotional design cue that can induce positive emotions and motivations in learners, given that displayed instructor enthusiasm comprises positive-activating emotions (Keller et al., 2014, 2016) that elicit positive affective-motivational states in learners through the emotional contagion effect (Hatfield et al., 1993; Moè, 2016; Plass et al., 2020). Liew et al. (2017) revealed that nonverbal cues, including facial expression, gesture, and voice conveying enthusiasm elevated learners' positive emotions and intrinsic motivation.

Likewise, across three experiments, Wang et al. (2022) found that a pedagogical agent's positive emotional tone exhibiting a smiling facial expression and an enthusiastic voice evoked higher positive emotions and motivation than a neutral emotional tone displaying a neutral facial expression and a calm voice. From the Cognitive Affective Model of E-learning model perspective, Horovitz and Mayer (2021) showed that an instructor's expressed happiness was associated with higher learners' positive affect (happiness) and motivation. On the other hand, Beege et al. (2020) revealed that an enthusiastic voice than a neutral voice enhanced learners' positive-activating emotional states in the pre-study but not in the main experiment. The researchers argued that enthusiasm cues through voice alone without other nonverbal cues such as facial expression and body gestures might not be strong enough to influence learners to report higher positive emotional states consciously. This study extends the affective-motivational effects of voice enthusiasm to a modern text-to-speech voice by testing the following hypotheses:

> *H3(a): Learners will report significantly more positive emotions than their baseline positive emotions after listening to Alexa's enthusiastic voices.*
> *H3(b): Learners listening to Alexa's enthusiastic voices will report significantly more positive emotions after the learning engagement than learners listening to Alexa's neutral voice.*
> *H3(c): Learners listening to Alexa's enthusiastic voices will report significantly higher intrinsic motivation than learners listening to Alexa's neutral voice.*

4.4 Alexa's voice enthusiasm effect on cognitive load ratings

Research has revealed that emotional design features (e.g., warm colors and anthropomorphic images) have a robust effect in decreasing perceived difficulty, plausibly due to the "what is beautiful is usable" effect (Tractinsky et al., 2000); that is, the attractive aesthetics in a multimedia learning environment compel learners to perceive the learning topics or materials as easier to process (Brom et al., 2018; Wong & Adesope, 2021). Enthusiastic voices can be considered an appealing emotional design cue (Beege et al., 2020); hence, they may decrease perceived difficulty. However, the scant research generally demonstrated that voice enthusiasm was not associated with reduced perceived difficulty, including intrinsic and extraneous cognitive load ratings (Liew et al., 2017, 2020; Wang et al., 2022). Contrariwise, Beege et al. (2020) found evidence that voice enthusiasm could lead to decreased extraneous cognitive load ratings (perceived difficulty in processing the learning materials due to the instructional design features), albeit only for learners experiencing low mental load. On the other hand, voice enthusiasm led to increased extraneous cognitive load ratings for learners experiencing high mental load. This study observes voice effects research (Beege et al., 2020; Davis et al., 2019; Liew et al., 2020) in distinguishing perceived difficulty to intrinsic cognitive load ratings (perceived difficulty based on the learning topic) and extraneous cognitive load ratings (perceived difficulty based on the instructional design). This study explores the potential of Alexa's voice enthusiasm in reducing perceived difficulty by evaluating the following hypotheses:

H4(a): Learners listening to Alexa's enthusiastic voices will report significantly lower intrinsic cognitive load ratings than learners listening to Alexa's neutral voice.
H4(b): Learners listening to Alexa's enthusiastic voices will report significantly lower extraneous cognitive load ratings than learners listening to Alexa's neutral voice.

The literature indicates that voice enthusiasm effect on germane cognitive load ratings (subsumes generative processing contributing to schemata construction and understanding of the learning materials) is not robust and may be susceptible to confounding factors. For instance, Beege et al. (2020) revealed that voice enthusiasm could enhance germane cognitive load, but only for learners experiencing low than high mental load. Wang et al. (2022) found that a pedagogical agent's positive emotional tone, including an enthusiastic voice, was associated with higher germane load ratings in only one of three conducted experiments. Further, the voice enthusiasm effect on germane cognitive load ratings should be considered within a non-native language speaker context (see Sect. 3.4). Empirical findings have indicated that multimedia learning narrations using strong-prosodic voices, including an enthusiastic voice, were associated with lower germane cognitive load among non-native English speakers (Davis et al., 2019; Liew et al., 2020). As this study's design involves non-native English speakers engaging with the multimedia lesson delivered in English, the following hypothesis will be examined:

H4(c): Learners listening to Alexa's enthusiastic voices will report significantly lower germane cognitive load ratings than learners listening to Alexa's neutral voice.

4.5 Alexa's voice enthusiasm effect on learning performance

Studies based on the Cognitive Affective Model of E-learning and the Positivity Principle revealed that a virtual agent or human instructor's expressed positive than neutral or negative emotional tones in a multimedia learning environment did not improve learning performance, plausibly because the immediate posttests were not sensitive enough for assessing deep learning effects (Horovitz & Mayer, 2021; Lawson et al., 2021a, b; Wang et al., 2022). Conversely, Liew et al. (2016) and Liew et al. (2020) revealed that voice enthusiasm led to better transfer scores measured through immediate posttest. Beege et al. (2020) found that voice enthusiasm was associated with higher learning scores assessed through an immediate posttest; specifically, recognition posttest (multiple-choice) for learners experiencing low mental load. This study aims to contribute insights into the mixed findings by examining the following hypothesis:

H5: Learners listening to Alexa's enthusiastic voices will perform significantly better on the posttest than learners listening to Alexa's neutral voice.

5 Method

5.1 Research design

This study adopted a between-subjects experimental design in which learners engaged with a multimedia learning environment narrated by one of the four Alexa's voices: (1) neutral (no enthusiasm), (2) low-enthusiasm, (3) medium-enthusiasm, and (4) high-enthusiasm. The experiment was conducted through Alchemer, an online survey tool incorporating the multimedia lesson, survey, and posttest.

5.2 Multimedia learning environment and Alexa's voices

Adobe Animate was used to develop a 196-s animation explaining how a distributed denial-of-service (DDoS) attack occurs (see Fig. 3). The learning content script, presented in Appendix A, had a Flesch-Kincaid Grade Level of 9.7, aligning with the university student's comprehension difficulty. We converted the learning content to Alexa's spoken voice narrations through the text-to-speech function in the Amazon Alexa Developer Console environment (see Fig. 4). We programmed the text-to-speech application to set Alexa's emotional tone to neutral, low-enthusiastic, medium-enthusiastic, and high-enthusiastic. The Alexa's voice samples are presented below:

1. Alexa's neutral voice (sample: https://youtu.be/pTzkbuVQfOw)
2. Alexa's low-enthusiastic voice (sample: https://youtu.be/84Reh8xkxhk)
3. Alexa's medium-enthusiastic voice (sample: https://youtu.be/GJkTwzdVjFs)
4. Alexa's high-enthusiastic voice (sample: https://youtu.be/mIJ_K_3evys)

