A Multimodal Human-Computer Interaction System and Its Application in Smart Learning Environments

Jiyou Jia(✉), Yunfan He, and Huixiao Le

Department of Educational Technology, School of Education, Peking University, Yiheyuanlu 5,
Beijing 100871, China
{jjy,heyunfan,lehuixiao}@pku.edu.cn

Abstract. A multimodal human-computer interaction system is composed of the comprehensive usage of various input and output channels. For the information input, apart from the traditional keyboard typing, mouse clicking, screen touching, the latest speech and face recognition technology can be used. For the output, the traditional screen display, the latest speech and facial expression synthesis and gesture generation can be used. After literature review of related works, this paper at first presents such a system, MMISE (Multimodal Interaction System for Education), about its architecture and working mechanism, POOOIIM (Pedagogical Objective Oriented Output, Input and Implementation Mechanism) illustrated with practical examples. Then this paper introduces this system's pilot applications in the epidemic time of novel coronavirus in 2020.

Keywords: Multimodal human-computer interaction · Intelligent tutoring system · Smart learning environment · Mathematics instruction · Learning of English as a foreign language

1 Concept Definition and Research Question

A multimodal human-computer interaction system "seeks to leverage natural human capabilities to communicate via speech, gesture, touch, facial expression, and other modalities, bringing more sophisticated pattern recognition and classification methods to human–computer interaction" (Turk 2014). With the rapid advance in artificial intelligence in the past decade, including natural language processing, machine learning and pattern recognition, the multimodal human-computer interaction research is leveraging keyboard-tipping, mouse clicking, speech, touch, vision and gestures. A multimodal human-computer interaction system should comprehensively use various input and output channels. For the information input, apart from the traditional keyboard typing, mouse clicking, screen touching, the latest speech and face recognition technology can be utilized. For the information output, the traditional screen display, the latest speech synthesis, facial expression synthesis and gesture generation can be utilized.

Two questions arise when a system deals with multiple channels' input and output. The first is how to receive and analyze the multiple input information, and the second is how to generate the appropriate multiple output information.

S. K. S. Cheung et al. (Eds.): ICBL 2020, LNCS 12218, pp. 3–14, 2020.
https://doi.org/10.1007/978-3-030-51968-1_1

2 Related Work

A multimodal human-computer interaction system facilitates the human-like interaction. The human-like interaction between the computer and the user supported by multiple modal technology has been applied in educational technology for a long time, especially in the field of pedagogical agent and intelligent tutoring system (Jia 2015). Johnson and Lester (2018) expected that the anthropomorphic and natural interaction between the pedagogical agent and the student could make the learning easier, more motivating and more participatory. A pedagogical agent is a human-like animation or avatar in a multimedia leaning system, and can support the student's learning by playing part of an expert, a teacher, a tutor, or a friend, which can give a lecture, suggestion, or question (Liew et al. 2017; Schroeder and Adesope 2014). The most played roles by the pedagogical agent include tutors, teachers and trainers, but seldom learning companion, as found by some systematic reviews (Poria et al. 2017; Terzidou et al. 2016).

The effect of a pedagogical agent with human-like interaction has been studied early in the 1990s. Lester et al. (1997) explained the function of a pedagogical agent as persona effect based on the experiment, and believed that the human-like agent, even with a weaker interaction function, could effectively promote the learning. Social presence was first proposed by Short et al. (1976), and defined by Gunawardena (1995) and Lowenthal (2009) as the extent how a virtual agent is regarded as a real human that communicates and connects with the agent. Social clues expressed by a pedagogical agent in a multimedia learning environment like the voice, emotion and facial expression can inspire the learners' social contact schema (Louwerse et al. 2009). This argument was evidenced by the fMRI research conducted by Schilbach et al. (2006). If the social contact schema is stimulated, the learner can process the information coming from the computer deeply and facilitate the more meaningful learning (Mayer and DaPra 2012). Nass and Moon (2000) found more similarity between the computer and the human being can stimulate more social contact of the human users.

