IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 19, 2017, accepted December 6, 2017, date of publication January 10, 2018, date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2791861

Towards Emotionally Aware AI Smart Classroom: Current Issues and Directions for Engineering and Education

YELIN KIM[1], (Member, IEEE), TOLGA SOYATA[1], (Senior Member, IEEE), AND REZA FEYZI BEHNAGH[2]

[1]Department of Electrical and Computer Engineering, State University of New York at Albany, Albany, NY 12222, USA
[2]Department of Educational Theory and Practice, State University of New York at Albany, Albany, NY 12222, USA

Corresponding author: Tolga Soyata (tsoyata@albany.edu)

ABSTRACT Future smart classrooms that we envision will significantly enhance learning experience and seamless communication among students and teachers using real-time sensing and machine intelligence. Existing developments in engineering have brought the state-of-the-art to an inflection point, where they can be utilized as components of a smart classroom. In this paper, we propose a smart classroom system that consists of these components. Our proposed system is capable of making real-time suggestions to an in-class presenter to improve the quality and memorability of their presentation by allowing the presenter to make real-time adjustments/corrections to their non-verbal behavior, such as hand gestures, facial expressions, and body language. We base our suggested system components on existing research in affect sensing, deep learning-based emotion recognition, and real-time mobile-cloud computing. We provide a comprehensive study of these technologies and determine the computational requirements of a system that incorporates these technologies. Based on these requirements, we provide a feasibility study of the system. Although the state-of-the-art research in most of the components we propose in our system are advanced enough to realize the system, the main challenge lies in: 1) the integration of these technologies into a holistic system design; 2) their algorithmic adaptation to allow real-time execution; and 3) quantification of valid educational variables for use in algorithms. In this paper, we discuss current issues and provide future directions in engineering and education disciplines to deploy the proposed system.

INDEX TERMS Educational technology, emotion recognition, smart classroom, deep learning, real-time computing, mobile-cloud computing, meta-cognition.

I. INTRODUCTION

Imagine a smart classroom in the year of 2024, in which a student or a prospective teacher is wearing haptic gloves and practicing a presentation in front of their peers or other students. Although this setup is similar to today's training/practice environment, they will have had the advantage of performing "*live*" presentations in which a machine intelligence-driven system provides instant feedback to them through their haptic gloves and a feedback visualization dashboard in *real-time* on how to adapt their body language, voice intonation, and other non-verbal behavior to be a more effective and emotionally-intelligent teacher and presenter. What allows this instant feedback (in "*presentation mode*") is a system that accomplishes highly sophisticated tasks *within a fraction of a second*, by digitizing the multi-modal audio/visual data of the presenters (e.g., through cameras and microphones) and transmitting them into the cloud through a high-speed network, processing them to determine the behavioral state using sophisticated computationally-intensive algorithms on Graphics Processing Unit (GPU) clusters, and a deep-learning based network that had been previously trained; such training is assumed to have been done on a similar presentation, in which the machine (in "*training mode*") receives feedback from the crowd via a mobile app that is installed on the listeners' mobile devices.

The scenario described in the previous paragraph is no longer science fiction and presents our vision in this paper; we present current issues and research directions that utilize the state-of-the-art research in behavior recognition, deep

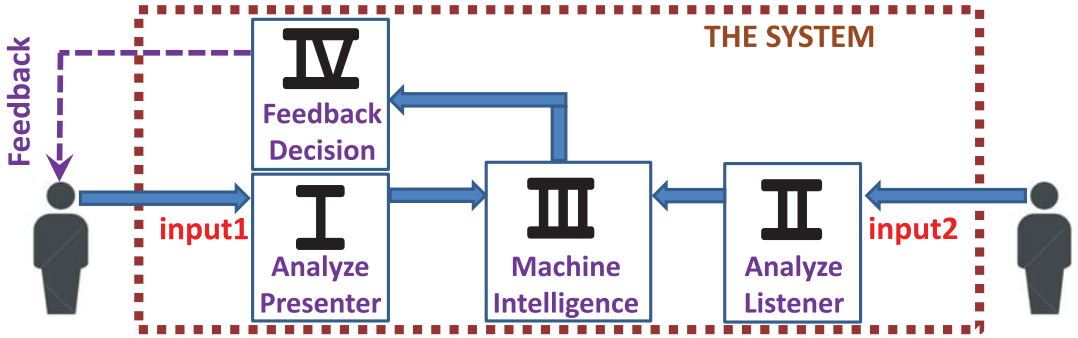

FIGURE 1. Overview of a conceptualized smart classroom system, which receives its input from humans (input1 and input2) and outputs its computed feedback response to one of the humans. While the internal computations of the system are based purely on quantitative values, the connection between the "human" and the "machine" require complete quantification of the human behavior, which is not trivial.

learning, computational architecture, and educational theory to conceptualize a "system" that we hypothesize to be the a part of future smart classrooms [1]. A successful implementation of such a system can have a transformative impact not only on classroom-based education, but also the way we train our sales personnel, doctors, security personnel, and soldiers who are deployed in a foreign country, more generally any setting where humans communicate with each other to exchange knowledge or information. Indeed, human communication does not only depend on verbal communication; a significant percentage of it involves non-verbal communication, such as voice intonation and body gestures [2]–[4]. While these aspects are under-studied in our current education system, their underlying theory can be quantified surprisingly well and can be incorporated into machine-intelligence-based education. Therefore, we further argue that a machine intelligence-driven system can allow students to receive valuable feedback during their practice presentation in front of the "machine," while avoiding presentation anxiety [5] or embarrassment due to an imperfect presentation.

The described application requires a totally new system design paradigm, depicted in Fig. 1, which receives its inputs from two human sources —which are potentially subjective— and outputs its feedback to one of them. Such a design introduces four significant research challenges: **(a)** while both of the inputs into the system are received from humans, a purely-quantitative basis must be established for both of them to allow them to be used in machine learning algorithms that only work with well-defined quantitative data. Therefore, a "box" must exist in the system that turns potentially subjective presenter input (Box I), as well as the listener input (Box II) into pure quantitative values; **(b)** a machine intelligence platform must be designed (Box III) that is capable of "learning" the relationship between these quantified inputs; while existing research investigates Multimodal Learning Analytics (MLA) [6]–[9], in which the acquired multi-modal presentation data is applied into a data analytics algorithm in a "*feed-forward*" fashion to determine the effectiveness of the presenters, the additional "*feed-back*" loop in Fig. 1 drastically increases the underlying system complexity; **(c)** another machine learning platform (Box IV) must be designed to learn each presenter's available cognitive resources to tailor the timings and modalities of the feedback to individual presenters, thereby *personalizing* the machine-generated feedback; **(d)** finally, the entire system (Box I–IV) must be capable of operating in *real-time* [10], to allow the presenter to make changes *during* the presentation. Existing algorithms that can be used for Boxes I–IV work two to three orders-of-magnitude slower than real-time; therefore, a new set of algorithms must be developed to achieve the required computational acceleration.

The main contributions of this paper include:

- integration of multimodal sensing, emotion recognition, deep learning, high-performance GPU computing, and feedback systems,
- quantification of important human variables in a smart classroom, such as crowd scores and behavioral cues,
- demonstration and verification of our proposed system design with a template smart classroom at SUNY Albany,
- bridging the gap between engineering and education with an education theory-based system design, and
- review of the state-of-the-art technologies and proposal of future directions for AI-enabled smart classrooms.

The remainder of this paper is organized as follows. In Section II, we introduce our proposed system design in detail. In the following three sections, we survey the state-of-the-art in technologies that are necessary to realize this system: Section III studies the techniques that allow the quantification of potentially-subjective human-based metrics such as crowd scores. Section IV investigates the techniques that can quantify non-verbal human communication metrics, such as human emotions, facial expressions, and general body language, as well as voice-related metrics. Section V elaborates on techniques that enable the required high-intensity computations in real-time. In Section VI, we develop a detailed algorithmic/computational infrastructure for our proposed system and provide a feasibility evaluation in Section VII. We point out the open challenges and outline directions