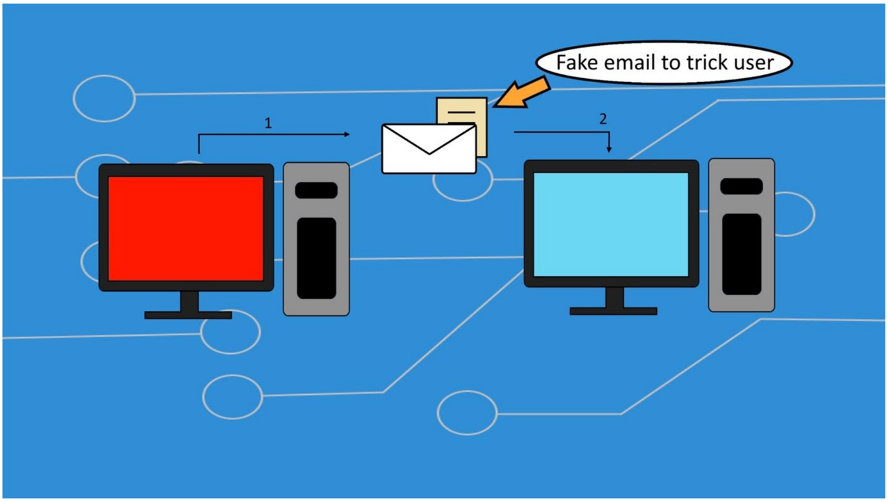
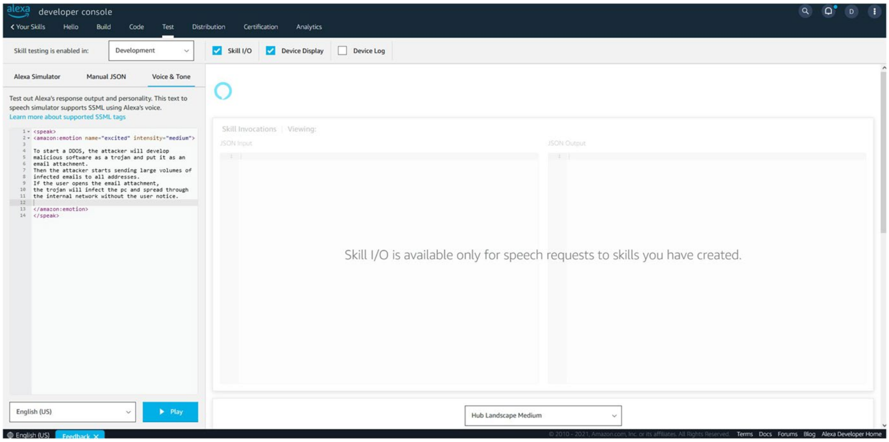


Fig. 3 The multimedia learning environment



Test out Alexa's response output and personality. This text to speech simulator supports SSML using Alexa's voice.

```
<speak>
<amazon:emotion name="excited" intensity="medium">

To start a DDOS, the attacker will develop
malicious software as a trojan and put it as an
email attachment.
Then the attacker starts sending large volumes of
infected emails to all addresses.
If the user opens the email attachment,
the trojan will infect the pc and spread through
the internal network without the user notice.

</amazon:emotion>
</speak>
```

Skill I/O is available only for speech requests to skills you have created.

Fig. 4 The Amazon Alexa Developer console

5.3 Instruments

This study utilized a ten-point Likert scale survey for assessing learners' prior knowledge concerning the instructional topic (average of items' scores) comprising three items: (1) How much knowledge about Distributed Denial of Service (DDOS) do you have? (2) How much understanding of Distributed Denial of Service (DDOS) do you have? and (3) How much familiarity with Distributed Denial of Service (DDOS) do you have? Learners' English proficiency was measured by averaging the items' scores of a ten-point Likert scale survey with four items: (1) "How proficient are you in the English language?" (2) "How skillful are you in the English language?,"

(3) "How good are you at understanding the spoken dialogue of the English language?," and (4) "How familiar are you with listening to the spoken English language?".

Alexa's perceived voice enthusiasm scores (average of items' scores) were assessed using a ten-point Likert scale asking learners to indicate their agreement with the following items: (1) Alexa's voice was full of enthusiasm, and (2) Alexa's voice showed enthusiasm. Following Emotional Design studies (Liew et al., 2017), learners' positive emotions were measured as a total score based on the ten items defining various positive feelings through the Positive Affect Scale (PAS). This study used the seven-point Likert scale intrinsic motivation scale consisting of eight items to measure learners' intrinsic motivation scores (sum of all items' scores) (Liew et al., 2017; Plass et al., 2020). Similar to recent research examining voice effects on cognitive load (Davis et al., 2019), this study utilized Leppink's eleven-point Likert scale (Leppink et al., 2013) to measure intrinsic load (three items), extraneous load (three items), and germane load (four items). The reliability analysis indicated that all the abovementioned scales were reliable ($\alpha > 0.7$).

Learning performance was assessed through immediate retention and transfer posttests. The first retention question asked why a perpetrator would conduct a DDoS attack — one score was awarded for each correct answer, including monetary, fun, or political factors. The second retention question asked about the DDoS attack's seven steps, with one score granted for each right step. The transfer test comprised the three questions: (1) If you are an attacker, how would you make a DDOS attack more effective? (2) If you are a server administrator, how would you prevent (avoid) a DDOS attack? and (3) If you are a computer user, how would you prevent (avoid) the risk of your personal computers or other personal devices from getting infected and becoming botnets? One mark was awarded for each acceptable answer, such as infusing a Trojan virus in an unsecured website or on the router (transfer question 1); updating antivirus and firewall, promoting security awareness, or doing a routine check-up (transfer question 2); and ignoring suspicious emails, activating antivirus and firewall, or performing regular updates on antivirus (transfer question 3). Two examiners followed a scoring guide rubric to score the answers blind to the conditions and resolved any score difference through discussion.

5.4 Participants and the online experiment

Learners' diverse cultural traits, linguistic profile, prior subject knowledge, and educational background can differently influence the effects of features aimed at evoking affective-motivational and cognitive changes in a multimedia learning environment (Brom et al., 2017, 2018; Davis et al., 2019; Wong & Adesope, 2021). Therefore, this study aimed to involve learners with similar educational profiles, linguistic ability, and prior knowledge. To this end, we sampled only business majors from a large private Asian university that uses English as a medium of instruction. The non-IT majors generally had low prior knowledge of the instructional topic (novice learners) while sharing similar English proficiency and academic background. Section 6.1 describes more thoroughly the demographic data of this study's learners.

We posted an invitation to participate in the study in exchange for an e-voucher, instructions, and the Alchemer's (online survey platform) access link on the university's business student social media groups on Facebook, Google Classroom, and Whatsapp. The instructions asked the learners to: (1) use desktop or laptop instead of mobile devices, (2) use headphones, earphones, or high-quality speakers, and (3) engage with the learning activity in a non-distractive environment. Business majors accessed the link to Alchemer, logon into the online survey platform to engage with the multimedia lesson, and completed the self-reported surveys and posttest. The following describes in more detail the experimental process via the Alchemer platform:

1. The platform had an auto-detection system that authorized only log-on through desktop or laptop than mobile devices.
2. Learners log on to Alchemer using their university's student email account. We set the platform to allow only one login per student to prohibit multiple attempts.
3. Learners read and indicated informed consent by clicking on the checkbox.

4. As an audio test, learners listened to the platform's spoken code ("LAPPY") and must type in the correct code into the system to proceed to the next section.
5. Learners filled out the survey on demographics, DDoS prior knowledge, and baseline positive emotions (Positive Affect Scale).
6. Learners were automatically randomized by the platform to engage with either one of the four multimedia lessons featuring (1) Alexa's neutral voice version, (2) Alexa's low-enthusiastic voice version, (3) Alexa's medium-enthusiastic voice version, or (4) Alexa's high-enthusiastic voice. Learners could replay the multimedia lesson within ten minutes but could not view it once they had accessed the next section.
7. Learners filled out the survey on Alexa's emotional tone, Alexa's persona ratings (Agent Persona Inventory), learner's emotional state (Positive Affect Scale), intrinsic motivation, and cognitive load (Leppink's scale: intrinsic, extraneous, and germane load).
8. Learners answered the posttest questions via text box. The time limit for the posttest was three minutes for Retention Question 1 and eight minutes each for Retention Question 2, Transfer Questions 1, Transfer Questions 2, and Transfer Questions 3. This section reminded learners not to acquire answers externally and assured learners that their performance would not affect their grades outside of the experiment.
9. The last section of the survey platform thanked and debriefed the learners.

Two hundred forty-four ($n=244$) valid learner's data was collected after discarding those who stayed on the respective multimedia lesson section for less than the animation's 196 s. All learners were aged between 17 and 25 and reported business majors.