The previous studies on pedagogical agent used traditional keyboard text input, natural language and eye or face recognition as input channels. Graesser and his research team applied LSA (Latent Semantic Analysis) and other natural language processing technology in designing intelligent tutoring systems that talk with the student via text or spoken voice in dialogue or trialogue form (Graesser 2016; Graesser et al. 2018).

Lepper and Chabay (1988) found through classroom observations the teachers spent almost as much time on the students to help them with the affective goals as with the knowledge goal. The latest advancement of multimodal emotion recognition and micro emotion recognition improves the performance of affective computing (Wu et al. 2016). Recently the design of pedagogical agents to simulate the human teachers in affective intervention has been a research trend. From the perspective of multimedia learning, the pedagogical agent which can identify learners' emotions can effectively convey more social clues. Positive emotions can help learners focus on current tasks, motivate the learners, and improve the learning effect (Pekrun 2006). Liew et al. (2017) found through the meta-analysis about 30 experiments that the usage of emotional recognition and expression can improve the students' learning motivation, knowledge retention and knowledge transformation.

The output channels of pedagogical agents include text, spoken voice, and animation. Atkinson (2002) found that the pedagogical agents using voices could improve the students' learning performance more effectively than the agents using just the texts. The real human voice used by the pedagogical agents could have better effect than the synthesized voice (Mayer and DaPra 2012). IBM developed the Waston Tutor, which could generate human-like voices with different tones corresponding to the instructional goals and communicate with the students in spoken voices (Afzal et al. 2019). Those findings coincide with the social presence theory arguing that the human voice as a more effective social clue can motivate the learners' social communication schema.

The animating avatar used as a pedagogical agent can behave like a real human being, convey more social clues, and thus improve the learners' performance (Dehn and van Mulken 2000). There are two kinds of animating avatars, the anthropoid or animal-like cartoon (Yilmaz and Kiliç-Çakmak 2012), and the real human avatar (Yung and Paas 2015). Although the real human avatar seems more normal, Schroeder et al. (2013) found in their meta-analysis the anthropoid or animal-like pedagogical agents had significant and positive effect on learning performance, but the real human-like agents did not have significant effect on the learning performance. The reason may be that the learners could be aware of the existence of another human subject and lose their control of the learning (Kim et al. 2016).

The gender of the pedagogical agent is also an important factor in the human-computer interaction. The stereotype of human gender in the reality can be reflected in the relation between the learner and the pedagogical agent. Kim and Baylor (2007) found that the students liked the male agents more than the female agents, and achieved better learning gains by using the male agents. Makransky et al. (2019) found that the students got better learning result by using the same gender agents.

Some researchers have attempted to design the multiple channel output responding to learners' recognized expression. The AutoTutor system (D'mello and Graesser 2013) could recognize the learner's affective status, and then respond with appropriate voice, expression and gesture. Prendinger and Ishizuka (2005) detected the learners' affect by using the skin conductivity sensor and then gave corresponding help and concern, and found the learners felt less pressure.

The above reviewed related studies demonstrate that multimodal input or output technology has been applied in pedagogical agent design. However, the comprehensive consideration of multiple channel inputs and outputs and their application in educational scenarios have not been found in our literature review. This study attempts to fill in this gap. We design a multimodal interaction system for education based on the latest advancement in artificial intelligence, illustrate its workflow and mechanism with practical pedagogical scenarios, and introduce its technical evaluation and pilot application in the epidemic time of novel coronavirus outbreak.

3 The Architecture and Mechanism of the MMISE System

We designed and developed a web-based multimodal human-computer interaction system, MMISE (Multimodal Interaction System for Education). Implemented as a client/server architecture, it is comprised of the client program and the server system, as shown in Fig. 1. The client program deals with the capture, detection and recognition of various input signals on the one side, on the other side generates the synthesized voice and avatar animation. It can be an independent program written in Python or Java, or a webpage written with JavaScript and downloaded from the server. It runs in the client operating system like Windows in personal computers or Android in tablet computers or smartphones. The server system mainly deals with the input-output response mechanism. This client/machine architecture can fully use the computing capability of the user's client machine, and relieve the burden of the server machine.

