2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

From Textbook to Teacher: an Adaptive Intelligent Tutoring System Based on BCI

Tao Xu[1], Xu Wang[1], Jiabao Wang[1] and Yun Zhou[2]

Abstract— **In this work, we propose FT^3, an adaptive intelligent tutoring system based on Brain Computer Interface(BCI). It can automatically generate different difficulty levels of lecturing video with teachers from textbook adapting to student engagement measured by BCI. Most current studies employ animated images to create pedagogical agents in such adaptive learning environments. However, evidence suggests that human teacher video brings a better learning experience than animated images. We design a virtual teacher generation engine consisting of text-to-speech (TTS) and lip synthesis method, being able to generate high-quality adaptive lecturing clips of talking teachers with accurate lip sync merely based on a textbook and teacher's photo. We propose a BCI to measure engagement, serving as an indicator for adaptively generating appropriate lecturing videos. We conduct a preliminary study to build and evaluate FT^3. Results verify that FT3 can generate synced lecturing videos, and provide proper levels of learning content with an accuracy of 73.33%.**

I. INTRODUCTION

In 2020, facing the sudden attack of the Covid-19 pandemic, online learning are widely used due to its higher teaching flexibility, lower cost, and sharing of high-quality resources. Current online learning does not support personalized learning. Many studies have proved that adaptive learning environments that provide personalization of the instruction process could facilitate reaching their full pedagogical potentials for students. A satisfactory adaption representation is a perquisite for a successful personalization.

Most current studies [1] employ animated images to create pedagogical agents to instruct students in such adaptive learning environments. However, evidence shows that the presence of human teacher image or video in learning environments brings students a better learning experience than animated images [2]. The video of human teacher gains more trust from students, helps to create a learning atmosphere similar to a real classroom, enhances students' sense of social presence, stimulates students' learning motivation and learning investment, and effectively promotes learning performance. Although videos of human teacher improve learning experience, the workload of preparing amounts of teaching video clips as learning contents used for adaption may overwhelm teachers. AI technology provides a solution to liberate teachers from heavy work of making videos but concentrating on designing instruction. The text to sound technology converts text to real human voice smoothly and deep learning helps synthesize lip movement based on sound merely.

Engagement indicates directly whether students are involved in learning, thus several attempts have been made to investigate on monitoring and enhancing engagement levels, and creating adaption rules in adaptive learning environment. Since Electroencephalogram (EEG) reflects inner brain activities directly and accurately [3], EEG-based Brain-Computer Interface (BCI) provides a feasible solution to measure engagement [4], [5]. Szafir and Mutlu [4] proposed a BCI to monitor user engagement and design adaptive agents to improve engagement levels. These agents used gestures, volume cues to evoke attention. Hu et al. distinguished attention into three levels by using CFS feature selection approach and k-nearest-neighbor (KNN), obtained a high accuracy of classification.

In this paper, we propose an intelligent adaptive tutoring system called FT^3, short for From Textbook to Teacher. It generates lecturing video of virtual human teacher adaptively merely with lecturing textbook and a photo of teacher and employs an EEG-based BCI to detect student engagement as an indicator to adapt different difficulty levels of learning content. This system consists of an automated virtual teacher generation engine, an EEG-based BCI for detecting student engagement, and a learning content adaption engine. The automated virtual teacher engine is designed based on TTS and lip synthesis method, which can create lecturing video with accurate teacher's lip sync movement based on the adaptive lecture text and teacher photo so that the student watches personalized lecturing video clips. EEG-based BCI detects and monitors student engagement in real-time and sends the engagement value to adaption engine simultaneously. Learning content adaption engine allocates lecture text at appropriate difficulty levels to the automated virtual teacher generation engine. Finally, we conduct a preliminary study to build and evaluate FT^3 system. The learning contents selected from standardized tests have different difficulty levels, used for finding thresholds of engagement index for each participant in a pilot calibration experiment. Then we design an adaptive English teaching experiment to evaluate FT^3. Results prove the feasibility of FT^3 and show that it can provide proper levels of learning content with an accuracy of 73.33%.