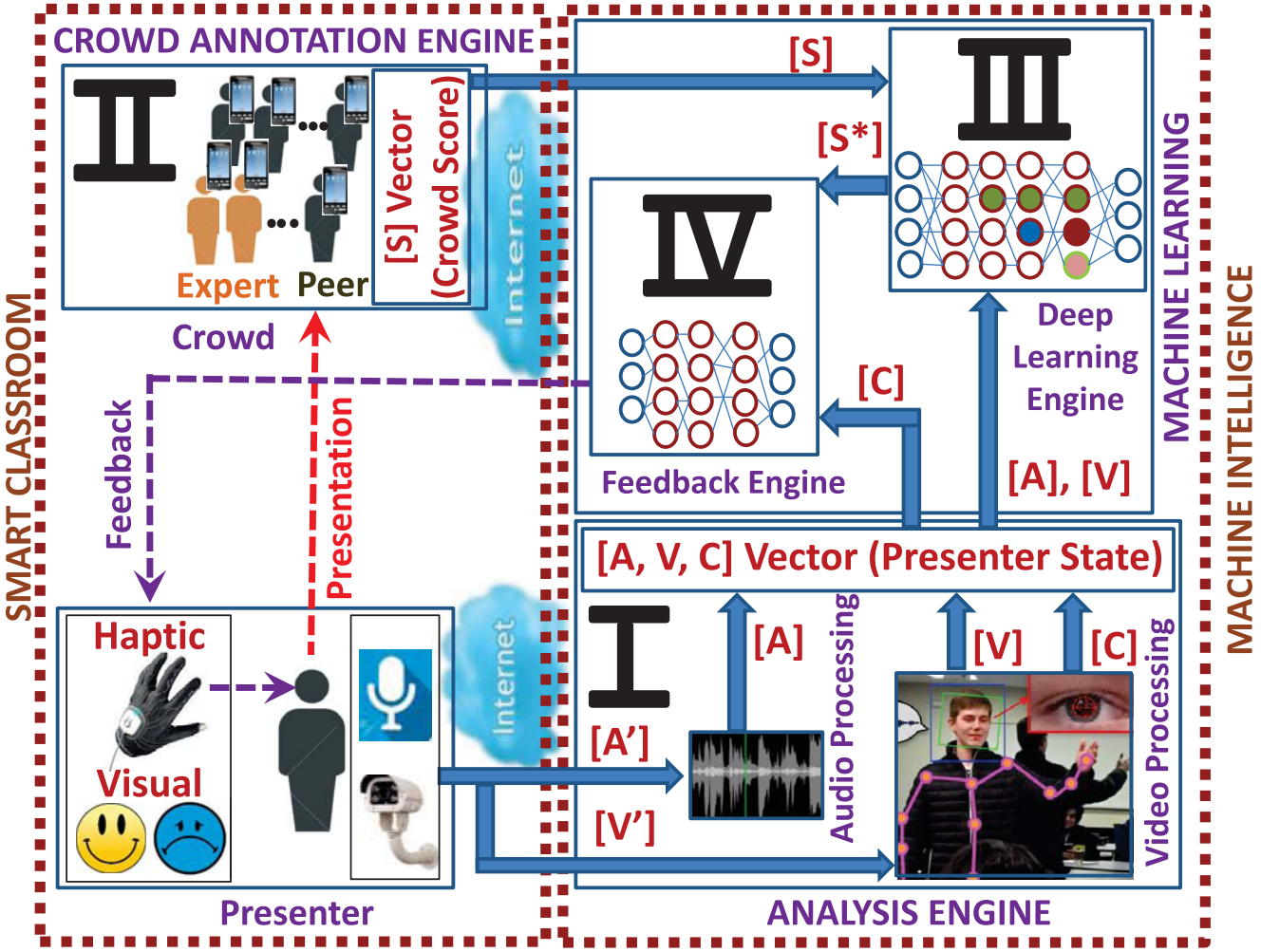

FIGURE 2. Proposed smart classroom system, which consists of (Box I) code to extract Audio ([*A*]), Visual ([*V*]), cognitive load ([*C*]) vectors from a presentation, (Box II) smart classroom crowd, composed of peers and experts, for computing crowd scores ([*S*]) (Box III) deep learning algorithms that learn "best practices" in *Training Mode* and estimate crowd scores ([*S*∗]) in *Presentation Mode*, and (Box IV) a machine learning engine that provides real-time haptic or visual feedback to the presenter.

for future research in Section VIII. We provide concluding remarks in Section IX.

II. SYSTEM ARCHITECTURE

Our proposed work to realize the system in Fig. 1 relies on two hypotheses we propose:

H1: It is possible to quantify the presenter and listener input (Boxes I, II), despite their subjective nature, and

H2: It is possible for machine intelligence (Box III) to learn the relationship between the presenter behavior and its resulting effect on the presentation quality in *Training Mode* and convey this information to the presenter (Box IV) without distracting them in real-time in *Presentation Mode*.

A. PROPOSED SYSTEM

To test both of our hypotheses, we propose a system design shown in Fig. 2, in which a presenter's *raw* audio ([A']) and video ([V']) data are captured during a presentation and transmitted to the cloud. A set of multi-modal sensing and behavior recognition algorithms in the *Analysis Engine* (Box I) convert this raw information into processed audio ([A]) and visual ([V]) feature vectors to quantify the behavioral cues of the presenter such as *vocal expressions*, *facial movements*, and *body gestures*. The same video stream is also used to compute the presenter's cognitive load ([C] vector) using techniques that correlate pupil dilations or other facial expressions to the cognitive load [11], [12]. The *Crowd Annotation Engine* (Box II) turns the votes of the crowd (who are either *expert* or *peer* listeners) into robust and practical quantitative metrics (such as the proposed Crowd Score Vector [S]). The *Deep Learning Engine* (Box III) learns what type of behavioral patterns such as open hand gestures —quantified by the [A] and [V] vectors— result in the highest crowd scores (quantified by the [S] vector). The *Feedback Engine* (Box IV) makes real-time suggestions to the presenter during

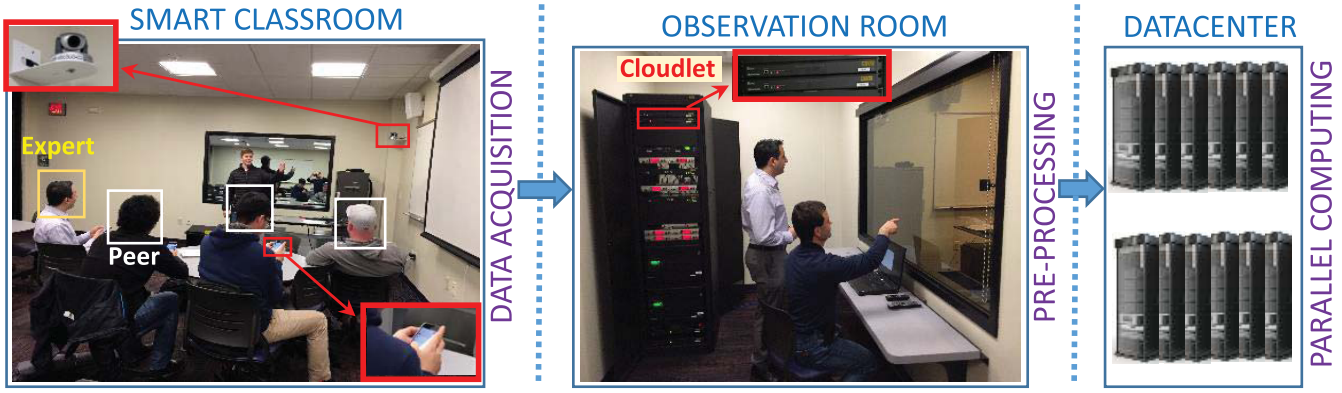

FIGURE 3. Our proposed smart classroom using a three-component system architecture: (i) *data acquisition* component to receive raw audio/video information, (ii) *pre-processing* component to perform initial computations on the received data, and (iii) *massively-parallel computing* component to perform the high-intensity computations, required by the proposed algorithms.

a presentation, using either a haptic glove and visual feedback by a monitor screen in front of the presenter, by taking into account the cognitive load of the presenter (quantified by the $[C]$ vector). These suggestions allow the presenter to adjust his/her body language, voice intonation, or hand gestures to improve the presentation.

B. DESIGN CHALLENGES

To realize the system in Fig. 2, the following design challenges must be overcome: First, a methodology is required to quantify multimodal cues from the presenter ($[A]$, $[V]$, $[C]$ vectors in Box I) and the potentially subjective input from the listeners (the $[S]$ Vector in Box II). Particularly, the computation of $[S]$ must be investigated based on psychometric studies in education. Established methods in *crowd-sensing* to eliminate outliers [13] and potentially incorrect or biased scores [14], [15] can aid in determining an accurate $[S]$ vector.

Second, the design of a deep neural network (Box III) is required, which can learn the complex non-linear relationship among $[A]$, $[V]$, and $[S]$ vectors in *Training Mode* during a presentation, and emulate a crowd by generating an estimated crowd score vector ($[S^*]$) in *Presentation Mode*.

Third, the design of a parametric feedback engine (Box IV) requires the understanding of the best visual and haptic feedback options, which must be based on multiple parameters that allow the system to adapt to the skills and moment-to-moment available cognitive resources of a diverse set of presenters. Because this feedback is highly subjective, it must be personalized by using the cognitive load (the $[C]$ vector), to determine the modality of the provided feedback. This will create an atmosphere where the presenters can adjust the machine to their own pace and *personalize* its usage; this is a step towards *personalized education*, which avoids overly-generalized teaching methods that might cause some of the slower learners to be left behind.

Finally, a set of parallel programming techniques that can be used to allow Boxes I, III, and IV to run in real-time must be developed. As we will detail in Section VII-B, this requires two to three orders-of-magnitude computational acceleration as compared to single server implementations. Although mere functionality that is required for Box I exists using programs such as openSMILE speech feature extraction [16] and the vision tools DRMF [17] and CERT [18], and for Boxes III and IV using the deep learning tools Caffe [19] and Nvidia DIGITS [20], achieving real-time performance requires the re-formulation of their underlying computations by decomposing them into online/offline components [21] as well as applying multi-cloud-server parallelization techniques [22], [23].

C. A TEMPLATE SMART CLASSROOM

A smart classroom that we conceptualize is shown in Figure 3, which is housed at SUNY Albany. Each classroom is currently equipped with several cameras on the wall and wireless microphones at the podium, as well as an observation room. Each observation room includes a rack with necessary hardware to receive and record audio and video data, as well as a one-way mirror behind which researchers and educators can observe the teacher, presenter, or the classroom non-intrusively. This audio/video information can be used to analyze affective, cognitive, verbal, and non-verbal cues of the presenter. The classroom in Fig. 3 consists of three components: **(i)** a *data acquisition* component inside the classroom to receive raw audio ($[A']$) and raw video ($[V']$) data and provide feedback to the presenter through a haptic glove or a second screen that relays visual cues, **(ii)** a *pre-processing* component inside the observation room to perform the necessary initial computations, and **(iii)** a *massively-parallel computing* component at the data center to implement all high-intensity computations.