6 Data analyses and results

6.1 Learners' profile

Table 1 presents the learners' profiles. The demographics data generally conforms with the representative profile of undergraduate business majors in a large Malaysian private university sampled in this study. All learners were aged between 17 and 25. Most of the learners were females, which aligns with the higher number of females than male undergraduates within business courses in Malaysian higher learning institutions. The majority of this study's learners were of the Chinese ethnicity, cohering with Malaysian private universities and colleges being predominantly populated with Malaysian Chinese students. The data indicate that about half of the learners were pursuing their Diplomas while the others were in their Degree programs. Concerning the majors of the business courses, the learners in this study mostly reported general business administration (Diploma program), followed by banking and finance, international business, marketing, human resource, knowledge management, and accounting. Expectedly, the business majors generally had low IT knowledge — the data in Sect. 6.3 shows that the mean and median of the learners'

Table 1 Learners' profile

Learners' profile	Percentage
Gender	
Male (n = 77)	31.6%
Female (n = 166)	68%
Undisclosed (n = 1)	0.4
Race	
Chinese (n = 203)	83.2%
Malay (n = 17)	7.0%
Indian (n = 18)	7.4%
Other (n = 6)	2.5%
Age	
Range	17–25
Standard Deviation	1.4
Average	20 years
Current education level	
Pre-University (n = 2)	0.8%
Diploma (n = 127)	52.0%
Bachelor's Degree (n = 115)	47.1%
Study major	
Accounting (n = 3)	1.2%
Banking and Finance (n = 42)	17.2%
Human Resource (n = 18)	7.4%
International Business (n = 41)	16.8%
Knowledge Management (n = 12)	4.9%
Marketing (n = 30)	12.3%
Business studies (general) (n = 98)	40.2%
State / City	
Johor (n = 58)	23.8%
Kedah (n = 4)	1.6%
Kelantan (n = 1)	0.4%
Melaka (n = 19)	7.8%
Negeri Sembilan (n = 32)	13.1%
Pahang (n = 2)	0.8%
Penang (n = 8)	3.3%
Perak (n = 5)	2.0%
Perlis (n = 1)	0.4%
Sabah (n = 3)	1.2%
Sarawak (n = 4)	1.6%
Selangor (n = 69)	28.3%
Terengganu (n = 4)	1.6%
Undisclosed (n = 34)	13.9%

prior knowledge scores regarding the instructional topic of the denial-of-service attack were 2.66 (on a ten-point Likert scale), affirming that the participants were novice learners.

Approximately 45 percent of the learners were from the southern region of Peninsular Malaysia, comprising Johor, Negeri Sembilan, and Melaka states. This geographical profile aligns with the fact that the university's campus, which this study's learners were sampled, is located in Peninsular Malaysia's southern region. About 28 percent of the learners were from Peninsular Malaysia's central region, i.e., Selangor state. Seven percent of the learners were from Peninsular Malaysia's northern part comprising Perlis, Kedah, and Penang Perak states. Approximately 3 percent of the learners were from Peninsular Malaysia's east coast region comprising Kelantan, Pahang, and Terengganu states. Another 3 percent of the learners were from East Malaysia, comprising Sabah and Sarawak states.

As this study's learners were mostly Malaysian Chinese, it is worth pointing out that they are likely to have attended Chinese vernacular schools that observe a predominant usage of the Mandarin language within the educational settings, despite the English language being compulsory taught as a subject. Implicatively, the sample learners were regarded as mostly non-native English speakers. However, it is also worth noting that learners were required to possess International English Language Testing System scores ranging from 5.0 to 7.0 or equivalent for admissions at the private university. Thus, the sample learners were sufficiently adept in writing, speaking, and listening to the English language. The data in Sect. 6.4 support this, indicating the mean and median of learners' self-reported English proficiency scores as 5.83 and 5.50, respectively (on a ten-point Likert scale). Nonetheless, the learners' English proficiency standards could vary depending on socio-economic profiles and geographical locations. Malaysian learners from middle-upper and upper-income backgrounds and major cities or urban areas tend to have higher English proficiency than others. In light of this, we examined whether the English proficiency levels differ between Alexa's voice conditions in the subsequent analysis (Sect. 6.4) to check if English proficiency standards can confound this study's findings.

6.2 Descriptive data

Table 2 shows the means and standard deviations of the dependent measures.

6.3 Do learners' prior knowledge about the instructional topic differ between the Alexa's voice conditions?

The mean and median of the learners' prior knowledge scores were both 2.66 (on a ten-point scale), indicating that this study's participants were novice learners. A one-way ANOVA revealed that the learners' prior knowledge about the instructional topic did not significantly differ between the four Alexa's voice conditions, $F(3,240)=0.97$, $p=0.41$. Thus, the random assignment effectively created conditions equivalent in the learner's prior knowledge about the instructional topic.

Table 2 Means and standard deviations of the measures

	Alexa's neutral voice ($n=62$) M (SD)	Alexa's low-enthusiastic voice ($n=65$) M (SD)	Alexa's medium-enthusiastic voice ($n=67$) M (SD)	Alexa's high-enthusiastic voice ($n=50$) M (SD)	Total ($n=244$) M (SD)
Prior Knowledge	2.50 (1.30)	2.84 (1.20)	2.57 (1.29)	2.72 (1.09)	2.66 (1.23)
English Proficiency	5.62 (1.49)	5.73 (1.95)	5.79 (1.68)	6.29 (1.98)	5.83 (1.78)
Positive emotions before learning engagement	32.11 (6.81)	32.08 (7.96)	32.16 (6.95)	31.96 (7.62)	32.09 (7.29)
Perceived Alexa's enthusiasm	5.32 (1.98)	5.52 (1.83)	6.01 (2.04)	6.23 (2.09)	5.75 (2.00)
Facilitating Learning	4.38 (1.12)	4.41 (1.08)	4.43 (1.10)	4.35 (1.12)	4.40 (1.10)
Credibility	4.88 (1.24)	5.16 (1.15)	5.15 (1.25)	4.90 (1.07)	5.03 (1.19)
Human-like	3.83 (1.46)	3.62 (1.38)	4.01 (1.49)	3.77 (1.38)	3.81 (1.43)
Engaging	4.02 (1.41)	4.11 (1.26)	4.26 (1.32)	4.14 (1.41)	4.13 (1.34)
Positive emotions after learning engagement	32.92 (7.66)	33.06 (7.79)	33.84 (7.58)	33.00 (8.46)	33.23 (7.80)
Intrinsic motivation	4.58 (1.14)	4.39 (1.04)	4.46 (1.15)	4.55 (1.35)	4.49 (1.16)
Intrinsic Load	5.40 (1.66)	4.68 (2.04)	4.96 (1.87)	4.80 (2.45)	4.96 (2.01)
Extraneous Load	3.77 (2.40)	2.89 (2.02)	3.19 (1.95)	2.84 (1.99)	3.19 (2.12)
Germane Load	6.20 (1.57)	6.20 (1.46)	5.96 (1.80)	5.96 (1.87)	6.09 (1.67)
Retention	1.39 (1.56)	1.49 (2.00)	1.31 (1.48)	1.82 (2.29)	1.48 (1.83)
Transfer	4.26 (3.44)	3.65 (3.75)	4.12 (3.27)	3.86 (4.28)	3.98 (3.65)

6.4 Do learners' English proficiency level differ between the Alexa's voice conditions?

The mean and median of learners' self-reported English proficiency scores were 5.83 and 5.50, respectively (on a ten-point scale). A one-way ANOVA found that the learners' English proficiency scores did not significantly vary between the four Alexa's voice conditions, F(3,240) = 1.49, $p = 0.22$. This data implies that the random assignment effectively created conditions equivalent in the learner's English proficiency.

6.5 Do learners' baseline positive emotions differ between the Alexa's voice conditions?

A one-way ANOVA found that learners' baseline positive emotions did significantly not differ between the four Alexa's voice conditions, F(3,240) = 0.00, $p = 0.99$. Thus, the random assignment effectively created conditions equivalent in the learners' baseline positive emotions.

6.6 Do learners infer more voice enthusiasm from Alexa's enthusiastic voices than Alexa's neutral voice

A series of independent-samples t-tests was conducted to compare learners' perceived Alexa's enthusiasm between Alexa's neutral voice and each Alexa's enthusiastic voice (low-, medium-, and high-enthusiasm). Learners did not regard Alexa's low-enthusiastic voice as more enthusiastic than Alexa's neutral voice, t(125) = 0.57, $p = 0.57$, d = 0.09. However, compared to Alexa's neutral voice, learners reported significantly higher perceived voice enthusiasm for Alexa's medium-enthusiastic voice, t(127) = 1.95, $p = 0.05$, d = 0.34, and high-enthusiastic voice, t(110) = 2.35, $p = 0.02$, d = 0.45. Thus, learners could infer significantly more voice enthusiasm from Alexa medium- and high-enthusiastic voices than from Alexa's neutral voice. However, learners did not sense significantly more voice enthusiasm from Alexa's low-enthusiastic voice than Alexa's neutral voice. H1, which predicted that learners could infer significantly more voice enthusiasm from Alexa's enthusiastic voices than Alexa's neutral voice, was endorsed for Alexa's medium- and high-enthusiastic voices.