*This research was supported by the National Natural Science Foundation of China(62077036), the National Key Research and Development Program of China(2018AAA0100500), the Shaanxi Natural Science Foundation(2019JQ-628)

[1] Tao Xu, Xu Wang and Jiabao Wang are with the School of Software, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, 710072, P.R.China. xutao@nwpu.edu.cn, 1149693659@mail.nwpu.edu.cn, wjiabao@mail.nwpu.edu.cn

[2] Yun Zhou is with the School of Education, Shaanxi Normal University, Xi'an, 710062, P.R.China. Corresponding authors, zhouyun@snnu.edu.cn

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

II. FT^3 SYSTEM STRUCTURE

As shown in Fig.1, FT^3 consists of three parts, including an automated virtual teacher generation engine, an EEG-based BCI for Engagement Assessment, and a learning content adaption engine.

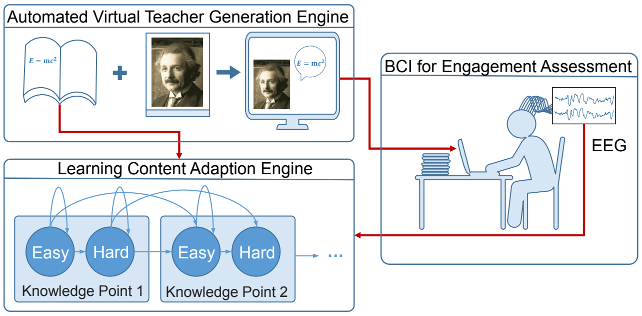

Automated Virtual Teacher Generation Engine
E = mc²
E = mc²
BCI for Engagement Assessment
EEG
Learning Content Adaption Engine
Easy
Hard
Knowledge Point 1
Easy
Hard
Knowledge Point 2
...

Fig. 1. The system structure of FT^3

III. AUTOMATED VIRTUAL TEACHER GENERATION ENGINE

Automated virtual teacher generation engine creates personalized lecturing video clips for students, and these video clips include a talking teacher. To build this engine, we propose an approach to generate a talking teacher based on the adaptive lecture text and an teacher photo, with two steps, including converting text to sound and syncing lip movement, as shown in Fig.2. The first step employs text-to-speech (TTS) technology to convert text to human voice that will be used as an input for the second step. The second step leverages a photo of a teacher, the generated voice, and lip synthesis technology to create lecturing video. In these videos, a talking teacher with accurate and realistic lip movements teaches the adaptive knowledge provided by learning content adaption engine.

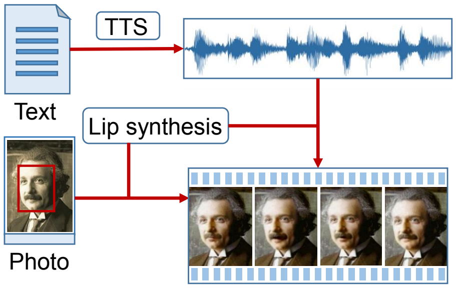


Fig. 2. The work flow of automated virtual teacher generation engine

A. *From Textbook to Speech*

As one of the most important technologies in Nature Language Processing (NLP) and human-computer interaction, TTS synthesizes human voice from digital text. Current TTS methods include two types: traditional method and deep learning based method. The traditional method consists of a text analysis, an acoustic model and a speech synthesizer. The deep learning based method is an end-to-end model, creating voices directly from text based on data with tens of thousands of samples per second of audio. Compared with traditional method, deep learning based method becomes the mainstream method of TTS, producing sound more naturally.

In FT^3 system, we employ Deep Voice 3 [6] to synthesize the sound. It can convert various text features (words, phonemes, accents) into various acoustic features, such as mel-spectrogram, linear scale logarithmic amplitude sound spectrum, or a set of vocoder characteristics like fundamental frequency, amplitude-frequency envelope, and non-periodic parameters. These acoustic features are then used as the input of the sound waveform synthesis model. First, a fully convolutional encoder converts text features into internal learned representations. Then, a fully convolutional causal decoder decodes the learned representations using a multi-hop convolutional attention mechanism (in an automatic regression mode) into low-dimensional sound representations (mel-spectrogram). Finally, a converter, a fully convolutional post-processing network, predicts the acoustic characteristics through the hidden state passed by the decoder, thereby synthesizing the signal waveform. Differing from the decoder, the converter is non-causal. Thus it is able to employ the future decoder context to predict the output.