D. TRAINING MODE AND PRESENTATION MODE

Our system is expected to operate in one of the following two modes:

1) TRAINING MODE

In Training Mode, the presenter's behavioral cues and listeners' ratings are recorded in order to train the machine intelligence system. In this mode, the system is run while the *crowd*, which consists of a set of *peers* and a few *experts*, listens and scores the presentation, as shown in the data acquisition part of Figure 3. The crowd evaluates the presentation along with several other factors (e.g., presenter's body language and affect) based on the educational theories (detailed in Section VI-C) using a mobile application on their mobile devices.

2) PRESENTATION MODE

In this mode, presenters present, while being video and audio recorded. They receive real-time feedback regarding their body language, vocal tone, and affect (Section VII-B) while presenting through a haptic glove and visually through a screen on the podium (Section VI-C). The visual feedback will consist of easily understandable diagrams, bars using salient colors and shapes. In this mode, the machine essentially emulates Box II (expert and peer ratings, Section VI-A) based on all the input received in the *training mode*.

E. ENVISIONED DEPLOYMENT/TEST ENVIRONMENT

While our system is able to give *offline feedback* (i.e., at the completion of a presentation) to the presenter after the development of Box III, an *online feedback* (i.e., *real-time, during the presentation*) is only available after the development of Box IV. The development of Boxes III and IV are tightly intertwined; their design is an iterative process, gradually improving each one based on the results of several iterations of presentation studies. This test involves not only a set of recruited study participant students, who study to be a teacher, but also instructors as 'experts'.

As detailed in Section II-D, in the *Training Mode* all crowd participants hold smart phones with the Crowd scoring app installed and connected to the central computer. The design of the *Crowd Annotation Engine* (Box II) requires that quantitative metrics must be produced that quantify the presenter's voice, affect, and body language by first collecting raw human ratings —from experts and peers— through surveys administered via the smartphone app in real-time, which will then be quantified into metrics (such as the proposed Crowd Score Vector [S]) that can be integrated into the machine learning algorithm. These ratings are representative and summative quantitative measures of the presenter's vocal, affective, and body language (gestures, eye-contact), along with how they deliver the content (ease of comprehension, flow). Development of such a scoring mechanism for this engine is challenging due to the potential subjectivity of the crowd.

On the other hand, in the *Presentation Mode*, there is no crowd in the smart classroom and the presenter wears a haptic glove and watches a computer screen on the podium. Real-time machine feedback regarding his/her presentation is relayed to him/her at appropriate times through the haptic glove and visually on the computer screen.

III. QUANTIFICATION OF CROWD SCORING AND COGNITIVE LOAD

The proposed system design is driven by the two hypotheses we proposed in Section II, both of which state that the input from the ''human'' can be turned into purely quantitative values (Box I and II) and a feedback can be given to the human in real time (Box IV), which is computed from the quantitative output of the ''machine'' (Box III) and adjusted to be suitable for a human's cognitive absorption. In this section, we study the different aspects of the aforementioned human-machine interface, where the subjectivity of the human input (and output) poses significant challenges in the design of a machine intelligence-based system. We describe a mechanism for receiving a simplified quantitative score from the crowd —in real-time— while the presenter is presenting. Noting that this score may be biased due to the subjectivity of the members of the crowd, we describe methodologies to deal with the crowd bias.

In Section III-A, we describe a methodology to receive a set of individual scores from the listeners and statistical schemes to turn them into an aggregated *crowd score* (the [S] vector that is explained in Fig. 2) that captures the overall crowd rating, which is immune to individual biases and manipulations. In Section III-B, we elaborate on multiple options to provide real-time —yet cognitively-digestible— feedback to the presenter. In Section III-C, we describe a scheme to measure the cognitive load of the presenter to adapt the feedback (modality, amount, and timing) to the individual presenter.

A. CROWD SCORING

1) PRESENTATION FEEDBACK IN EDUCATION

Fostering 21st century student outcomes [24] such as *communication* is in the forefront of educational goals today. Communication is defined as the ability to articulate thoughts and ideas effectively, both orally and nonverbally. Traditionally, teachers support this by having students present and receive feedback from the teacher (as an expert) and peers. The feedback is typically provided to the student verbally at the end of the presentation. However, this type of feedback is not always systematic, could be cognitively overloading, and does not address all the key components (i.e., body language, affect, voice, intonation) of an effective and engaging presentation [25]. Additionally, there could be discrepancies among feedback from different classmates, or between peers and the teacher for various reasons [26]–[28]. Our conceptualization of the ''Crowd Score'' relies on the fact that the impact of these discrepancies among individuals will average out when a large crowd is used for scoring. Thus, the goal of the system is to make the presenter adjust to the *crowd behavior*, rather than pay unnecessary attention to *individual behavior*, which can be biased; the machine can learn to under-weight —or

outright eliminate— scores that it deems to be statistically less-important or biased.

2) QUANTIFYING THE CROWD SCORES

We aim to quantify the potentially subjective feedback from experts and peers into an aggregate value (the S vector, in Box II) relying on foundational educational theories on feedback and effectiveness of verbal and nonverbal communication [29]. To achieve this goal, we build on research that focuses on developing effective feedback mechanisms. Research in this area has relied on the principle that feedback leads to reflection on actions and improvements in performance [29] and allows humans to ''research their performance,'' (i.e., monitor and self-regulate [30], [31], and modify elements of their performance to achieve better outcomes). The cycle of self-regulation in humans entails forethought (planning and goal setting), performance and strategy use, and reflection (metacognition, evaluation and adaptation) [32], [33]. Feedback received by the student at any stage leads to enhanced planning, metacognition, monitoring of performance, and use of appropriate strategies to remedy problems by modifications to actions on-the-fly [34]–[36].

3) ELIMINATING LABEL NOISE

Human annotations are inherently subjective, due to difference in power relationship with the presenter (teacher vs. peer), prior experience of the scorer (novice vs. expert), or sheer intra-individual inconsistency. In general, the noise that is introduced to the crowd score by this subjectivity is termed *label noise* and can be categorized into two distinct types: (i) subjectivity in crowd scores and (ii) intentional misinformation. While the former former is caused by an opinion of an individual that might be skewed from the crowd, the latter is due to mis-intention, i.e., attempting to distort the outcomes to avoid the success of others. Both issues must be accounted for to design a successful system that is somehow immune to noise introduced by any of the two sources above.

One remedy to for eliminating bias can be to use well-established statistical methods that can improve the inter-evaluator agreement, such as Fleiss' Kappa statistics and algorithms that can eliminate outliers [37], intentional bad scorers, and untrustworthy scores. Additionally, quantitative reliability metrics should be established to account for peer vs. expert scoring differences [26]–[28]. Other measures to alleviate or minimize crowd bias include eliminating bias at the level of sampling, by recruiting a diverse (i.e., genders, races, other demographics) and random sample from the population, to ensure representativeness of the larger population, and to increase confidence in generalizing the findings of experiments to similar populations [38].

4) ENSURING TRUSTWORTHINESS

The field of Mobile Crowd Sensing (MCS) has studied the Sensing-as-a-Service (S^2aaS) platform [13]–[15], [39], in which a crowd of volunteering smartphone users perform a sensing task. Because the crowd may consist of users with malicious intent, recent studies have formulated methodologies to eliminate biases in the sensed input and provide values that are as close to correct values as possible. Although this platform was originally designed to be used in *Smart City* applications [40], [41] as well as Medical Cyber Physical Systems for remote health monitoring [42]–[44], it can be suitable for the proposed smart classroom, where the *user reputation* and *data trustworthiness* concepts in the MCS paradigm identify malicious and subjective user behavior, which are readily applicable to the design of Box II to eliminate user biases during scoring.

B. FEEDBACK TO THE PRESENTER

In the context of classroom presentations, tools such as wireless earpieces, haptic bracelets, and monitors have been used to provide immediate automated or human-generated feedback on body language and voice pitch to students presenting and in other contexts such as mentoring novice counselors by experts [45]–[50].

1) OPEN LEARNER MODELS

Visual feedback dashboards or Open Learner Models (OLMs) have been used extensively in the context of computer-based learning environments to provide feedback to users [34], [51]–[53]. An example of such a dashboard is Fig. 4, where students can review the degree of over- and under-confidence of their self-assessments visually during learning by looking at the color-coded progress bars. Researchers have investigated the benefits of allowing users to view these dashboards [53]–[55], indicating that the mere displaying of feedback raises the awareness of the users, allowing them to reflect on aspects of their performance and areas where they need to spend time to master [56]–[60]. To our knowledge, very few studies have examined how these interfaces can be implemented in classrooms [61].

2) FEEDBACK MODALITY

Given the benefits of providing students real-time feedback, the challenge is to create an effective interface to present this feedback in order to support the user. Many forms of representation have been adopted and modified to present feedback.

In classrooms, tools such as wireless earpieces, haptic bracelets, and monitors have been used to provide immediate feedback on body language and voice pitch to students presenting, one-on-one interviews, and in other contexts such as mentoring novice counselors by experts [45]–[50]. For instance, Damian *et al.* [47] presented a real-time hand and speech feedback system in a public speaking setting using a head-mounted device. Skill meters (progress bars, pie charts) are the most common visualization tool used in feedback dashboards [62]. Magic wands, smiley faces, and icons have been used to represent feedback as well [63].