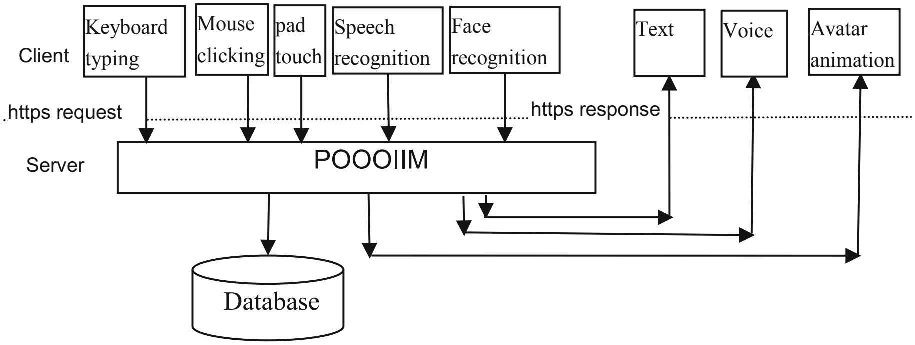


Fig. 1. The architecture of MMISE.

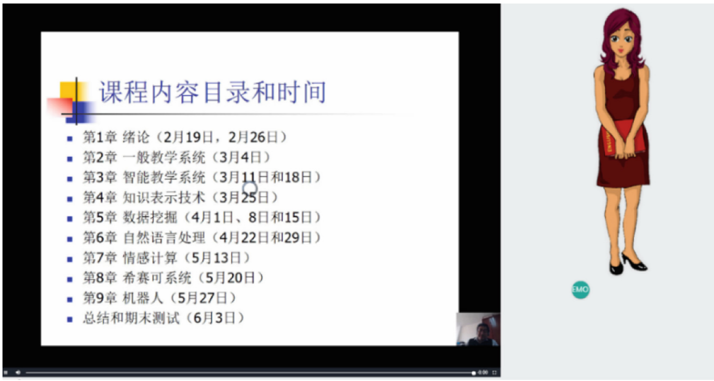
Even if all the inputs in various formats can be detected technologically nowadays, which of them should be captured and recorded into the database for further analysis, depends both on the pedagogical requirement and ethical consideration. Similarly, the output channels should be determined based on the system's pedagogical design objective.

The system can be implemented in three ways: as an independent webpage, as a webpage embedded in other webpages, or as an independent client program running in Windows or other operating system. The user can use the independent webpage or the independent client program, if he or she can concentrate himself or herself on the learning objects. The user should use the webpage format embedded in other webpages, if he or she is easy to lose the concentration or focus on the learning object. Which way to be used is dependent on the pedagogical objective of the system design.

Overall, the input channels, output channels and the implementation approach are all dependent on the pedagogical objectives. This is the mechanism of the MMISE, can be called pedagogical objective oriented output, input, and implementation mechanism, or in abbreviation, POOOIIM or PO^{3}I^{2}M. This mechanism can be explained with the following five examples.

The first example is a learning companion system with a MMISE that is hoped to improve the student's concentration on the learning content displayed on the screen, such as a video lecture on demand or a live video lecture. The student's face captured by the video camera is the most important input channel. For this functionality the continuance of facial identification is an important indicator for the student's concentration status of facing the screen. If the MMISE cannot recognize and identify the student's face continuously for a specified period, for example one minute, it should give out some information in the form of spoken voice, because the student does not face the screen and can't read the text message or watch the animation. In this case, the MMISE should be embedded into the lecture display program.

In Fig. 2, the live lecture during the special epidemic time of China was given to the students from Peking University who were located anywhere in the whole country. In the left division, the live video lecture with the lecturer's speech was synchronously displayed to the student, for example, Jack, while in the right division, the learning companion Emina was watching the student' facial expression. Besides, the companion expressed the same emotion with the student. If the student looked happy by watching the lecture, Emina would also looked happy. If the student watched the video in full screen, the companion Emina was hidden, but was still watching the student. If the student left the screen for one minute, Emina would spoke "Where are you now, Jack?".