B. *From Photo to Teacher*

Lip movement synthesis studies realizing human-like lip movements through mapping from acoustic speech signals to lip image sequence, crucial for generating digital natural human-like face [7]. Inspired by work in [8], we adopt Wav2Lip model to generate lip-sync video. Using Generative Adversarial Network (GAN), this model consists of generator, pre-trained lip-sync expert, and a visual quality discriminator. In the model, generator creates a lip-synced frame with sound and lecture text, including three parts: audio encoder, face encoder and face decoder. A pre-trained lip-sync expert is used to detect out-of-sync video-audio content in raw, unconstrained samples. Visual quality discriminator is used to force the generator to continuously produce accurate and realistic lip movements. This model underlies the development of photo-to-teacher part in our automated virtual teacher generation engine, through which we synthesize lecturing video with a teacher photo and previously generated voices.

IV. EEG-BASED BCI FOR ENGAGEMENT ASSESSMENT

Engagement reflects directly whether student pay attention to learning materials. EEG is an electrophysiological monitoring method for recording brain electrical activity, being considered as a very promising method of cognitive physiological evaluation due to its objectivity, real-time, and accuracy. In this paper, we propose to use EEG-based BCI to identify student engagement and integrate it into adaptive intelligent teaching system so that the system can improve learning experience and enhance learning effects through optimizing engagement.

As shown in Fig.3, three main parts underlie our EEG-based BCI, that is, EEG data preprocessing, engagement value calculating, and threshold generating.

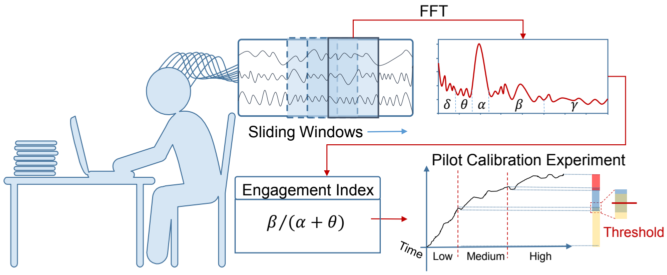


Fig. 3. EEG-based BCI for Engagement Assessment

First, we employ the sliding window to process the raw EEG data. The sliding window lasts one second and the stride is 0.5 second. The power spectrum analysis is used to obtain frequency features. The EEG data segments of each channel is transformed into the frequency domain by the Fast Fourier Transform (FFT). The power of α, β, δ, and θ is calculated as the accumulated power of the spectrum within their waves frequencies band.

Second, we adopt engagement index as the main indicator to make learning content adapt, which is widely used to reflect human learning situation, including learning and memory Tasks, and math problem solving [9]. This index was first proposed by Popes et.al [10] at NASA, which is proportional to the power of β wave and inversely proportional to the power of α wave and θ wave as below. The value of engagement index of segments can be calculated according to this equation (1) and the power of α, β, and θ getting from the power spectrum analysis.

$$Engagement\ index = \frac{\beta}{\alpha + \theta} \tag{1}$$

Third, we generate thresholds distinguishing the engagement level into low, medium and high, through a calibration experiment as described in details in Section 6. In this experiment, we select questions at different difficulty levels from standardized tests, and use these tests to trigger low, medium and high engagement levels. We record EEG signals during the answering process and calculate engagement values over time based on EEG data.

To generate the low limit threshold to distinguish low and medium, and high limit threshold to distinguish medium and high, we first smooth these values using exponentially weighted average, as shown in the following equation:

$$S_t = \begin{cases} Y_1, & t = 1 \\ \alpha Y_t + (1-\alpha) \cdot S_{t-1}, & t > 1 \end{cases} \tag{2}$$

where: the coefficient α represents the degree of weighting decrease, a constant smoothing factor between 0 and 1. Y_t is the value at a time period t. S_t is the value at any time period t.