Automatic behavioral feedback loops during multi-person, face-to-face social interactions, are also developed with an

FIGURE 4. An example dashboard deployed in an intelligent tutoring system providing visual feedback (bar graphs, numbers, colors) to students on over- and underconfidence of the self-assessments they make while learning about the human circulatory system [34].

aim to help the users to perceive their speaking time for balanced group discussions [64]. They use wearable devices, such as Myo armbands, headphones, and Google Glass, to capture multimodal behaviors of multiple users and deliver tactile (Myo armband), auditory (headphones), visual head-mounted (Google Glass), and visual remote (common monitor) stimulation. The study found that users are skeptical towards the usefulness of such systems, which opens new research questions to our proposed project. The authors also found the Google Glass and vibro-tactile feedback delivery devices to be the most disturbing.

Researchers [53]–[55] have investigated the benefits of allowing users to view these dashboards, indicating that the mere displaying of feedback raises the awareness of the users, allowing them to reflect on aspects of their performance and areas where they need to spend time to master [56]–[60]. However, the feedback is not always inspected carefully by students [58]. Tanimoto [65] argues that if the feedback presented is complex and difficult to understand, the learner may refuse to use the information presented.

Examining students' preference for modality of feedback presentation, Bull [53] found that students would like feedback to be about their weak points, current performance level, misconceptions, and to help them reflect on their performance. In another study, Law *et al.* [66] investigated student preference among eleven types of visualizations.

Thus far in the literature no studies have examined integrating human ratings into the feedback provided to the presenter. Additionally, Our work differs in that we focus on an education environment, where the presenter has to deliver specific content to the audience in a real classroom setting.

3) FEEDBACK PERSONALIZATION

Existing literature also indicates that humans have a preference for modes of visual presentation of information that are salient, easy to comprehend, and most relevant [52]. Fung *et al.* [67] also presented a novel automated framework that can analyze smiles, movement, and volume modulation of a person in a public speaking environment, using a webcam camera and web browser. This framework also connected Mechanical Turk workers to provide interpretations, ratings, and comment rankings to the presenter.

C. COGNITIVE LOAD

According to Cognitive Load Theory [68], [69], the human brain and information-processing system has limitations in capacity and duration when dealing with novel information; only a limited amount of cognitive processing can be carried out in the working memory at any moment in time [70]. These limitations are key factors influencing student learning and performance. In conditions of high cognitive load on the working memory, minimal resources become available for meaningfully interpreting information and this creates inadequate conditions for learning.

1) TYPES OF COGNITIVE LOAD

There are three types of cognitive load: *intrinsic load*, referring to the inherent difficulty associated with any particular task, which cannot be altered with instruction, *extraneous load*, which is induced by the manner in which the information is presented (i.e., design), and *germane load*, which is dedicated to processing and comprehending new information, thus enhancing learning. Extraneous and intrinsic cognitive loads are not ideal, because they are due to poor instructional designs and complexity of information. Ultimately, the goal of any instruction is to reduce extraneous cognitive load to free up working memory [71].

In instructional environments, external guidance or feedback is provided to enhance student performance. However, if a student has a high level of cognitive load during the task, there will be little if any available cognitive resources for him/her to interpret and apply the feedback [72]. Hence, designing effective feedback presentation mechanisms can greatly enhance interpretation of the feedback by the humans who receive it. This form of personalization of instruction plays a major role in the any education environment [73], which is superior to traditional classroom instruction in that

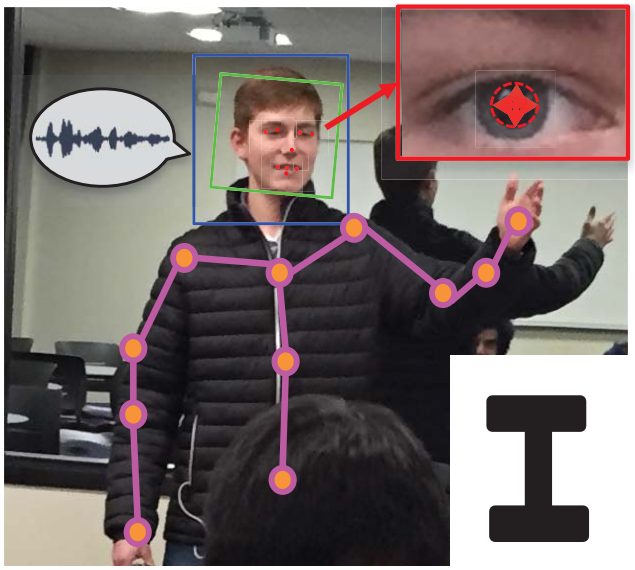

FIGURE 5. Demonstration of Box I in a SUNY Albany classroom. We capture nonverbal cues, such as facial expressions, body movements, speech prosody, and eye tracking system captures pupil dilation during a presentation.

the needs and capabilities of individual students are evaluated and appropriate levels of practice, task difficulty, and support are provided. In this regard, researchers have been developing intelligent tutoring systems [62], [74], [75] for the past two decades for individualizing instruction, strategy suggestions, and feedback to individual students.

2) COGNITIVE LOAD MEASUREMENT

Different methods have been proposed in the literature for measuring cognitive load, such as various rating scales [76] and data from pupillary response [77]. Researchers have used infrared cameras in eye-trackers or high resolution video cameras (e.g., the 5 MP camera used to generate Fig. 5) to record pupil dilation as a proxy for cognitive load. Variations in pupil dilation indicate working memory load, however, such physiological measures do not distinguish between different *types* of cognitive load. Surveys have been proposed as good alternatives to distinguish between germane, intrinsic, and extraneous cognitive loads [78]. Because pupils dilate while cognitive processing and brain activation increases [77], [79]–[83], it can be used to compute the cognitive load ($[C]$ vector). Additionally, fairly well established techniques can be used that correlate pupil dilations and other facial expressions to the cognitive load [11], [12]. Pupil dilation is associated with hedonic valence and emotional arousal as well [84]–[86], where higher arousal or intense emotions (positive or negative) lead to larger mean pupil diameter than neutral emotional state [86]. However, since pupil dilation is highly influenced by variations in the intensity of the light source in the environment or the brightness of the screen or the stimulus, these factors need to be controlled in experiments using this methodology [87]. In addition to increased pupil dilation, higher cognitive load leads to longer eye fixations and shorter saccades (rapid movement of eye from one point to the other) [88].

Note that the pupillary response is sensitive to both cognitive load and emotional activation. Care must be taken to distinguish between these two types of activation (i.e., cognitive vs. affective). An oral presentation setting is an inherently stressful task, which will cause high "affective activation." On the other hand, a high cognitive demand content, delivered during the presentation will result in a high cognitive load. Pupil dilations will indicate both of these components, although separating these two sources is challenging. One potential solution is to use machine intelligence algorithms for source separation with controlled experiment settings, which establish a baseline for an individual presenter.

A second factor that must be carefully handled when measuring pupil dilations is the pupillary response to luminescence. Because pupil dilations are also modulated by the variance in luminescence in the classroom, this can be a source of false measurements. Although *luminescence* can also be treated as a parameter in the machine intelligence algorithms, one simpler solution is to continuously compute a baseline luminescence during the course of a lecture and associate it with a baseline pupil dilation.

IV. COMPUTATIONAL BEHAVIOR ANALYSIS AND AUTOMATIC EMOTION RECOGNITION

In this section, we study algorithmic and statistical approaches to computationally represent and analyze human behavior. These approaches can be used to develop a *Multimodal Sensing and Analysis Engine* (Box I) that turns the recorded *raw* audio ($[A']$) and video ($[V']$) data of the presenter into audio ($[A]$) and visual ($[V]$) feature vectors to quantify the non-verbal communications components of the presenter such as *vocal expressions*, *facial movements*, and *body gestures*.

We provide an overview of the previous work in multimodal sensing and emotion recognition that considers three fundamental modalities (audio, visual and cognitive cues), which are the essential components of human interaction [89]–[91]. A demonstration of how multimodal sensing and emotion recognition systems use facial and vocal signals during a dyadic conversation is shown in Fig. 5.

A. EMOTION QUANTIFICATION

Emotion recognition systems focus on computational approaches for understanding and recognizing expressive behaviors of emotion. In this paper, we assume that the perceived emotion label of human annotators describe the emotional cues in a given dataset, and use this label as a ground truth emotion of the dataset. In this section, we discuss *categorical* and *dimensional* [92], [93] affective labeling approaches that can quantify emotion of given data.

1) CATEGORICAL MODELING

of emotion uses a finite number of emotion categories to describe the emotional phenomena, such as *Angry, Happy,*

TABLE 1. Behavioral and affective cues to quantitatively measure the effectiveness and success of presentation based on crowd listeners' scores, which will be used to compute [S] vectors.

Behavioral Cues		Affective Cues	
Vocal Quality:	clarity, projection, loudness of the presenter's voice	**Valence:**	positive vs. negative attitudes and tones
Eye Contact:	consistent eye contact with audience members	**Excitement:**	excitement, passion of the presenter about the topic
Hand Gestures:	appropriate and meaningful use of hand gestures	**Engagement:**	engaging the audience with the lively presentation
Head Gestures:	appropriate use of head gestures (e.g., nodding)		
Body Postures:	confidence in body language and open posture		

Sad, Frustrated. This modeling is based upon "basic" emotions that are universally recognizable, proposed by Ekman [94].