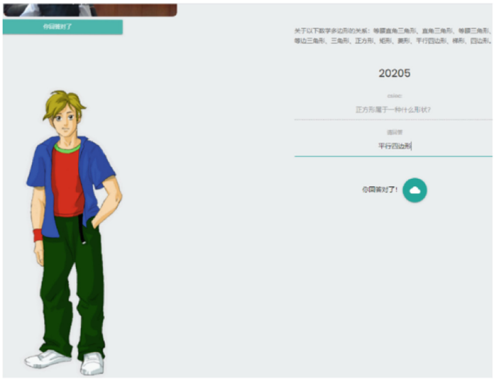
课程内容目录和时间

- 第1章 绪论（2月19日，2月26日）
- 第2章 一般教学系统（3月4日）
- 第3章 智能教学系统（3月11日和18日）
- 第4章 知识表示技术（3月25日）
- 第5章 数据挖掘（4月1日、8日和15日）
- 第6章 自然语言处理（4月22日和29日）
- 第7章 情感计算（5月13日）
- 第8章 希赛可系统（5月20日）
- 第9章 机器人（5月27日）
- 总结和期末测试（6月3日）

Fig. 2. The screenshot of one live lecture companied by the avatar Emina.

The second example is a tutoring system with the MMISE (Zhang and Jia 2017). The student should input the answers either by clicking the correct one in the multiple choices or fill in the blanks in a given time period, for example in 10 min for the drill, because some students lag behind the required drill duration according to our previous analysis of online learners' drilling behavior (Le and Jia 2018). If the student does not click the mouse or type the keyboard during the given time, the MMISE should alert the student. Furthermore, it should determine the concrete output format, either spoken voice, text or animation. If it detects that the student leaves the screen for a longer time, it should give out spoken text. If the student doesn't leave the screen, the MMISE should keep quiet for a quiz writing system, but function as a tutor with appropriate hints or help in the format text, voice or animation for a tutoring system. Figure 3 shows the screenshot of one online math tutoring system with the learning companion Christoph recognizing the user's face. In this case, the MMISE could be embedded into the dialogue program to enhance the agent effect, and some important issues regarding the voice, the avatar gender and features should be considered, as suggested in the literature review.

The third example is a text dialogue system for language learning with the MMISE. The student's input text is an important indicator of his or her speech behavior. If the



你回答对了

关于以下数学多边形的关系：等腰直角三角形、直角三角形、等腰三角形、等边三角形、三角形、正方形、矩形、菱形、平行四边形、梯形、四边形。

20205

csiec:

正方形属于一种什么形状?

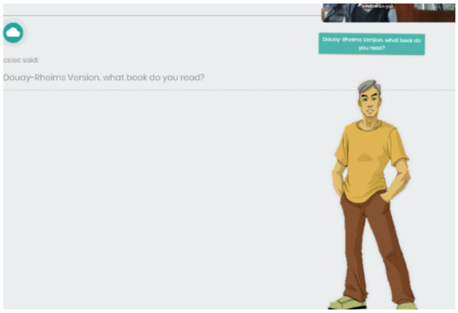
请回答

平行四边形

你回答对了!

Fig. 3. The screenshot of one online math tutoring system with the user's face recognized.

MMISE cannot capture the student' text for a given time period, for example, for ten minutes, it should give some text messages to the student. On the other hand, all the student' input texts should be responded appropriately by the MMISE. Figure 4 shows one screenshot of the CSIEC system (Jia 2009) where the user is sending message to the avatar via keyboard text input, and receives both text output and spoken utterance from the animation avatar Stephan. In this case, the MMISE could be embedded into the dialogue program or be used just as an independent webpage or program, because only text input and output are required for the dialogue system.



csiec said:

Douay-Rheims Version. what book do you read?

Douay-Rheims Version. what book do you read?

Fig. 4. The screenshot of the CSIEC system

The fourth example is a voice speech dialogue system for language learning with the MMISE. The student's speech is an important indicator for his or her speech behavior. If the MMISE cannot capture the student' speech for a given time period, for example, for ten minutes, it should give some utterances to the student. On the other hand, the student' each input should be responded appropriately by the MMISE. In this case, the MMISE should be embedded into the dialogue program, because the voice speech as the important input channel cannot be separated from the program.