6.7 Do learners report higher Alexa's persona ratings with Alexa enthusiastic voices than Alexa's neutral voice?

This study conducted a series of independent-samples t-tests to compare Alexa's persona ratings per the Agent Persona Inventory's four subcomponents (Baylor & Kim, 2009) between Alexa's neutral voice and each Alexa's enthusiastic voice (low-, medium-, and high-enthusiasm). Compared to Alexa's neutral voice, Alexa's low-enthusiastic voice did not significantly enhance Alexa's persona ratings concerning facilitating learning, t(125) = 0.13, p = 0.89, d = 0.03; credibility, t(125) = 1.29, p = 0.2, d = 0.23; human-like, t(123.6) = 0.85, p = 0.4, d = 0.15; and engaging, t(125) = 0.386, p = 0.7, d = 0.07. Similarly, learners did not accord Alexa's medium-enthusiastic voice with significantly more positive persona ratings regarding facilitating learning, t(125.92) = 0.21, p = 0.83, d = 0.05; credibility, t(127) = 1.21, p = 0.23, d = 0.22; human-like, t(127) = 0.69, p = 0.49, d = 0.12; and engaging, t(127) = 1, p = 0.32, d = 0.18 than Alexa's neutral voice. Lastly, Alexa's high-enthusiastic than Alexa's neutral voice did not affect Alexa's persona ratings differently for facilitating learning, t(110) = 0.15, p = 0.88, d = 0.03; credibility, t(110) = 0.06, p = 0.96, d = 0.02; human-like, t(110) = 0.24, p = 0.81, d = 0.04; and engaging, t(110) = 0.46, p = 0.65, d = 0.09. Collectively, these results did not support H2, which predicted that Alexa's enthusiastic voices could lead to significantly more positive Alexa's persona ratings than Alexa's neutral voice.

6.8 Do learners report higher affective-motivational ratings with Alexa enthusiastic voices than Alexa's neutral voice?

This study conducted a series of paired-samples t-tests to assess whether Alexa's voices could prompt learners to report an increase of positive emotions from their baseline positive emotions. The results indicated that learners did not experience significantly more positive emotions after engaging with the multimedia lesson featuring Alexa's neutral voice, t(61) = 1.37, p = 0.14, d = 0.11, low-enthusiastic voice, t(64) = 1.37, p = 0.17, d = 0.12, and high-enthusiastic voice, t(49) = 1.57, p = 0.12, d = 0.13. On the other hand, compared to the learners' baseline positive emotions before the learning engagement, there was a significant increase in learners' positive emotions after engaging with the multimedia lesson that featured Alexa's medium-enthusiastic voice, t(66) = 2.99, p = 0.00, d = 0.23. In other words, learners who listened to Alexa's medium-enthusiastic voice reported a significant increase of positive emotions from their baseline positive emotions; thus, H3(a) was only confirmed for Alexa's medium-enthusiastic voice but not for Alexa's low- and high-enthusiastic voices.

Next, we conducted a series of one-way analyses of covariance (ANCOVA) with learners' Positive Affect Scale scores measured before the learning engagement added as a covariate to compare learners' Positive Affect Scale scores measured after the learning engagement between Alexa's neutral voice and each Alexa's enthusiastic voice (low-, medium-, and high-enthusiasm). The results revealed that in comparison with Alexa's neutral voice, Alexa's low-enthusiastic voice, F(1,124) = 0.04, p = 0.84, η_p^2 = 0.00, medium-enthusiastic voice, F(1,126) = 1.27, p = 0.26, η_p^2 = 0.01, and high-enthusiastic voice, F(1,109) = 0.07, p = 0.79, η_p^2 = 0.00, did not evoke significantly higher positive emotions. Collectively, these results refuted H3(b) that learners listening to Alexa's enthusiastic voices would report significantly more positive emotions than learners listening to Alexa's neutral voice.

This study performed a series of independent-samples t-tests to compare learners' intrinsic motivation scores between Alexa's neutral voice and each Alexa's enthusiastic voice (low-, medium-, and high-enthusiasm). The results demonstrated that compared to Alexa's neutral voice, learners did not report higher intrinsic motivation with Alexa's low-enthusiastic voice, t(125) = 0.99, p = 0.32, d = 0.17, medium-enthusiastic voice, t(126.4) = 0.99, p = 0.87, d = 0.10, and high-enthusiastic voice, t(110) = 0.16, p = 0.55, d = 0.02. Collectively, these findings rejected H3(c) that Alexa's enthusiastic voices than neutral voice would significantly enhance learners' intrinsic motivation ratings.

6.9 Do learners report different cognitive load ratings with Alexa enthusiastic voices than Alexa's neutral voice?

A series of independent-samples t-tests was conducted to compare learners' cognitive load ratings via Leppink's scale differentiating intrinsic, extraneous, and germane cognitive load ratings (Leppink et al., 2013) between Alexa's neutral voice and each Alexa's enthusiastic voice (low-, medium-, and high-enthusiasm). The results showed that compared to Alexa's neutral voice, Alexa's low-enthusiastic voice significantly lowered learners' intrinsic cognitive load ratings, t(125) = 2.18, p = 0.16, d = 0.39. However, Alexa's medium-enthusiastic voice, t(127) = 1.42, p = 0.16, d = 0.25, and Alexa's high-enthusiastic voice, t(83.03) = 1.48, p = 0.14, d = 0.29 did not differently impact learners' intrinsic cognitive load ratings than Alexa's neutral voice. Thus, H4(a), which predicted that Alexa's enthusiastic voices than neutral voice would significantly reduce learners' intrinsic cognitive load ratings, was only endorsed for Alexa's low-enthusiastic voice.

Compared to Alexa's neutral voice, learners' extraneous cognitive load ratings were significantly decreased with Alexa's low-enthusiastic voice, t(125) = 2.26, p = 0.03, d = 0.40, and with Alexa's high-enthusiastic voice, t(110) = 2.21, p = 0.03, d = 0.42. However, Alexa's medium-enthusiastic voice did not affect learners' extraneous cognitive load ratings differently than Alexa's neutral voice, t(117.76) = 1.50, p = 0.13, d = 0.27. Hence, H4(b), which predicted that Alexa's voice enthusiasm would significantly reduce learners' extraneous cognitive load ratings, was only affirmed for Alexa's low-enthusiastic voice and Alexa's high-enthusiastic voice.

Compared to Alexa's neutral voice, learners' germane cognitive load ratings were not impacted differently by Alexa's low-enthusiastic voice, t(125) = 0.01, p = 0.99, d = 0.00, medium-enthusiastic voice, t(127) = 0.80, p = 0.42, d = 0.14, and high-enthusiastic voice, t(110) = 0.76, p = 0.45, d = 0.14. Therefore, these results refuted H4(c), which predicted that Alexa's enthusiastic voices would significantly reduce learners' germane cognitive load ratings than Alexa's neutral voice among non-native English speakers.

6.10 Do learners perform better on the posttest with Alexa enthusiastic voices than Alexa's neutral voice?

This study performed a series of independent-samples t-tests to compare learners' retention scores and transfer scores between Alexa's neutral voice and each Alexa's enthusiastic voice (low-, medium-, and high-enthusiasm).

Concerning retention scores, learners did not perform better with Alexa's low-enthusiastic voice, t(120.11) = 0.33, p = 0.74, d = 0.06, medium-enthusiastic voice t(127) = 0.30, p = 0.76, d = 0.05, or high-enthusiastic voice, t(83.10) = 1.14, p = 0.26, d = 0.22, than Alexa's neutral voice. Likewise, for transfer scores, Alexa's low-enthusiastic voice, t(125) = 0.33, p = 0.35, d = 0.17, medium-enthusiastic voice t(127) = 0.24, p = 0.82, d = 0.04, and high-enthusiastic voice, t(110) = 0.55, p = 0.59, d = 0.10, did not lead to better performance than Alexa's neutral voice. These results rejected H5 that Alexa's enthusiastic voices would significantly enhance learners' posttest performance than Alexa's neutral voice, plausibly because using immediate rather than delayed posttest may not provide adequate sensitivity to discern deep learning effects (Horovitz & Mayer, 2021; Lawson et al., 2021a, b; Wang et al., 2022).