2) DIMENSIONAL MODELING

assumes that emotion can be represented using continuous values, described in emotion dimensions. Two widely used dimensions are *valence* (positive vs. negative) and *arousal* (calm vs. excited) [95]. Other dimensions, such as dominance, have been proposed, however previous studies showed that valence and arousal dimensions can capture most of emotional phenomena [96].

Although these labeling approaches have shown to be effective in developing emotion recognition systems, the primary challenge in quantifying human emotion is the *subjectivity* in emotion expression and perception. To handle subject-dependent perception and labeling, a common practice in the research community is to take an average of multiple annotators' perceived emotion labels, discarding inconsistently labeled data. Table 1 summarizes our suggested emotional and associated behavioral cues that can be used in smart classroom systems.

B. AUDIO-VISUAL FEATURES

A computational system that can automatically sense and recognize related signals, such as facial expressions, body gestures, speech, and pupil dilation, can be used calculate the A, V , and C vectors for Boxes I and III. In this section, we elaborate on state-of-the-art speech and visual features that have been shown to be effective in emotion recognition systems. Audio-visual emotion recognition is a process of predicting emotion from behavioral cues of a speaker, such as speech [97]–[102], facial cues [103], [104], and bodily expressions [102], [105]. Surveys that study this topic in detail can be found in [92], [97], [98], and [105]–[109].

For speech-based emotion recognition, Interspeech 2013 Paralinguistic features [110] (6,373 features in total) have shown to be useful in emotion recognition in previous work. The low-level descriptors (LLDs) include 4 energy related LLDs, such as loudness, RMS energy, zero-crossing rate, 55 spectral LLDs such as MFCC, spectral energy, spectral variance, and 6 voicing related LLDs such as F0, probability of voice, log harmonic-to-noise ratio (HNR), jitter, and shimmer. The statistical functions such as mean value of peaks, amplitude, linear regression coefficients, and percentage of non-zero frames are calculated from these LLDs.

For the facial features, Action Unit (AU) features are widely used. CERT [18] can extract AU features that include 7 AU features for the upper face and 14 AU features for the lower face at each frame. Statistical functions can be computed at the segment level. For the body movement features, MODEC algorithm [111] has been proposed to estimate per-frame upper body gestures.

C. MULTIMODAL EMOTION RECOGNITION

Although the two metrics —[A], [V]— to quantify the low-level features can be extracted using readily-available feature extraction tools (e.g,. openSMILE [16] and CERT [18]), a research challenge exists in efficiently *combining* the multimodal cues to extract perceptually meaningful information without causing a dimensionality crisis that is known to overload machine learning algorithms [112].

Recent research has shown that multimodal approaches can improve the overall emotion recognition accuracy. Many studies investigate how to combine these multiple modalities and build a classification system that can effectively fuse this information using ensemble [113], filtering [114], and Bayesian network [115], [116] methods. The computational cost and efficiency in these studies is often neglected, thereby making their adaptation to real-time applications challenging. Furthermore, computationally measurable ground truth labels for education literature are under-studied.

Deep learning approaches for multimodal fusion have also shown to be effective in emotion classification [116], [117]. Deep learning models have been applied various fields, such as speech and language processing, computer vision, and automatic emotion recognition [118]–[120]. These deep network can automatically generate feature representation by learning the high-level dependencies of input dimensions [121]–[123]. Our previous work has shown that audio-visual emotion recognition can be improved using deep learning [116], [124]. Stuhlsatz *et al.* [125] also used a Generalized Discriminant Analysis (GerDA)-based deep learning models for speech emotion recognition. Brueckner and Schuller [126] has also used a two-layer deep network, achieving significant improvement in a Interspeech 2012 challenge [127].

Deep Belief Networks (DBNs) are a type of Deep Neural Network (DNN), which are widely used; within the context of our proposed smart classroom, we use DBNs for multimodal sensing. The structure of a DBN is depicted in Fig. 6, which

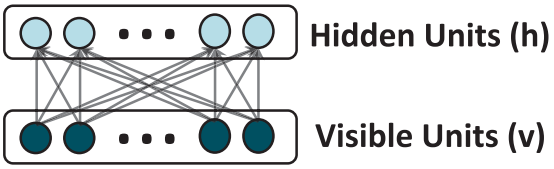

FIGURE 6. A Restricted Boltzmann Machine (RBM) is a building block for constructing Deep Belief Networks (DBNs).

contains multiple stacked Restricted Boltzmann Machines. As shown in Fig. 6, the input layer of the RBM consists of v, visible units (i.e., observation vectors of data), while there are h are hidden units inside the RBM that are learned during the training of RBM. The h hidden units are activated based on the weighted sum of the visible units using the W weights and biases. RBM training is efficient [128], because of the absence of connections between units within each layer. We use a type of RBM, namely a Gaussian RBM, which uses the following energy function. Let $\{v^{(1)}, \ldots, v^{(m)}\}$ be a given a set of observation vectors (features of training instances):

$$\begin{aligned}\mathbb{EX}(\mathbf{v}, \mathbf{h}) = &\frac{1}{2\sigma^2}\sum_i v_i^2 \\ &- \frac{1}{\sigma^2}\left(\sum_i c_i v_i + \sum_j b_j h_j + \sum_{i,j} v_i W_{ij} h_j\right) \\ &+ \lambda \sum_{j=1}^{K}\left|u - \frac{1}{m}\sum_{l=1}^{m}\mathbb{E}\left[h_j^{(l)}|v^{(l)}\right]\right|^2, \end{aligned} \tag{1}$$

where $\mathbb{EX}[\cdot]$ is the conditional expectation given v, λ is a regularization parameter, and u is the target activation of the hidden unit [123].

Deep sparse autoencoders have also been shown to effectively learn discriminative features in various domains, including emotion recognition [129]. These neural networks attempt to learn a sparse representation of inputs by reconstructing the input at its output. The objective function is to minimize the reconstruction error, and one can use the mean squared error function as follows:

$$J_{sparse}(W, b) = J(W, b) + \beta \sum_{j=1}^{N} KL(\rho||\hat{\rho}_j), \tag{2}$$

where N is the number of hidden units, ρ is a sparsity parameter, KL represents the Kullback-Leibler (KL) divergence, and $J(W, b)$ is defined as:

$$\sum_{j=1}^{N} \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}. \tag{3}$$

The deep learning methods above can be adapted for learning educationally important cues captured in smart classrooms. However, how multiple modalities interact over time, and what quantitative methods we can use to control for high variabilities in data induced by such cross-modal interactions, remain open questions.

V. HIGH PERFORMANCE REAL-TIME COMPUTING

In our proposed system, a substantial amount computation must be performed in real-time. The intensity of the computations do not allow them to be computed in the observation room of the smart classroom and access to computational resources in a datacenter or cloud is required. In this section, we provide background on previously proposed computational architectures to realize our real-time computation goal for such high-intensity computations. In Section V-A, we survey methods that allow the underlying computations to be parallelized, regardless of the utilized computational hardware. In Section V-B, we the MOCHA architecture proposed in [130], which is a VM-less mobile-cloudlet-cloud architecture; MOCHA is the architecture our system is based on and allows parallelization of the computations by utilizing multiple cloud servers.

A. PARALLELIZATION OF COMPUTATIONS

Each one of the required algorithms in our system problem inherently possesses serial and parallel characteristics based on Amdahl's law [131]–[133] and only the parallel portion of a problem can be accelerated using parallel hardware/software.

1) ONLINE/OFFLINE COMPUTATIONAL DECOMPOSITION

One approach to speeding up computations is to reformulate the original computations to require two separate phases: *online* and *offline*. While the offline portion does not depend on the real-time data, the online portion does. Because of this, the offline portion can be computed before the actual real-time computations take place; therefore, the computation time of the offline portion is not exposed in the real-time portion.

Liu *et al.* [21] describe a methodology to accelerate online internet search by separating the Personalized PageRank (PPR) algorithms into two components, the *offline* component that can be computed prior to the actual computations, and an *online* component that must be computed at the time of the search queries, i.e., *real-time*. The goal of this decomposition is to accelerate the online computations (which are *experienced as elapsed time by users*), even at the expense of slower offline computations and/or higher memory requirement. This methodology is suitable for the computations in our system [16]–[20], because the added memory requirement is an acceptable tradeoff and the intensive offline computations can be performed at a school datacenter at night, which is when the computational resources are nearly idle.

2) HARDWARE-BASED ACCELERATION

The nature of the computations in Boxes I, III, and IV determine which computational resources (e.g., CPU vs. GPU) they can use efficiently. Generally, the best application performance is achieved when a proper CPU-GPU decomposition of computations is formulated, as evidenced by the *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* [134],

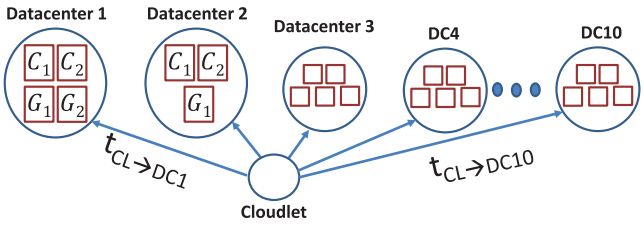

FIGURE 7. Graph representation of the cloud resources; C_1, G_1 denote the instantaneous CPU, GPU loads and $t_{CL\rightarrow DC1}$ denotes the network delay from the cloudlet (CL) to the first datacenter (DC1).

where the winner group had to research an extensive set of techniques to determine the best CPU-GPU mapping. Therefore, multiple techniques suggested by this group [135], [136] can be used to find the best CPU/GPU computation structure for optimum performance. Typically, the best approach it so use CPUs for computations that require significant single-thread performance and GPUs for computations that enjoy a high degree of parallelization [137].