The fifth example is a quiz writing system with the MMISE in online learning. The identification of the examinee and the examination process are critical to the guarantee the fairness of the examination, as cheating often happens in MOOC and other online learning systems (Northcutt et al. 2016). The MMISE should work as the proctors or the examiner in the online examination. With the understanding and the agreement of the examinee, all the input data including keyboard typing, mouse action, speech and face movement should be recorded into the database for future review and analysis. To avoid disturbing the examinee's thinking and question answering process, no output is needed to give out throughout the examination. In this case, the MMISE as the monitor or the proctors should be embedded into the quiz writing system.

For the above examples, the textual output can be continuously presented to the user on the screen. But the voice and animation output should be given in some intervals in order to prevent their overlapping.

4 The Software Systems Used to Implement the MMISE

The advancement of artificial intelligence, including NLP (Natural Language Processing) and machine learning such as deep learning, enables both open source and commercial software systems to be used by implementing the MMISE. We just list some systems.

The continuous speech recognition can be realized by Julius (https://github.com/julius-speech/julius), an "Open-Source Large Vocabulary Continuous Speech Recognition Engine", as a high-performance, small-footprint large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and developers. Based on word N-gram and context-dependent HMM, it can perform real-time decoding on various computers and devices from micro-computer to cloud server." (Lee et al. 2019). The English model (LM+DNN-HMM) can be downloaded from the Julius Model page (https://sourceforge.net/projects/juliusmodels/files/).

The face, landmark recognition and facial expression recognition can be realized by the FaceAPI.js (https://github.com/justadudewhohacks/face-api.js), a JavaScript API for Face Detection, Face Recognition and Face Landmark Detection of a photo file, a video file, or live webcam capture. It can be embedded in an existing webpage or realized in an independent webpage located in an https webserver server. For face detection, we use the SSD (Single Shot Multibox Detector) model, which computes the locations of each face in an image and returns the bounding boxes together with its probability for each face. The size of the quantized detection model is about 5.4 MB. For face landmark detection, we use the 68 point face landmark detector. The default model file has a size of only 350 kb. For face recognition, a ResNet-34 like architecture is implemented to compute a face descriptor (a feature vector with 128 values) from any given face image, which is used to describe the characteristics of a person's face. We determine the similarity of two arbitrary faces by computing the Euclidean distance. The model achieves a prediction accuracy of 99.38% on the LFW (Labeled Faces in the Wild) benchmark for face recognition. The size of the quantized recognition model is roughly 6.2 MB. For face expression recognition, we use a lightweight and fast model with a file size of roughly 310 kb. It has been trained on a variety of images from publicly available datasets as well as images scraped from the web.

The speech synthesis technology has been matured. It can be realized by commercial providers, such as Chinese iFlyTek (https://www.iflytek.com/), or online Lifelike Text to

Speech ReadSpeaker (https://www.readspeaker.com), whose web version “webReader allows content interaction on a personal level and offers important learning tools that help understanding and improve retention and make it easier to access online content on the go”, and whose “speechCloud API is an online text-to-speech API for making desktop/web/mobile applications and Internet-connected devices talk.” It can also be realized by open source TTS (Text To Speech) projects, from the old FreeTTS (https://sourceforge.net/projects/freetts/), to Google TTS API, MicrosoftSpeech SDK, and the recent Mozilla TTS (https://github.com/mozilla/TTS), “a part of Mozilla Common Voice. TTS aims a deep learning based Text2Speech engine, low in cost and high in quality.” The latest development of TTS is the Real-Time Voice Cloning which claims to “Clone a voice in 5 s to generate arbitrary speech in real-time” (https://github.com/CorentinJ/Real-Time-Voice-Cloning) (Corentin 2020).

The avatar technology can be implemented through 2D or 3D animation software, like live2d (https://www.live2d.com), or game engine Unity (https://unity.com/). We have developed five avatars using the live2d animation software: Christoph, Stephan, Ingrid, Emina, and Christine, and applied them in the above program examples.