6.11 Summary of findings

Table 3 summarizes Alexa's voice enthusiasm effects on social, affective-motivational, cognitive, and learning outcomes.

7 Discussion

Extending the Cognitive Affective Model of E-learning model to modern text-to-speech voice, this study found that learners inferred more expressed enthusiasm projected by Alexa's medium- and high-enthusiastic voices than Alexa's neutral voice. Noteworthy, the learners did not assign higher enthusiasm ratings for Alexa's low-enthusiastic voice than Alexa's neutral voice. Therefore, a modern text-to-speech voice enthusiasm's intensity seems to be a significant factor. Alexa's low-enthusiastic voice might not have exhibited a strong enough enthusiasm cue for learners to recognize the emotional tone differently from Alexa's neutral voice. Generally, this study highlights the modern text-to-speech vocalizer's potential in conveying synthetic emotional cues, such as voice enthusiasm, that are sufficiently natural, realistic, and expressive for learners to decipher the emotional tones accurately.

However, Alexa's voice enthusiasm did not enhance Alexa's persona ratings as predicted by the Positivity Principle (Lawson & Mayer, 2021), which was surprising in light of recent studies highlighting the prospect of positive and active emotional tones (such as enthusiasm) for elevating instructor's persona ratings compared to negative and passive emotional tones (Lawson et al., 2021a, b) and neutral/calm emotional expression (Liew et al., 2017, 2020). This observation could be attributed to the notion that while Alexa's neutral voice does not exude high-activating (high-arousal) emotional expression like Alexa's enthusiastic voices, the neutral voice, regardless, has a warm and friendly tone expressing positive valence and favorable social signals to the learners, thereby eliciting similar persona ratings for Alexa's enthusiastic and neutral voices. Indeed, learners assigned Alexa's persona ratings relatively high for facilitating learning, credibility, human-like, engaging across all of Alexa's voices (means indicating above 3.5 out of 7), thereby substantiating the

Table 3 Summary of Alexa's voice enthusiasm effects

	Alexa's low-enthusiastic voice	Alexa's medium-enthusiastic voice	Alexa's high-enthusiastic voice
Perceived voice enthusiasm	No effect	Exuded more recognizable enthusiasm cues than Alexa's neutral voice	Exuded more recognizable enthusiasm cues than Alexa's neutral voice
Alexa's persona ratings	No effect	No effect	No effect
Positive emotions	No effect	Induced a significant increase of positive emotions from baseline positive emotions	No effect
Intrinsic motivation	No effect	No effect	No effect
Cognitive load ratings	Decreased intrinsic load and extraneous load than Alexa's neutral voice	No effect	Decreased extraneous load than Alexa's neutral voice
Learning performance	No effect	No effect	No effect

premise that learners could build positive social connections with contemporary text-to-speech voices, irrespective of synthetic enthusiasm cues. Implicatively, an instructor's persona ratings may not necessarily be influenced by arousal (active/passive) as much as valence (positive/negative) cues expressed by a modern text-to-speech voice.

Based on the Integrated-Cognitive Affective Model of Learning with Multimedia (Plass & Kaplan, 2015), voice enthusiasm can be regarded as an emotional design feature in a multimedia learning environment that potentially evokes more positive emotions and intrinsic motivation in learners (Beege et al., 2020; Liew et al., 2017; Wang et al., 2022). This premise resonates with the literature on instructor's enthusiasm in classroom settings where an enthusiastic teaching style exudes positive-activating emotions such as joy, excitement, and pleasure toward the learning subject (Keller et al., 2016), subsequently impacting learners' emotional states, interest, and motivation (Keller et al., 2014, 2016; Kunter et al., 2011, 2013; Moè, 2016). However, this study's results rendered partial empirical support for this notion. Notably, while Alexa's medium-enthusiastic voice prompted learners to report a significant increase in positive emotions compared to their baseline positive emotions, each Alexa's enthusiastic voice did not cause learners to experience more positive emotions compared to Alexa's neutral voice. At the same time, Alexa's enthusiastic voices did not elevate learners' intrinsic motivation ratings compared to Alexa's neutral voice. Collectively, this findings contravene prior research demonstrating that learners' positive emotions and intrinsic motivation could boost when a pedagogical agent expressed more enthusiastic than neutral/calm non-verbal cues (Liew et al., 2017; Wang et al., 2022). On the other hand, this study parallels Beege et al. (2020)'s main experiment revealing that an enthusiastic voice than a neutral voice did not lead learners to report higher positive-activating emotions. As noted by Beege et al. (2020), enthusiasm cues derived from voice alone may not be strong enough to prompt learners to consciously perceive and thus assign more positive affective-motivational ratings, unlike enthusiasm cues exhibited through full-range nonverbal expressions encompassing facial display, voice, and body gesture (Liew et al., 2017; Wang et al., 2022).

This study revealed a nuanced picture concerning the cognitive load effects of Alexa's voice enthusiasm. More specifically, Alexa's low-enthusiastic voice reduced learners' intrinsic cognitive load and extraneous cognitive load ratings compared to Alexa's neutral voice. Whereas, Alexa's high-enthusiastic than neutral voice contributed to lower learners' extraneous cognitive load ratings. Interestingly, similar effects were not manifested with Alexa's medium-enthusiastic voice; that is, the medium-enthusiastic voice did not differently affect learners' perceived difficulty ratings regarding the learning topic (intrinsic cognitive load) and how the information was presented (extraneous cognitive load) compared to Alexa's neutral voice. Situated within the Emotional Design model, learners might have perceived Alexa's low- and high-enthusiastic voices as aesthetically pleasing, prompting the learners to superficially regard the instructional subject and the multimedia presentation as easier to understand due to the "what is beautiful is usable" bias (Brom et al., 2018; Tractinsky et al., 2000). Nevertheless, the Emotional Design interpretation can not be regarded as definitive, constrained by the weak effects of Alexa's voice enthusiasm on learners' affective-motivational ratings.

Thus, Alexa's voice enthusiasm effects in decreasing perceived difficulty (intrinsic and extraneous cognitive load ratings) might also be attributed to other mechanisms, such as the Dr. Fox effect, in which an enthusiastic instructor's charisma and expressiveness can "seduce" learners to believe superficially that they have learned significantly from the course (Keller et al., 2016) or that voice enthusiasm can capture learners' attention more effectively while reducing their tendency to become distracted by other stimuli (Moe et al., 2021). The latter point might be particularly relevant to this online study's context (non-laboratory experiment), in which learners might have been exposed to environmental distraction that competes for cognitive resources during the learning engagement, thereby influencing intrinsic and extraneous cognitive load ratings. This study demonstrated that Alexa's voice enthusiasm did not impact learners' cognitive germane load ratings, failing to reproduce prior findings that strong-prosodic voices (e.g., enthusiastic voice) than weak-prosodic voices (e.g., calm voice) could decrease germane load among non-native English speakers (Davis et al., 2019; Liew et al., 2020). Hence, this finding implies that the voice enthusiasm effect on germane cognitive load may not be robust due to susceptibility to confounding factors. In support of this point, Wang et al. (2022) found that voice enthusiasm influenced learners' germane cognitive load ratings in one experiment but not in the other two experiments, signifying an inconsistent enthusiasm effect on germane cognitive load.

Alexa's enthusiastic voices did not lead to better learning performance than Alexa's neutral voice. From the Cognitive Affective Model of E-learning model and the Positivity Principle perspective (Lawson & Mayer, 2021; Lawson et al., 2021a, b, c), Alexa's voice enthusiasm did not lead to more positive social connections and therefore did not lead to enhanced learning performance. On the other hand, this finding could also be considered through the Integrated-Cognitive Affective Model of Learning with Multimedia and Emotional Design hypothesis (Beege et al., 2020; Domagk et al., 2010; Liew et al., 2017; Plass & Kaplan, 2015). Alexa's voice enthusiasm did not evoke more positive emotions and intrinsic motivation and thus did not promote learning performance. While Alexa's voice enthusiasm influenced intrinsic and extraneous cognitive load ratings to some degree, Alexa's enthusiastic voices did not lead to better learning performance than Alexa's neutral voice, even though there were significant correlations between intrinsic and extraneous cognitive load ratings and retention scores. Noteworthy, this study's finding could have been attributed to the immediate posttest, which might not have effectively detected voice enthusiasm effects on learning outcomes (Wang et al., 2022). A delayed posttest may be more sensitive for discerning voice effects on learning performance concerning deep learning (Lawson et al., 2021b) and long-term knowledge acquisition (Davis et al., 2019).