B. HIGH PERFORMANCE MOBILE-CLOUD COMPUTING

Our proposed system can be conceptualized as being a mobile-cloud computational system, where the mobile devices that are used by the crowd constitute the ''mobile'' portion and the computational work is offloaded to ''cloud'' servers.

1) COMPUTATIONAL OFFLOADING

For our proposed system to be usable in a scenario where a school uses multiple cloud servers (i.e., multiple datacenters), located at a wide geographic distribution, to perform the required intense computations by the system, an infrastructure, such as the one shown in Fig. 7), can be utilized. Each server at a datacenter is modeled as a set of CPU resources with instantaneous loads $C_1, C_2 \cdots$ and GPU resources with instantaneous loads $G_1, G_2 \cdots$ A cloudlet that interfaces with all of the mobile devices serves as the key device to either intelligently schedule the offloading tasks or perform pre-computations [138]. The network delay from the cloudlet to each physical datacenter location is modeled as $t_{CL\rightarrow DCn}$, where n is the total number of datacenter locations. In this case, we propose to perform the task scheduling by the cloudlet using two sets of inequalities: **(a)** *delay inequalities* must be satisfied to ensure the computation before the allowed *real-time* target, and **(b)** *computational load inequalities* must be satisfied to ensure that the server resources are not overloaded. A set of Mixed Integer Linear Programming (MILP) can be derived for the solution of the optimum scheduling, which is commonly used for mobile-cloud computing scheduling tasks [139].

2) MOBILE CLOUD HYBRID ARCHITECTURE (MOCHA)

The MObile Cloud Hybrid Architecture (MOCHA) [22], [130] prescribes a three-level mobile-cloudlet-cloud acceleration architecture, in which the acquisition source (mobile) utilizes a nearby pre-computation device (cloudlet) to parallelize the task and outsource it to a large number of cloud servers (cloud) [140]. This architecture is proposed to allow computationally-intensive mobile-cloud applications, such as mobile-loud face recognition [141], [142] or military applications [130], to be performed in real-time [10], [138].

Our proposed smart classroom computational infrastructure in Fig. 3 is modeled based on MOCHA with the three levels being the *data acquisition* (inside the classroom), *data pre-processing* (inside the observation room, which is a part of the smart classroom physical space), and the *parallel computing* (which is at the datacenter).

While the generic massively parallel computing techniques are suitable for computations that can be performed using a *single* server, the techniques that are introduced within the context of MOCHA allow distributed parallelization across *multiple servers*, which can also be in different geographic regions; this is necessary to make our proposed system practical for schools that do not own their datacenter and intend to use public cloud computing platforms, such as Amazon EC2 [143].

3) ACCELERATION AS A SERVICE (AXAAS)

A novel service model, Acceleration as a Service (AXaaS) [23], [144], [145], describes a platform in which Telecom Service Providers (TSPs) rent computational acceleration as a monthly service (Giga FLOPS per month), much like the data services (Giga Bytes per month). This service model describes how intense peak computational requests can be supported by the TSP using their datacenters (e.g., Verizon Terremark [146]) through a monthly computational service subscription. We argue that such a service can also allow schools to purchase a service subscription for performing the computational that are performed during training and presentations for a smart classroom. This can eliminate the need for the schools to invest in large datacenters, even a cloud computing subscription.

4) VOLUNTEER CROWD-COMPUTING

Although we envision the *cloudlet* pre-computation device as being a small server that is placed on a rack inside the observation room of a smart classroom (as shown in Fig. 3), an emerging *volunteer computing* paradigm describes an architecture where a set of volunteering mobile device users contribute computational resources to a large computational task, such as protein folding [147], [148]. This allows a substantial computational acceleration for tasks that can be massively parallelized. Although volunteer computing is not suitable for the actual computations of Boxes I, III, and IV, it can be suitable for the *pre-precessing*, which can potentially eliminate the cloudlet and render the system more suitable for schools with a restricted equipment budget.

5) SECURITY AND PRIVACY OF SENSED DATA

Generally, when public cloud services are used, the security and privacy of the acquired data becomes a concern.

While known cryptography-based techniques can be used [149] to ensure the safe (i.e., encrypted) transportation of the data from the mobile devices to the cloud and back, more advanced techniques also allow secure computations in public clouds, such a Fully Homomorphic Encryption [150]–[152] or Paillier Encryption [153]. Despite their exciting nature, these encryption algorithms are extremely compute-intensive, which renders them impractical [154]–[156]. Typically, the best encryption to use is Advanced Encryption Standard (AES) [157], which provides a good balance between computational intensity and security, although it does not allow computations on encrypted data. Well established Elliptic Curve Cryptography (ECC) provides an interesting alternative, especially in Wireless Sensor Networks, where power consumption is the primary concern.

VI. SYSTEM COMPONENTS

The conceptual smart classroom system depicted in Fig. 2 requires the development of Boxes I–IV, as described in Section II. In this section, we provide a detailed analysis of each box, in terms of the algorithms and technologies required for its design.

A. DATA ACQUISITION FROM THE PRESENTER (BOX I) AND CROWD LISTENERS (BOX II)

The design of Boxes I and II requires extensive consideration of the subjectivity of human scores and the received feedback.

1) CROWD SCORE ACQUISITION

One of the most important research challenges to tackle in the design of our system is the testing our hypothesis **H1** (introduced in Section II), which states that *it is possible to quantify the crowd scores* and create an $[S]$ vector that captures the general crowd consensus related to ''best presentation practices.'' To provide a template for the crowd to evaluate (i.e., *score*) a presentation, we propose a set of questions in Table 1 for the crowd to answer —using a mobile app— in real-time during a presentation.

The **selection criteria of the questions in Table 1** was motivated by the following:

- they are based on the foundation of educational theories presented in [158]–[160] and provide meaningful *psychometrics* for evaluating presenter performance,
- they can provide a quantitative answer that is monotonically increasing, i.e., 3 is always higher than 2, etc.,
- they are reasonably statistically independent, i.e., the answer to each question yields a high additional information content, and lastly,
- there are as few questions as possible, to avoid distracting the crowd. A different set –or number– of questions may impact the cognitive load and inter-rater agreement among the crowds. Rather than asking all of the eight questions at the same time to each listener in the crowd, Box II must be designed to randomly cycle through these questions so that each listener can answer a much fewer number of the questions (e.g., three questions for eye contact ('Eye), hand gestures ('Hand'), and head gestures ('Head') in Fig. 8) at a given prompting interval, e.g., 5 minutes in the *Training Mode* of a presentation.

Table 1 lists the key components of verbal, nonverbal, and affective behavior that has been shown to influence how humans communicate. There are a host of measures and surveys in the literature that purport to measure these components [161], [162]. We have selected the minimum number of items (one item per component) to measure crowd perception regarding each one of the components. This makes the survey quick, clear, and easy to answer and avoids imposing unwanted cognitive load or frustration to the crowd responders. To quantify the responses to these questions, we will be using a 5-point monotonically increasing Likert scale (1.Poor, 2.fair, 3.average, 4.good, 5.excellent).

2) THE $[S]$ VECTOR

Figure 8 depicts our proposed Crowd Annotation Engine (Box II), which takes the answers from N experts and M peers, and combine these answers to generate the $[S]$ vector as follows. First, individual answers of each crowd listener is stored in vectors named $[E']$ for N experts, e.g., $[E']_{Joe}$ and $[E']_{Emma}$ shown in Figure 8, and $[P']$ for the M peers, e.g., $[P']_{Jill}$, $[P']_{Louis}$, $[P']_{Adri}$, and $[P']_{Peter}$. Next, combined $[P]$ and $[E]$ vectors are calculated from individual $[P']$ and $[E']$ vectors, after filtering each person's answer against biasing using an outlier detection algorithm (e.g., [37]) and a statistical method to improve inter-rater agreement (e.g., Fleiss' Kappa statistics). Finally, the $[S]$ vector is computed as $[S]=\alpha\cdot[E]+\beta\cdot[P]$.

3) AUDIO ($[A]$), VISUAL ($[V]$), COGNITIVE ($[C]$) CUE ACQUISITION

The design of Box I involves the adaptation of the existing multimodal sensing and emotion recognition methodologies to create a unified system that automatically captures salient behavioral patterns of the presenter. The three modalities that we consider in this paper —audio, visual, and cognitive cues— are the essential components of human interaction [89]–[91]. They affect and regulate how we communicate with each other, and how we perceive, judge, and react to the outside world [89]–[91]. A computational system that can automatically sense and recognize related signals, such as facial expressions, body gestures, speech, and pupil dilation, can calculate the $[A]$, $[V]$, $[C]$ vectors for use in Box I. A demonstration of our proposed method to extract these $[A]$, $[V]$, $[C]$ vectors during a presentation is shown in Figure 5.