5 The Pilot Application of the MMISE

During the epidemic time of the novel coronavirus outbreak in 2020, all Chinese people including students should stay at home. The schools could not run as usually, and the students could not go to the schools. Therefore all the teachers and the students were separated from each other and might be distributed around the whole country. However, in order to maintain the ordinary teaching plan, the educational authorities encouraged and even required the teachers and students to adopt online learning. The teacher used the computer at home or in the school office. Almost all of the students used their computers at home. The access to the Internet was the necessary condition for the online learning. Fortunately, all most all families in China nowadays have the smart phones, and many families have wired Internet connection via optical fiber. The phones can access the Internet via 3G, 4G or even 5G network provided by the ISP (Information Service Provider) and can function as a hot spot to construct a WLAN (wireless local area network), through which the desktop computer, notebook computer or tablet computer can access the Internet. The connection to the Internet could be better and more stable if the family had a wired network through optic fabrics.

Two forms of online instruction approaches are used. The first is instant, synchronous and bidirectional lecture. The second is asynchronous and one-way learning resource browsing and activity participation.

The synchronous lecture was usually implemented in an online conference or meeting system such as Zoom, Classing, Dingding, Tencent, and others. Normally the teacher used the broadcasting function to present the lecture slides like PowerPoint with synchronous aural explanation just as in the traditional classroom or in the lecture hall. The students at home watched the lecture broadcasting as well as the teacher’s facial expressions, and listened to the teacher’s speech. The teacher could also watch the students’ facial expressions through video transmission. However, the teacher’s main function was holding the lecture, and could not observer the students’ actions for all the time as in the classroom, especially with a large class volume. Because the network bandwidth was limited during the lecture time, the lecture broadcasting together with bidirectional video signals was seldom used. The teacher could not watch the students' facial expression, and could not feel their attitudes and reactions immediately. In such a virtual classroom, the students sitting alone at home could hardly feel the coexistence of other classmates. Only those students with stronger self-regulation could follow the teacher throughout the whole lecture time.

The asynchronous learning resource browsing and activity participation often happened in a course management system such as Moodle or Blackboard. The learning resources include teaching slides file in PowerPoint or other formats, required readings in PDF, Word or other formats, and web pages. The online learning activities include the quiz comprised of multiple choice, blank-filling or other types of questions, the assignment to submit a document or online texts, the feedback or survey, the discussion forum, the Wiki, and online examinations. In the traditional school teaching, similar activities are mostly completed in the paper format. Compared with the traditional paper-format activities, the online learning activities often require the students to have a stronger control over their online learning behavior, because many contents other than the learning content in the Internet, for example, the games, are more attractive to the student, especially the pupils in the schools. If the parents or other adult relatives accompanied the young student or the school pupil, the student could complete the online learning activities on time.

Both forms of pure online learning are different from the traditional classroom learning and the blended learning of traditional classroom learning with computer supported learning including online learning, because the students could not meet the teachers and their classmates on site, and lacked the teachers' face-to-face companion and guidance.

A virtual avatar representing the teacher and accompanying the student's online learning is needed for this special period of virus outbreak. For this practical purpose, we have applied the programs in the examples given in Sect. 3 in the graduate and undergraduate courses in Peking University.

We also applied the program in a web-based intelligent mathematics instruction system for school pupils, "Lexue 100" (https://www.lexue100.com), with the Chinese meaning Happy Learning for 100%. It is a web-based intelligent instruction system for school mathematics, developed by Beijing Lexue 100 Online Education Co., Ltd., and equipped with the OLAI model proposed by the authors' team (Jia and Yu 2017). More than 70,000 quizzes have been designed for the different versions of mathematics textbooks that are used in different provinces and metropolis in China. Writing quizzes is the main learning activity in this system. Each quiz is composed of a series of gap-filling or single-choice questions with predefined standard answers. As soon as one student submits the trial answer to the system, the trial answer can be compared with the standard answer, and the corresponding quiz score and feedback are instantly provided to the student. Users are allowed to pass the quiz only if every answer gets right, meaning that if the first try of one student is wrong, the student will have to try again until the answer hits the point. As we analyzed in previous studies (Jia and Zhang 2019), the school students, especially the students with learning difficulties, are easy to lose their concentration by writing the quiz alone at home, and may copy the answer from other classmates. The learning avatars facilitated with MMISE are embedded in the quiz activity, just like those shown in the above examples, and can improve the student's concentration on the quiz by watching the student's face, and giving appropriate help hints.