We set α as 0.9 to obtain a smooth curve. We then map the value of this smooth curve to a one-dimensional space. As shown in Fig.3, we calculate a low limit by averaging the maximum and minimum of overlapping parts of low and medium values, and a high limit by averaging the maximum and minimum of overlapping parts of medium and high values. Due to the big differences in EEG data between individuals, the threshold varies as the participant changes.

V. Learning Content Adaption Engine

In our system, we recommend the student an optional adaptive learning path instead of an obligatory recommended learning content, and s/he determines to proceed in which way by her/himself. The recommended unit is learning content and each learning content is labelled with a level as low, medium or high. We assume that the engagement value increases when the difficulty levels of tasks increasing, and design and conduct a pilot calibration experiment to test this hypothesis.

With the purpose of matching the difficulty levels of learning content to the level of the student, we define three rules to support adaption for learning content adaption engine, including "Repeat", "Step in", and "Step over". Results from [11] show that the student should be provided with medium difficulty level of learning materials so that s/he can solve the problem and proceed. Learning materials at low difficulty level may result in boring and high difficulty level may result in frustrating. And medium level is an appropriate level. Therefore, the principle of adaption in FT^3 is to adjust engagement level from low to medium or from high to medium. "Repeat" refers to repeating current learning content. When the engagement level increases and stays high for a while, "Repeat" is activated and the learning content is presented again. "Step in" refers to moving onto the next learning content but at the same difficulty level. "Step in" is activated when engagement is at medium level. The "Step over" refers to jump to the learning content at the level more difficult when engagement index is low. These rules provide optional learning paths for students but students themselves determine how to proceed.

VI. Experiment Design and Evaluation

To build and evaluate the FT^3, we design two experiments, including a pilot calibration experiment and an adaptive English teaching experiment. The pilot calibration test is used for calculating engagement thresholds and testing the hypothesis in Section 5. The adaptive English teaching experiment is conducted for evaluating the proposed FT^3 System.

In this work, we adopt Emotive Epoc+ to collect EEG signals. Emotive Epoc+ is a wireless portable dry EEG device to collect EEG signals of 14 channels, including AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, based on the international 10-20 system. The sampling rate was 256Hz.

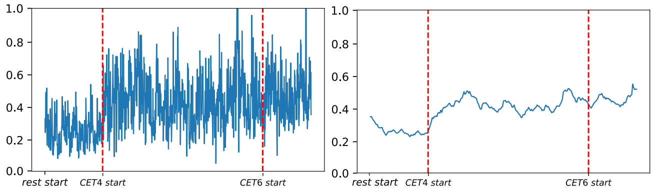
We create an adaptive English teaching system based on our methodology and employ text from the College English Test (CET) as learning content. CET is a Chinese national English test, aiming at estimating English levels of undergraduate and postgraduate. It has two levels: CET-4 in low and CET-6 in high.

A. Pilot Calibration Experiment

The purpose of the pilot calibration experiment is double-folded: to find the thresholds and to test the hypothesis in Section 5.

Five healthy graduate student (two females and three males) participated in the pilot calibration experiment. The procedure contains two steps. First, the tester showed four scenery images lasting 10 seconds respectively to participant for collecting EEG data of resting state, which corresponds to the low difficulty level of test and low engagement level . Then, seven listening comprehensive materials (four from CET-4 and three from CET-6) were chosen to trigger engaged states at different levels. After listening each material, two questions were asked to answer. CET-4 tests were used to trigger medium engagement level and CET-6 was related to high engagement level.

We visualized raw engagement values and smoothed these values of one participant. Through smoothing and mapping methods, we identified thresholds of three engagement levels (low, medium and high) for each participant. As shown in Fig.4, it was obvious that the engagement value increases when the difficulty levels of tasks increasing, which proved our hypothesis in Section 5.



(a) Raw data

(b) Smooth data

Fig. 4. The curve of engagement index of one participant

B. Adaptive English Teaching Experiment

In this experiment, we selected a photo of Albert Einstein to create virtual talking teacher and four pieces of lecture text for adaption. Two pieces were from the CET-4 representing medium-level learning content, while the other two from CET-6 representing hard-level content.