B. IMPLEMENTATION OF MACHINE INTELLIGENCE (BOX III)

We now describe how Box III, the machine intelligence component of our system, functions to automatically estimate crowd score $[S^*]$ from the audio ($[A]$) and visual ($[V]$) cues of the presenter and expert ($[E]$) and peer ($[P]$) scores from the crowd. The overview design of our proposed system is shown in Figure 9. Generative learning methods that

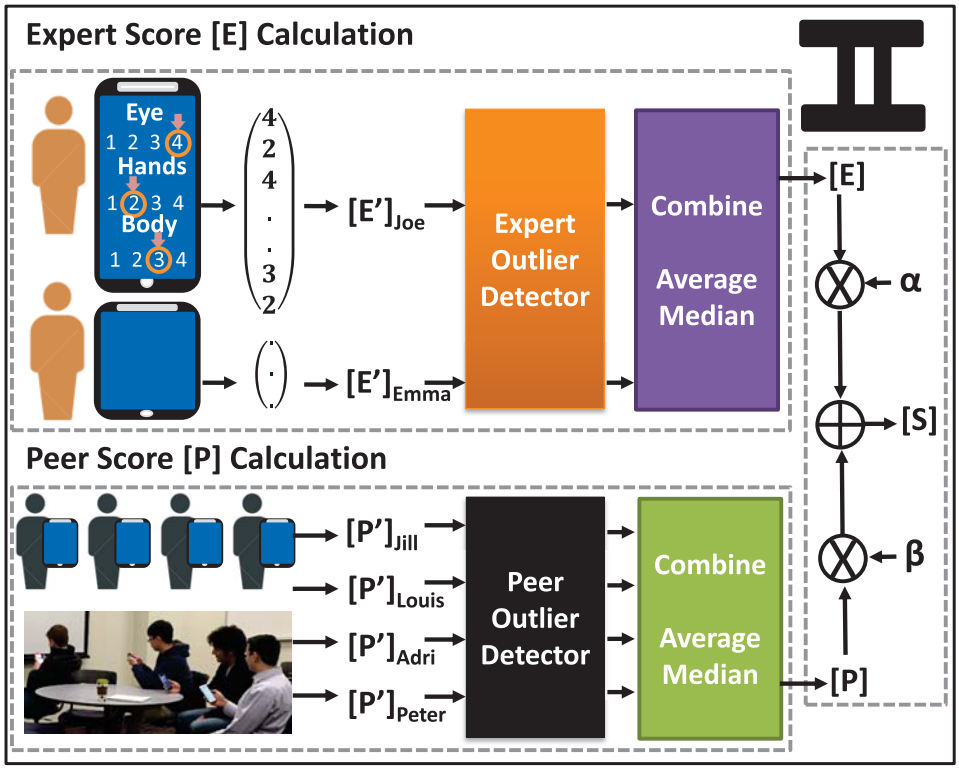

FIGURE 8. Demonstration of Box II with experts and peers using a mobile phone application to score the presentation. Raw inputs to the mobile app are filtered and combined into [*E*] and [*P*] vectors, which are eventually be used to compute the final [*S*] vector with two system parameters α and β.

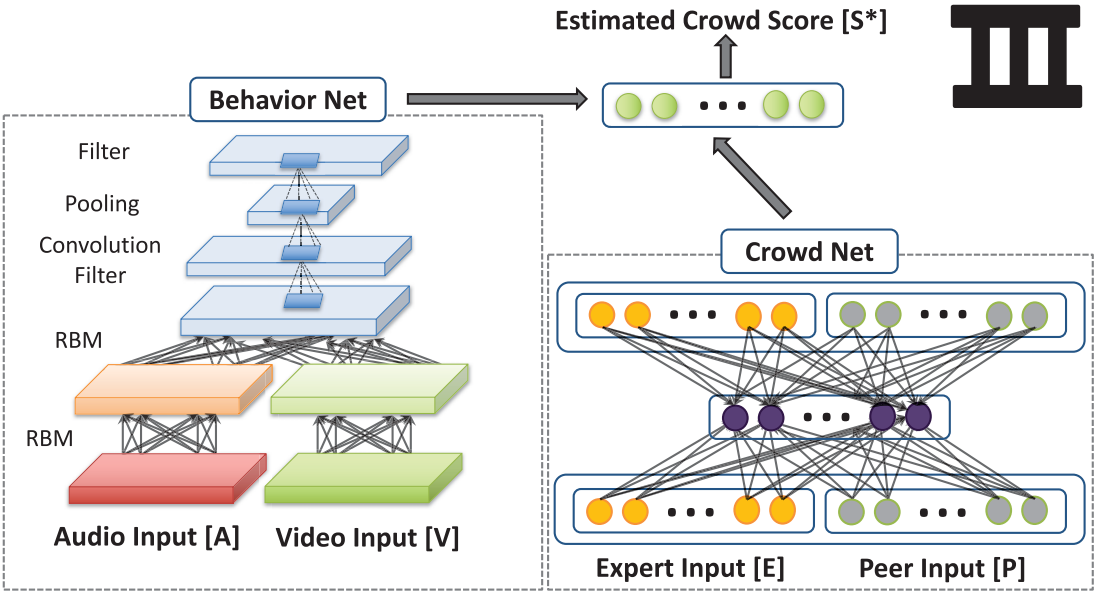

FIGURE 9. Our proposed machine intelligence system (Box III). It combines *Behavior Net*, a deep belief network combined with convolutional filter and max pooling operators, and *Crowd Net*, a sparse autoencoder, to automatically estimate crowd score [*S**] from [*A*, *V*, *E*, *P*] vectors.

underlie the state-of-the-art deep neural networks, namely, Deep Belief Networks (DBNs) [122] and Convolutional Neural Networks (CNNs) [136], are adapted to associate the relationships between these vectors. In contrast with prior work that attempts to directly find the relationships between audio and visual data, we propose to use the state-of-the-art

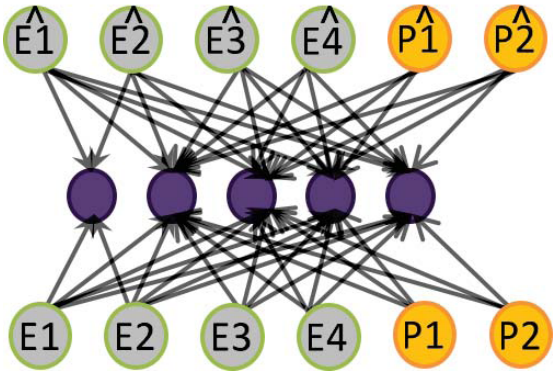

FIGURE 10. Sparse autoencoder for learning relationship between [*P*] and [*E*] vectors.

convolution filters and max pooling methods for enforcing efficiency and transform-invariance of the deep neural network. The proposed deep neural network-based design can learn the complex, non-linear interactions between each modality cues as well as presentation cues and crowd scores.

1) BEHAVIOR NET

We propose to first learn DBN-based high-level representation to fuse the multimodal information from audio and video modalities. We propose to use the first layer of the DBN includes two Gaussian-RBMs, each for audio and visual features, with sparsity regularization introduced in [123]. The upper layers are binary-RBMs, where the concatenation of the posteriors $Pr(h|v)$ of each audio and video RBM is used as the observed vector for the second layer. We employ a sigmoid function as an activation function of each node. We then use convolution filters and pooling layers, to learn features that are local and translation invariant [163]–[165].

2) CROWD NET

Sparse autoencoder (Fig. 10) fits well in the context of learning α and β for combining the peer ([P]) and expert ([E]) scores collected from Box II. An autoencorder is a neural network that sets the input and output to be equal, so that the network automatically learns sparse representation of the input data. Consider a set of expert inputs $\{E^{(1)}, \ldots, E^{(u)}\}$ and peer inputs $\{P^{(1)}, \ldots, P^{(w)}\}$, where $E^{(i)} \in \mathbb{R}^8$ and $P^{(i)} \in \mathbb{R}^8$, each dimension represents one of the eight-dimensional crowd score vectors that we defined in Section VI-A. We propose to learn the sparse representation that combines the information in both of the sets by learning a function $f_{w,b}(x) = \hat{x} \approx x$. For instance, as shown in Fig. 10, $x = P1, P2, E1, \ldots, E4$. It learns an approximation of the identity function so that the middle layer between the input and output layers learn the sparse representation of the input vectors.

3) ESTIMATED CROWD SCORE [S^*] VECTOR

We propose to combine the posterior probabilities of learned representation from the Behavior Net and Crowd Net and use the activation of this final layer as the estimated crowd score, [S^*]. This combination of the two networks enables us to automatically learn high-level complex interactions between the behavioral cues and crowd scores. How [S^*] deviates based on changes in the input [A, V, E, P] vectors must be investigated, in order to provide insight into how crowd score is associated with behavioral changes.

C. DESIGN OF THE PERSONALIZED FEEDBACK INTERFACE (BOX IV)

In this section, we describe the functionality of Box IV, i.e., the feedback interface between the machine and the presenter in both training and presentation modes. The difference between these two modes is that in *Training Mode*, Box III learns the crowd behavior, while Box IV learns the individual cognitive capabilities of a presenter. Alternatively, in *Presentation Mode*, Box III uses its knowledge to emulate the crowd behavior by supplying an estimated crowd score (the [S^*] vector) to Box IV, which, in turn, uses its knowledge of individual presenters to make suggestions to them in real-time. These suggestions are made through the haptic glove, as well as visual cues with a feedback delay —$t_{feedback}$— that matches the cognitive capabilities of individual presenters. After receiving the machine feedback, the presenter makes changes to his/her facial expressions, gesture, and voice intonation to improve the feedback; here, the definition of "improving" is set forth at the very beginning of the presentation, which can be turning frowny faces to smiley faces or reducing the vibrations in the haptic glove, etc.