6 Limitation and Further Study

This paper just proposes the framework of MMISE (Multiple Modal Interaction System for Education), illustrates its architecture and working mechanism with five pilot examples, and just begins to apply it in education. Its evaluation should be done soon after the user data and evaluation survey result can be collected.

Acknowledgement. The paper is one product of the national educational technology research key project of Central Educational Technology Center China, "Personalized instruction research based on artificial intelligence and big data technology" (Project No.: 176220009). The research is supported by the project "Lexue 100, Smart Education" of Beijing Lexue 100 Online Education Co., Ltd. The authors thank all the teachers and students who have participated in the program.

References

Afzal, S., et al.: The personality of AI systems in education: experiences with the Watson tutor, a one-on-one virtual tutoring system. Childhood Educ. **95**(1), 44–52 (2019)

Atkinson, R.K.: Optimizing learning from examples using animated pedagogical agents. J. Educ. Psychol. **94**(2), 416–427 (2002)

Corentin, J.: Real-time Voice Cloning (2020). https://matheo.uliege.be/handle/2268.2/6801. Accessed 11 Feb 2020

Dehn, D.M., van Mulken, S.: The impact of animated interface agents: a review of empirical research. Int. J. Hum. Comput. Stud. **52**, 1–22 (2000)

D'mello, S., Graesser, A.: AutoTutor and affective AutoTutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. (TiiS) **2**(4), 1–39 (2013)

Graesser, A.C.: Conversations with AutoTutor help students learn. Int. J. Artif. Intell. Educ. **26**(1), 124–132 (2016)

Graesser, A.C., Foltz, P.W., Rosen, Y., Shaffer, D.W., Forsyth, C., Germany, M.-L.: Challenges of assessing collaborative problem solving. In: Care, E., Griffin, P., Wilson, M. (eds.) Assessment and Teaching of 21st Century Skills. EAIA, pp. 75–91. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-65368-6_5

Gunawardena, C.N.: Social presence theory and implications for interaction and collaborative learning in computer conferences. Int. J. Educ. Telecommun. **1**(2), 147–166 (1995)

Jia, J., Yu, Y.: Online learning activity index (OLAI) and its application for adaptive learning. In: Cheung, S.K.S., Kwok, L., Ma, W.W.K., Lee, L.-K., Yang, H. (eds.) ICBL 2017. LNCS, vol. 10309, pp. 213–224. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59360-9_19

Jia, J.: An AI framework to teach English as a foreign language: CSIEC. AI Mag. **30**(2), 59–71 (2009)

Jia, J.: Intelligent tutoring systems. In: Spector, M. (ed.) Encyclopedia of Educational Technology, pp. 411–413. Sage, Thousand Oaks (2015)

Jia, J., Zhang, J.: The analysis of online learning behavior of the students with poor academic performance in mathematics and individual help strategies. In: Cheung, S.K.S., Lee, L.-K., Simonova, I., Kozel, T., Kwok, L.-F. (eds.) ICBL 2019. LNCS, vol. 11546, pp. 205–215. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21562-0_17

Johnson, W.L., Lester, J.C.: Pedagogical agents: back to the future. AI Mag. **39**(2) (2018)

Kim, S., Chen, R.P., Zhang, K.: Anthropomorphized helpers undermine autonomy and enjoyment in computer games. J. Consum. Res. **43**(2), 282–302 (2016)

Kim, Y., Baylor, A.L.: Pedagogical agents as social models to influence learner attitudes. Educ. Technol. 23–28 (2007)