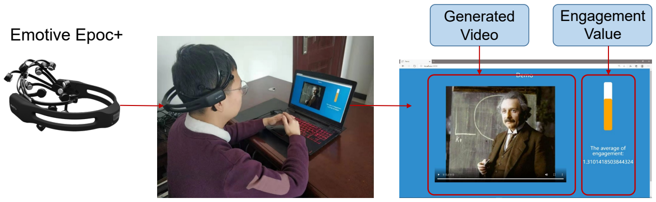


Fig. 5. The adaptive English teaching experiment

The interface of FT^3 system based on this adaptive English teaching experiment is shown in Fig.5, with the virtual talking teacher on the left side, and the real-time engagement index value on the right. The front-end used the Node.js based web development framework Express, supporting a remote access by using Socket.io in Node.js. Five same participants evaluated FT^3 system based on this adaptive English teaching experiment. We used a self-report to express her/his engagement levels for each learning content. In total, 15 learning video clips have been watched and learned. Results showed that the system could identify the engagement levels of 11 clips, with an accuracy rate of 73.33%.

VII. Conclusion and Future Work

In this work, we proposed and proved an intelligent adaptive tutoring system called FT^3 employing an EEG-based BCI to detect student engagement and using merely lecture text and photo to generate virtual human teacher for adaption representation. FT^3 system shed a light for using human teacher videos in cognitive state adaptive learning system. In the future, we will investigate profoundly how to present lecturing videos with human teacher with what frequency and what learning situation can best promote learning, and discuss the implications to inform the design of cognitive state sensing intelligent adaptive tutoring system.

References

[1] S. W. Chae, K. C. Lee, and Y. W. Seo, “Exploring the Effect of Avatar Trust on Learners’ Perceived Participation Intentions in an e-Learning Environment,” *International Journal of Human–Computer Interaction*, vol. 32, no. 5, pp. 373–393, May 2016, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2016.1150643.

[2] Q. Dunsworth and R. K. Atkinson, “Fostering multimedia learning of science: Exploring the role of an animated agent’s image,” *Computers & Education*, vol. 49, no. 3, pp. 677–690, Nov. 2007.

[3] T. Xu, Y. Zhou, Y. Wang, Z. Zhao, and S. Li, “Guess or Not? A Brain-Computer Interface Using EEG Signals for Revealing the Secret behind Scores,” in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–6.

[4] D. Szafir and B. Mutlu, “Pay attention! designing adaptive agents that monitor and improve user engagement,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’12. New York, NY, USA: Association for Computing Machinery, May 2012, pp. 11–20.

[5] J. Huang, C. Yu, Y. Wang, Y. Zhao, S. Liu, C. Mo, J. Liu, L. Zhang, and Y. Shi, “FOCUS: enhancing children’s engagement in reading by using contextual BCI training sessions,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’14. New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 1905–1908.

[6] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning,” Feb. 2018.

[7] E. Yamamoto, S. Nakamura, and K. Shikano, “Lip movement synthesis from speech based on Hidden Markov Models,” *Speech Communication*, vol. 26, no. 1, pp. 105–115, Oct. 1998.

[8] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild,” in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM ’20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 484–492.

[9] F. Cirett Galán and C. R. Beal, “EEG Estimates of Engagement and Cognitive Workload Predict Math Problem Solving Outcomes,” in *User Modeling, Adaptation, and Personalization*, ser. Lecture Notes in Computer Science, J. Masthoff, B. Mobasher, M. C. Desmarais, and R. Nkambou, Eds. Berlin, Heidelberg: Springer, 2012, pp. 51–62.

[10] A. T. Pope, E. H. Bogart, and D. S. Bartolome, “Biocybernetic system evaluates indices of operator engagement in automated task,” *Biological Psychology*, vol. 40, no. 1, pp. 187–195, May 1995.

[11] S. D’Mello and A. Graesser, “Dynamics of affective states during complex learning,” *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, Apr. 2012.