8 Implications for theory and practice

This study offers a first look into the social, affective-motivational, cognitive, and learning effects of a modern text-to-speech voice enthusiasm. As presented in Table 3, the data showed that different Alexa voice enthusiasm intensities produced distinct multimedia learning benefits and outcomes. In summary, Alexa's low-enthusiastic voice was not inferred by learners as more enthusiastic but could positively affect learners' intrinsic and extraneous cognitive load than the neutral voice. Alexa's medium-enthusiastic voice was recognized as more enthusiastic than the neutral voice and could induce learners to experience more positive emotions upon the learning engagement compared to their baseline positive emotions. Alexa's high-enthusiastic voice was inferred as more enthusiastic and could positively affect learners' extraneous cognitive load than the neutral voice.

Overall, the research outcomes contribute to contemporary theories surrounding voice effects in multimedia learning. First, our findings support an essential conjecture in the Cognitive Affective Model of E-Learning by affirming that learners can recognize expressed emotional tones by artificial text-to-speech voice, which extends from similar effects through human voice emotions (Lawson & Mayer, 2021). Based on the Social Agency Theory, the original voice principle opposes using robotic and monotonous voices produced by classical text-to-speech systems as they cannot evoke a sense of social presence in learners (Atkinson et al., 2005; Mayer, 2014). However, our findings reinforce Craig and Schroeder (2017)'s and (2019)'s proposition that modern text-to-speech engines can now effectively prime social connections with learners. From the Integrated-Cognitive Model of Learning with Multimedia theory perspective (Plass & Kaplan, 2015), our study renders initial support for framing synthetic text-to-speech voice enthusiasm as an Emotional Design feature because of its ability to elicit changes in learners' emotional states. The artificial voice enthusiasm decreased the learners' cognitive load, thereby supporting the Emotions as Facilitator perspective that emotional design attributes can promote multimedia learning while not imposing nonessential cognitive load (Plass & Kalyuga, 2019; Plass & Kaplan, 2015).

At this juncture, we acknowledge that the study's findings and theoretical scope did not provide additional insights into why and how different Alexa's voice enthusiasm intensities can elicit distinct social, affective-motivational, cognitive, and learning outcomes. Further theoretical exploration should be conducted to extend this study. Social cue's intensity or strength in a multimedia learning environment can influence learners' social schemata activation differently, consequently leading to diverse effects on cognitive load and the selection, organization, integration, and retrieval processes during multimedia learning (Schneider et al., 2021). A text-to-speech voice enthusiasm intensity may be somewhat linear to a predicted outcome — for instance, this study's learners could infer more enthusiasm tone from Alexa's medium- and high-enthusiastic voices but not from Alexa's low-enthusiastic voice. On the other hand, the data indicate that voice enthusiasm intensity effects on cognitive load factors are less straightforward, thereby warranting further empirical research to clarify the theoretical link between cognitive load and voice enthusiasm.

From an educational praxis perspective, Alexa's persona ratings data generally support that modern text-to-speech voices can effectively prime learners to build social connections with the instructor in the multimedia learning environment, irrespective of enthusiasm cues. Hence, contemporary vocalizers could be acknowledged as ideal for multimedia learning (Craig & Schroeder, 2017, 2019). Modern text-to-speech voice can replace human-recorded speech for an easier, more rapid, and cheaper multimedia learning production, especially amid the COVID-19 pandemic that catalyzes an escalating e-learning demand. Amazon Alexa is presently one of the scant English language text-to-speech engines that can infuse natural-sounding emotional tones like enthusiasm and disappointment emotional tones into the artificial voice. It is worth noting that the text-to-speech function is nestled within the Amazon Alexa Developer Console environment primarily used by developers working on AI voice assistants. Therefore, it is currently not packaged as an easy-to-use vocalizer system for general commercial use. Nonetheless, Amazon probably would extend Alexa's text-to-speech vocalizer service beyond the AI voice assistant developers' sphere for the broader consumer market in the future. Emerging text-to-speech service providers (e.g., Typecast.AI and Azure Neural Text to Speech) deliver high-quality synthetic emotional voices, including positive expressions representing happiness, hope, and excitement, as well as negative tones exuding sadness, anger, and terrified state. Because learning is fundamentally social, this technological advent potentially augment learning environments with synthetic voice emotions, including enthusiasm and disappointment, to align with how human instructors strategically express positive and even negative emotional tones for promoting learning in traditional classrooms (Keller et al., 2016; Kleef et al., 2011; Liew et al., 2022; Tunstall & Gsipps, 1996).

9 Limitations

This study's findings should be considered in light of this study's scope and limitations. First, this study employed an online instead of a laboratory experiment. While the online experiment promotes higher ecological and external validity, this method generally holds lower control for removing extraneous factors that can influence the findings. Thus, future laboratory experiments can deliver high internal validity in providing precise assessment regarding text-to-speech voice enthusiasm effects on social, affective-motivational, cognitive, and learning outcomes. Additionally, reproducing text-to-speech voice enthusiasm with a longer-duration multimedia learning material can clarify whether extended exposure to a text-to-speech enthusiastic voice can cause habituation (Beege et al., 2020) or adverse effects due to learners perceiving the synthetic emotion as annoying, uncanny, or fake (Liew et al., 2016). Future studies in this research stream should consider utilizing a delayed posttest, which is more sensitive than an immediate posttest for measuring voice effects on deep learning and long-term knowledge acquisition (Davis et al., 2019; Horovitz & Mayer, 2021; Lawson et al., 2021b).

We also recognize the research scope and drawbacks of our learners' profile. As this study's sample learners comprised mostly Malaysian Chinese who are non-native English speakers, the findings cannot be generalized to other cultures or native English speakers. We collected information about learners' English proficiency using a self-reported survey rather than an objective fluency test. This may not indicate a sufficiently accurate representation of the learners' English competence. Further, we did not ask about learners' familiarity with using and listening to text-to-speech voices, although this factor may influence the research outcomes differently. Because voice effects and sociocultural and linguistic factors are concentric (Davis et al., 2019; Mayer et al., 2003; Rey & Steib, 2013; Schneider et al., 2015), further research is needed to explore how text-to-speech voice enthusiasm affects learners from diverse cultures and demographic profiles differently.

10 Future outlook

We outline some research avenues for expanding the confluence between text-to-speech technology and multimedia learning. This research specified the text-to-speech technology's scope to create a speech for narrating a multimedia learning environment. Apart from this function, one of our reviewers pointed out that text-to-speech applications framed as assistive tools can be particularly valuable for special-needs learners with reading and writing challenges (Bone & Bouck, 2017; Evmenova & Regan, 2019). This domain advances new questions on how text-to-speech voice enthusiasm or other emotional tones can benefit and affect these learners' affective-motivational, cognitive load, and learning outcomes. Furthermore, text-to-speech applications have facilitated pronunciation and speaking exercises for language learning (Liakin et al., 2017; Qian et al., 2018; Shadiev et al., 2018; Zhang & Zou, 2022). Therefore, future research can explore how artificial voice emotions can aid language learning.

We recommend future works exploring other text-to-speech emotional tones, including negative ones, such as Alexa's synthetic disappointed or angry voice tones for multimedia learning. This recommendation hinges on the emerging literature accentuating the potential benefits of negative emotional tones in augmenting learning effort and performance in a multimedia learning environment (Jeong et al., 2017; Liew et al., 2022; Sullins et al., 2009), which simulates instructors' strategic use of positive and negative emotional expressions to promote learning in the educational milieu (Tunstall & Gsipps, 1996; Van Doorn et al., 2014). The state-of-the-art synthetic voice engines like Amazon Alexa, Typecast.AI, and Microsoft Azure, which provide myriad artificial emotional tones, open up prospects for researchers and practitioners to enrich multimedia learning through socio-emotional cues.