Le, H., Jia, J.: Analysis of learner timeout behavior in online tests of a bigdata set based on the OLAI concept. In: Cheung, S.K.S., Lam, J., Li, K.C., Au, O., Ma, W.W.K., Ho, W.S. (eds.) ICTE 2018. CCIS, vol. 843, pp. 285–294. Springer, Singapore (2018). https://doi.org/10.1007/978-981-13-0008-0_27

Lee, A., Kawahara, T.: Julius v4.5 (2019). https://doi.org/10.5281/zenodo.2530395

Lepper, M.R., Chabay, R.W.: Socializing the intelligent tutor: bringing empathy to computer tutors. In: Mandl, H., Lesgold, A. (eds.) Learning Issues for Intelligent Tutoring Systems. COGNITIVE SCIEN, pp. 242–257. Springer, New York (1988). https://doi.org/10.1007/978-1-4684-6350-7_10

Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., Bhogal, R.S.: The persona effect: affective impact of animated pedagogical agents. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 359–366 (1997)

Liew, T.W., Zin, N.A.M., Sahari, N.: Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment. Hum.-Centric Comput. Inf. Sci. **7**(1), 9 (2017)

Louwerse, M.M., Graesser, A.C., McNamara, D.S., Lu, S.: Embodied conversational agents as conversational partners. Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn. **23**(9), 1244–1255 (2009)

Lowenthal, P.R.: The evolution and influence of social presence theory on online learning. In: Kidd, T.T. (ed.) Online Education and Adult Learning: New Frontiers for Teaching Practices, pp. 124–134. IGI Global, Hershey (2009)

Makransky, G., Wismer, P., Mayer, R.E.: A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. J. Comput. Assist. Learn. **35**(3), 349–358 (2019)

Mayer, R.E., DaPra, C.S.: An embodiment effect in computer-based learning with animated pedagogical agents. J. Exp. Psychol. Appl. **18**(3), 239–252 (2012)

Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. J. Soc. Issues **56**(1), 81–103 (2000)

Northcutt, C.G., Ho, A.D., Chuang, I.L.: Detecting and preventing “multiple-account” cheating in massive open online courses. Comput. Educ. **100**, 71–80 (2016)

Pekrun, R.: The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educ. Psychol. Rev. **18**(4), 315–341 (2006)

Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)

Prendinger, H., Ishizuka, M.: The empathic companion: a character-based interface that addresses users’ affective states. Appl. Artif. Intell. **19**(3–4), 267–285 (2005)

Schilbach, L., et al.: Being with virtual others: neural correlates of social interaction. Neuropsychologia **44**(5), 718–730 (2006)

Schroeder, N.L., Adesope, O.O.: A systematic review of pedagogical agents’ persona, motivation, and cognitive load implications for learners. J. Res. Technol. Educ. **46**(3), 229–251 (2014)

Schroeder, N.L., Adesope, O.O., Gilbert, R.B.: How effective are pedagogical agents for learning? A meta-analytic review. J. Educ. Comput. Res. **49**(1), 1–39 (2013)

Short, J., Williams, E., Christie, B.: The Social Psychology of Telecommunication. Wiley, London (1976)

Terzidou, T., Tsiatsos, T., Miliou, C., Sourvinou, A.: Agent supported serious game environment. IEEE Trans. Learn. Technol. **9**(3), 217–230 (2016)

Turk, M.: Multimodal interaction: a review. Pattern Recogn. Lett. **36**, 189–195 (2014)

Wu, C.H., Huang, Y.M., Hwang, J.P.: Review of affective computing in education/learning: trends and challenges. Br. J. Edu. Technol. **47**(6), 1304–1323 (2016)

Yılmaz, R., Kılıç-Çakmak, E.: Educational interface agents as social models to influence learner achievement, attitude and retention of learning. Comput. Educ. **59**(2), 828–838 (2012)

Yung, H.I., Paas, F.: Effects of cueing by a pedagogical agent in an instructional animation: a cognitive load approach. Educ. Technol. Soc. **18**(3), 153–160 (2015)

Zhang, B., Jia, J.: Evaluating an intelligent tutoring system for personalized math teaching. In: Proceedings of International Symposium on Educational Technology 2017, pp. 126–130 (2017)