The voice principle discourages using foreign-accented voices because learners may negatively perceive these voices, consequently impeding learning (Mayer, 2014; Mayer et al., 2003). Nonetheless, an essential caveat is that the voice principle considers learners' linguistic and cultural profiles. Voice dialects can convey familiarity cues to learners with specific linguistic and cultural traits, enhancing learning interest, effort, and performance (Rey & Steib, 2013). Modern text-to-speech technologies, e.g., Vonage and Sanas enable natural-sounding voice variants with a wide array of dialects and accents (Dale, 2022). Thus, this technology advent gives rise to interesting opportunities for researchers and practitioners to explore how matching artificial voice accents and dialects with distinct learner linguistic and cultural traits can promote affective-motivational, cognitive, and learning outcomes in multimedia learning environments.

Appendix A: Narrated content in the multimedia lesson

Hi! I am Alexa, your virtual tutor!

In this video, we will learn what DDOS is and how DDOS attacks servers. DDOS refers to distributed denial of service. DDOS is a cyber-attack on a specific server or network to disrupt daily operations by sending fake requests to overwhelm the server or network. The reason causing DDOS is that the attacker does it for financial reasons, stealing classified data for sale to other parties, or political reasons like the attacker does not like the target, or the attacker does it for fun.

In this section, we will learn how DDOS attacks a server. Here we have a web server. This web server could be owned by companies that sell their products or show information about their company over the internet. Here we have some clients searching the website for a product, service, or information.

To begin a DDOS, the attacker will develop malicious software as a trojan horse and embed it into an email attachment. Then the attacker starts sending large volumes of infected emails to all addresses. If a user opens the email attachment, the trojan will infect the PC and spread through the internal network without warning. So now, all infected PCs will become an army of other infected devices to perform a DDOS attack. This army of infected devices is called botnets. Botnets could be hundreds or thousands of infected devices scattered worldwide.

Now, these botnets can be controlled like an army and waiting to receive instructions from the attacker, like a centralized command and control center. Then the attacker can send out attack commands to the controller that controls the botnets. The attacker can give an order to the controller to tell the botnets to attack a server at a specific time and date. Once the set time reaches, the attack will begin. The botnet will start constantly sending fake synchronize message requests to the server.

Once the server receives the spoofed synchronize, the server will send a synchronize acknowledge back. We know that the synchronize requests sent by the botnets are fake. Synchronize acknowledgments sent by the server will go to an unknown place, and the connections could not establish. This action will eat up server resources like CPU and network bandwidth. In the end, the server breaks down. At this moment, the attacker has access to the server to steal classified company information, purchasing records, or customer data stored in that server.

Acknowledgements This research was funded by Multimedia University Malaysia under IR Fund with ID: MMUI/210127

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

The authors declare that they have no competing interests.

References

Adnan, E., Pillai, S., & Chiew, P. S. (2019). The level of awareness and production of English lexical stress among English language teacher trainees in Malaysia. *Indonesian Journal of Applied Linguistics, 9*(1), 98–107.

Ali, N. L. (2013). A changing paradigm in language planning: English-medium instruction policy at the tertiary level in Malaysia. *Current Issues in Language Planning, 14*(1), 73–92.

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.

Ba, S., Stein, D., Liu, Q., Long, T., Xie, K., & Wu, L. (2021). Examining the effects of a pedagogical agent with dual-channel emotional cues on learner emotions, cognitive load, and knowledge transfer performance. *Journal of Educational Computing Research, 59*(6), 1114–1134.

Baylor, A. L., & Kim, S. (2009). Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior, 25*(2), 450–457.

Beege, M., Schneider, S., Nebel, S., & Rey, G. D. (2020). Does the effect of enthusiasm in a pedagogical Agent's voice depend on mental load in the Learner's working memory? *Computers in Human Behavior, 112*, 1–11.

Bone, E. K., & Bouck, E. C. (2017). Accessible text-to-speech options for students who struggle with reading. *Preventing School Failure: Alternative Education for Children and Youth, 61*(1), 48–55.

Brom, C., Hannemann, T., Starkova, T., Bromová, E., & Deˇchteˇrenko, F. (2017). The role of cultural background in the personalization principle: Five experiments with Czech learners. *Computers & Education, 112*, 37–68.

Brom, C., Starkova, T., & Mello, S. K. (2018). How effective is emotional design? A meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning. *Educational Research Review, 25*, 100–119.

Cambre, J., Colnago, J., Maddock, J., Tsai, J., & Kaye, J. (2020). Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Chan, K. Y., Lyons, C., Kon, L. L., Stine, K., Manley, M., & Crossley, A. (2020). Effect of on-screen text on multimedia learning with native and foreign-accented narration. *Learning and Instruction, 67*, 1–11.

Christensen, R., Knezek, G., & Tyler-Wood, T. (2014). Student perceptions of science, technology, engineering and mathematics (STEM) content and careers. *Computers in Human Behavior, 34*, 173–186.

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education, 114*, 193–205.

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.

Dale, R. (2022). The voice synthesis business: 2022 update. *Natural Language Engineering, 28*(3), 401–408.

Davis, R. O., Vincent, J., & Park, T. (2019). Reconsidering the voice principle with non-native language speakers. *Computers & Education, 140*, 1–12.

Debue, N., De, V., & Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology, 5*, 1–12.

Domagk, S. (2010). Do pedagogical agents facilitate learner motivation and learning outcomes? *Journal of Media Psychology, 22*(2), 84–97. https://doi.org/10.1027/1864-1105/a000011

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior, 26*(5), 1024–1033.

Evmenova, A. S., & Regan, K. (2019). Supporting the writing process with technology for students with disabilities. *Intervention in School and Clinic, 55*(2), 78–85.

Fountoukidou, S., Matzat, U., Ham, J., & Midden, C. (2021). The effect of an artificial agent's vocal expressiveness on immediacy and learning. *Journal of Computer Assisted Learning, 38*(2), 500–512.

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence, 20*(2), 507–537.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science, 2*(3), 96–100.

Hillaire, G., Iniesto, F., & Rienties, B. (2019). Humanising text-to-speech through emotional expression in online courses. *Journal of Interactive Media in Education*, 1-9. https://doi.org/10.5334/jime.519

Horovitz, T., & Mayer, R. E. (2021). Learning with human and virtual instructors who display happy or bored emotions in video lectures. *Computers in Human Behavior, 119*, 1–8.

Jeong, D. C., Feng, D., Krämer, N. C., Miller, L. C., & Marsella, S. (2017). Negative feedback in your face: examining the effects of proxemics and gender on learning. In *International conference on intelligent virtual agents* (p. 170–183). Springer.

Johnson, G., & Connelly, S. (2014). Negative emotions in informal feedback: The benefits of disappointment and drawbacks of anger. *Human Relations, 67*(10), 1265–1290.

Jungert, T., Levine, S., & Koestner, R. (2020). Examining how parent and teacher enthusiasm influences motivation and achievement in STEM. *The Journal of Educational Research, 113*(4), 275–282.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need. *Educational Psychology Review, 23*(1), 1–19.

Keller, M. M., Goetz, T., Becker, E. S., Morger, V., & Hensley, L. (2014). Feeling and showing: A new conceptualization of dispositional teacher enthusiasm and its relation to students' interest. *Learning and Instruction, 33*, 29–38.

Keller, M. M., Hoy, A. W., Goetz, T., & Frenzel, A. C. (2016). Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review, 28*(4), 743–769.

Kim, T., & Schallert, D. L. (2014). Mediating effects of teacher enthusiasm and peer enthusiasm on students' interest in the college classroom. *Contemporary Educational Psychology, 39*(2), 134–144.

Kleef, G. A. V., Doorn, E. A. V., Heerdink, M. W., & Koning, L. F. (2011). Emotion is for influence. *European Review of Social Psychology, 22*(1), 114–163.

Kunter, M., Frenzel, A., Nagy, G., Baumert, J., & Pekrun, R. (2011). Teacher enthusiasm: Dimensionality and context specificity. *Contemporary Educational Psychology, 36*(4), 289–301.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820.

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*. https://doi.org/10.1007/s40593-021-00282-y

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021a). Do learners recognize and relate to the emotions displayed by virtual instructors? *International Journal of Artificial Intelligence in Education, 31*(1), 134–153.

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021b). The positivity principle: Do positive instructors improve learning from video lectures. *Educational Technology Research and Development, 69*, 3101–3129.

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021c). Recognizing the emotional state of human and virtual instructors. *Computers in Human Behavior, 114*, 1–9.

Lee, H., & Mayer, R. E. (2018). Fostering learning from instructional video in a second language. *Applied Cognitive Psychology, 32*(5), 648–654.

Leppink, J., Paas, F., Vleuten, C. P. V. D., Gog, T. V., & Merriënboer, J. J. V. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*(4), 1058–1072.

Liakin, D., Cardoso, W., & Liakina, N. (2017). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning, 30*(3–4), 325–342.

Liew, T. W., Tan, S. M., & Kew, S. N. (2022). Can an angry pedagogical agent enhance mental effort and learning performance in a multimedia learning environment? *Information and Learning Sciences*, 1-22. https://doi.org/10.1108/ILS-09-2021-0079

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.

Liew, T. W., Zin, N. A. M., & Sahari, N. (2017). Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment. *Human-Centric Computing and Information Sciences, 7*(1), 1–21.

Liew, T. W., Zin, N. A. M., Sahari, N., & Tan, S.-M. (2016). The effects of a pedagogical agent's smiling expression on the learner's emotions and motivation in a virtual learning environment. *The International Review of Research in Open and Distributed Learning, 17*(5), 1–19.

Liu, Y., Jang, B. G., & Roy-Campbell, Z. (2018). Optimum input mode in the modality and redundancy principles for university ESL students' multimedia learning. *Computers & Education, 127*, 190–200.

Alonso Martin, F., Malfaz, M., Castro-González, Á., Castillo, J. C., & Salichs, M. Á. (2020). Four-features evaluation of text to speech systems for three social robots. *Electronics, 9*(2), 267. https://doi.org/10.3390/electronics9020267

Matthew, G. (2020). The effect of adding same-language subtitles to recorded lectures for non-native, English speakers in e-learning environments. *Research in Learning Technology, 28*(1), 16. https://doi.org/10.25304/rlt.v28.2340

Mayer, R. E. (2014). Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In *The Cambridge handbook of multimedia learning* (vol. 16, p. 345–370). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369

Mayer, R. E. (2020). Searching for the role of emotions in e-learning. *Learning and Instruction, 70*, 1–3.

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), 239–239.

Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology, 28*(5), 653–660.

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), 419–425.

Moè, A. (2016). Does displayed enthusiasm favour recall, intrinsic motivation and time estimation. *Cognition and Emotion, 30*(7), 1361–1369.

Moe, A., Frenzel, A. C., Au, L., & Taxer, J. L. (2021). Displayed enthusiasm attracts attention and improves recall. *British Journal of Educational Psychology, 91*(3), 911–927.

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

Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research, 19*(4), 504–527.

Osada, N. (2001). What strategy do less proficient learners employ in listening comprehension?: A reappraisal of bottom-up and top-down processing. *Journal of Pan-Pacific Association of Applied Linguistics, 5*(1), 73–90.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341.

Peters, J. (2019). *Alexa's voice can now express disappointment and excitement*. The Verge. Retrieved from https://www.theverge.com/2019/11/26/20984629/amazon-alexa-voice-disappointment-empathetic-happy-excited-newscaster-music-us-australia. Accessed 26 July 2022.

Pillai, S., & Ong, L. T. (2018). English (es) in Malaysia. *Asian Englishes, 20*(2), 147–157.

Plass, J. L., & Kaplan, U. (2015). Emotional design in digital media for learning. In S. Y. Tettegah & M. Gartmeier (Eds.), *Emotions, technology, design, and learning* (pp. 131–161). San Diego: Academic Press.

Plass, J. L., Bruce, D., Homer, A., Macnamara, T., Ober, M. C., Rose, S., . . ., & Olsen (2020). Emotional design for digital games for learning: The effect of expression, color, shape, and dimensionality on the affective quality of game characters. *Learning and instruction*, *70*, 1-13.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review, 31*(2), 339–359.

Poushneh, A. (2021). Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services, 58*, 1–10.

Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology, 22*(1), 69–96.

Rajadurai, J. (2006). Pronunciation issues in non-native contexts: A Malaysian case study. *Malaysian Journal of ELT Research, 2*(1), 42–59.

Ramli, N. F., & Talib, O. (2017). Can education institution implement STEM? From Malaysian teachers' view. *International Journal of Academic Research in Business and Social Sciences, 7*(3), 721–732.

Rey, G. D., & Steib, N. (2013). The personalization effect in multimedia learning: The influence of dialect. *Computers in Human Behavior, 29*(5), 2022–2028.

Rodero, E., & Lucas, I. (2021). Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society*, 1–19. https://doi.org/10.1177/14614448211024142

Schneider, S., Beege, M., Nebel, S., Schnaubert, L., & Rey, G. D. (2021). The cognitive-affective-social theory of learning in digital environments (CASTLE). *Educational Psychology Review*, 1–38.

Schneider, S., Nebel, S., Pradel, S., & Rey, G. D. (2015). Introducing the familiarity mechanism: A unified explanatory approach for the personalization effect and the examination of youth slang in multimedia learning. *Computers in Human Behavior, 43*, 129–138.

Shadiev, R., Hwang, W. Y., & Liu, T. Y. (2018). A study of the use of wearable devices for healthy and enjoyable English as a foreign language learning in authentic contexts. *Journal of Educational Technology & Society, 21*(4), 217–231.

Sullins, J., Craig, S. D., & Graesser, A. C. (2009). Tough love: The influence of an agent's negative affect on students' learning. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Artificial intelligence in education, building learning systems that Care: From knowledge representation to affective modeling* (pp. 677–679). Washington, DC: IOS Press.

Sweller, J., Merrienboer, J. J. V., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Thirusanku, J., & Yunus, M. M. (2014). Status of english in Malaysia. *Asian Social Science, 10*, 254–260.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with computers, 13*(2), 127–145.

Tunstall, P., & Gsipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22*(4), 389–404.

Van Doorn, E. A., Van Kleef, G. A., & Van Der Pligt, J. (2014). How instructors' emotional expressions shape students' learning performance: The roles of anger, happiness, and regulatory focus. *Journal of Experimental Psychology: General, 143*(3), 980–984.

Viegas, C., & Alikhani, M. (2021). Towards Designing Enthusiastic AI Agents. *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 203–205.

Wang, Y., Feng, X., Guo, J., Gong, S., Wu, Y., & Wang, J. (2022). Benefits of affective pedagogical agents in multimedia instruction. *Frontiers in Psychology, 12*, 1–14. https://doi.org/10.3389/fpsyg.2021.797236

Westlund, K., Jeong, J. M., Park, S., Ronfard, H. W., Adhikari, S., Harris, A., . . ., & L, C. (2017). Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in human neuroscience*, *11*, 1–20.

Wong, R. M., & Adesope, O. O. (2021). Meta-Analysis of emotional designs in multimedia learning: A replication and extension study. *Educational Psychology Review, 33*(2), 357–385.

Yap, T. S., & Pillai, S. (2018). Intonation patterns of questions in Malaysian English. *Asian Englishes, 20*(3), 192–205.

Yu, C. H. (2012). Examining the relationships among academic self-concept, instrumental motivation, and TIMSS 2007 science scores: A cross-cultural comparison of five East Asian countries/regions and the United States. *Educational Research and Evaluation, 18*(8), 713–731.

Zhang, R., & Zou, D. (2022). Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, 35(4), 696-742.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Tze Wei Liew[1] · **Su-Mae Tan**[2] · **Wei Ming Pang**[1,2] · **Mohammad Tariqul Islam Khan**[1] · **Si Na Kew**[3]

✉ Tze Wei Liew
twliew@mmu.edu.my

Su-Mae Tan
smtan@mmu.edu.my

Wei Ming Pang
weimingpang@hotmail.com

Mohammad Tariqul Islam Khan
tariqul.islam@mmu.edu.my

Si Na Kew
snkew@utm.my

[1] Human-Centric Technology Interaction SIG, Faculty of Business, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia

[2] Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia

[3] Language Academy, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia, Jalan Iman, 81310 Skudai, Johor, Malaysia