1) FEEDBACK VECTOR ([F])

We propose a Feedback Engine (Box IV) that outputs a feedback vector ([F]), consisting of multiple rows, where each row corresponds to a modality; for example, the intensity of the vibration of the haptic glove is the first row (e.g., with vibration intensities from 1 to 5), while the intensity of the smiley face can be another row (i.e., 1 = frowny face, 2 = slightly unhappy, 3 = neutral, 4 = slightly happy, and 5 = fully smiling) etc. Bar plots and their colors can be similarly selected, as shown on the right side of Fig. 12, where red and green ends of the spectrum show poor and excellent performance in the relevant aspect (e.g., eye contact), respectively. Bar graphs (i.e., skill-meters) can be used in conjunction with colors (as in Fig. 4) to render the feedback message easier to understand. A variety of feedback modalities, as well as their quantification might prove to be useful and must be investigated. A possible visual feedback modality could be an icon of a human with open vs. closed arms to indicate open versus closed posture [47].

Feedback saliency: An important factor to consider when constructing a feedback dashboard is saliency, which refers to the feedback being prominent and easy to notice and understand. The presenter's attention can be drawn to feedback by vibrations via the haptic glove. The icons, colors, and bars used in the feedback screen need to be meticulously designed in terms of size, appearance, color, and order to make them easy to notice and understand for the presenter.

Ease of understanding refers to when the message (feedback) induces the lowest degree of cognitive load to the presenter while carrying the most succinct and important message. An example of this could be a red human icon with arms crossed appearing on the screen, that indicates to the presenter that it is not appropriate to have a closed posture.

Feedback needs to be *tailored* to each presenter in terms of modality, contents, and timing. The timing is determined by the importance of the feedback and considering the presenter's cognitive load at the moment. If the presenter has a very high cognitive load at one moment, he/she will not be able to receive and *digest* and ultimately implement the feedback appropriately. What distinguishes our work from the existing works in the literature [47], [64], [166] is that we propose integrating crowd ratings (crowd scores) into the machine feedback echoed to the presenter. Additionally, we propose providing a broader range of nonverbal and affective feedback to the presenter.

2) COGNITIVE LOAD CURVE

Humans have a wide variety of learning styles, background knowledge levels, and self-regulation skills. These differences, along with other environmental factors (e.g., task difficulty, time available), contribute to differing levels of cognitive load for each individual while performing a task [167]. One of the ways of determining whether feedback has been effective is to examine if the individual has integrated the provided suggestions in their performance, i.e. *compliance*. If the presenter is *not compliant*, it can be due to either too high cognitive load or inattention to feedback. If the cognitive load is determined to be too high, the feedback latency ($t_{feedback}$) could be increased to allow the presenter to process the demands of the task at hand or the saliency and type of the feedback will be changed to re-evaluate *compliance*. One major reason for not integrating feedback is that the human's brain is already processing the task at hand or the task is difficult, and insufficient cognitive resources are available for receiving, processing, and implementing feedback. If the presenter's moment-to-moment level of cognitive load is not attended to, the feedback will be ineffective and wasted, besides frustrating the presenter and leading to their disengagement. In general, the contents, amount, and mode of feedback need to be tailored to individual presenters in real-time to maximize effectiveness and minimize cognitive burden on them.

Figure 11 highlights this phenomena and is termed the *Cognitive Load Curve*, which is different for each individual and it correlates the cognitive load (shown as a scalar value C for simplicity) to the speed at which the feedback is provided to an individual; feedback that is *too slow* (represented as the region where $C < C_{min}$ in Fig. 11) does not cause cognitive overload but it is useless, because an ideal feedback must be provided when the higher cognitive processing area of the brain (e.g., the pre-frontal cortex — PFC) [168] is ready to receive such feedback and incorporate it into its learning process. As we will detail in Section VII-B, this time is typically less than a second. The ideal feedback time depends on an individual; in Fig. 11, this area is represented by the region ($C_{max} > C > C_{min}$) where the cognitive load increases when the feedback time decreases, however, the feedback is useful in achieving the "learning" outcome. The third region, which is marked as *too fast* (the $C > C_{max}$ region) causes a significant cognitive overload, because processing information so fast requires substantial higher-level cognitive processing; this causes presenter frustration, as described previously. At a certain point, which is marked as C_{max} on the curve, this feedback becomes so fast that the presenter *gives up* on paying attention to it, hence the sudden drop on the load below C_{max}.

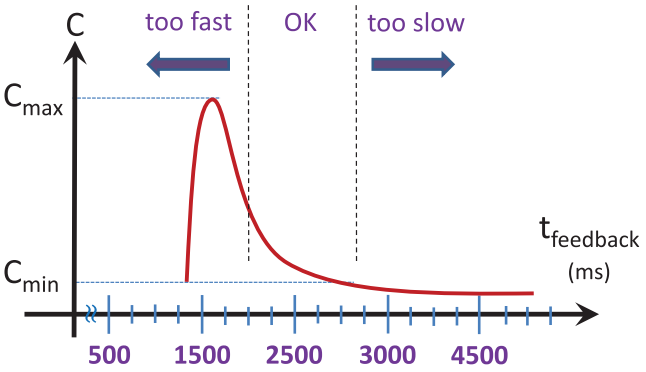

FIGURE 11. An example *Cognitive Load Curve* for a presenter. There is an optimum feedback response time ($t_{feedback}$) for each presenter; feedback that is too fast is ignored by the presenter, while too slow does not provide instructive value.

3) DESIGN OF THE FEEDBACK ENGINE (**BOX IV**)

From our preliminary discussion, we conclude that Box IV is required to provide a feedback vector $[F]$ to each individual presenter in a *time sequence*, i.e., a feedback vector at a given interval; therefore, we propose to apply the $[C]$ and $[S^*]$ vectors to its input by a time sequencer (e.g., every 100 ms as shown in Fig. 12). Previous research [112], [169], [170] has used a similar scheme in determining the importance of "*the same metric at different time intervals*" within the context of cardiac monitoring; the QT interval in a heart beat —using an ECG recording— was entered into a support vector machine (SVM) at different hours and used as separate *dimensions*. Our proposed scheme differs from the previous work, in that we overcome a *dimensionality problem* for large problem sizes [112].

We propose a similar idea to design Box IV, with one exception: A *Recurrent Neural Network (RNN)* eliminates the necessity to enter each time slot as a separate dimension, thereby easing the dimensionality pressure on the learning network; instead, an example application of an RNN, a **long-short term memory network (LSTM)**, includes an internal memory and is governed by internal *remembering* and *forgetting* functionality and has been in use since 2016 by Apple [171] for personal hand-writing recognition, which is one form of "understanding the characteristics of a person."

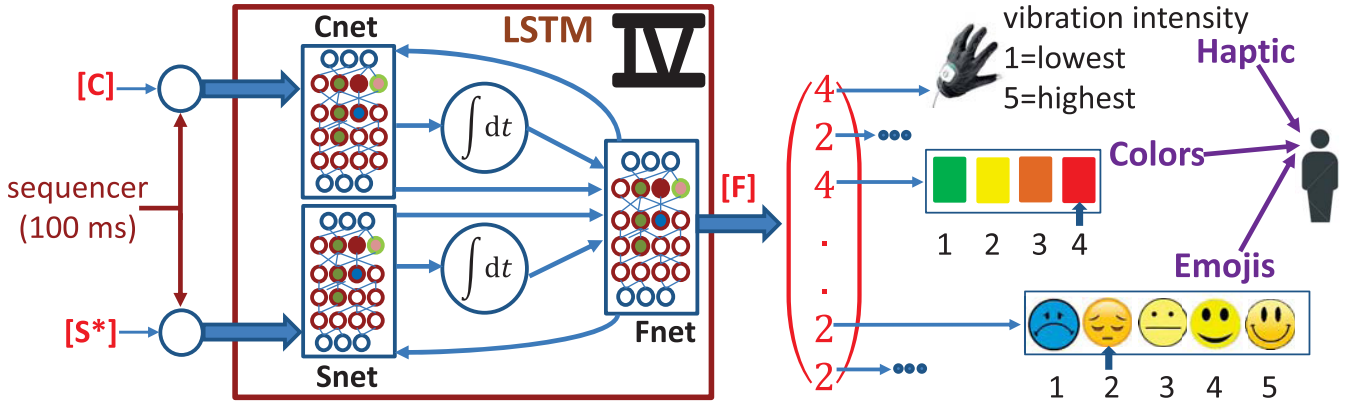

FIGURE 12. A long-short term memory (LSTM) recurring neural network (RNN) can implement the functionality desired for Box IV through remember/forget intervals, which matches the cognitive curve of individual presenters.

We suggest that they are an excellent candidate for the design of Box IV.

Figure 12 is our proposed application of an LSTM to designing our Box IV: the Snet and the Cnet networks are proposed to interface the sequenced estimated crowd scores ($[S^*]$) and the cognitive loads $[C]$, respectively, and their outputs are connected to the Fnet network, which combines them to generate the final feedback vector $[F]$. The two integrators are utilized to satisfy the "remembering" functionality and the shortcut connections from Snet and Cnet into Fnet implements the "forgetting" functionality. The *recurrence* feature, which is typical in any RNN, is implemented by the reverse connections from the Fnet back to Snet and Cnet. Different forget/remember intervals to implement the *optimum cognitive feedback time*, which matches the cognitive curve of individual presenters must be investigated for optimum performance.

VII. EVALUATION

In this section, we provide a quantitative evaluation of the proposed system. Our evaluation is based on estimated computation and communication delay values, obtained by running appropriate software that is representative of the component of a given part of the system.

A. EVALUATION METRICS FOR OUTCOMES

The success of the proposed system can be evaluated based on two different criteria.

1) TECHNICAL SUCCESS

(i) *accuracy* and **(ii)** *latency* of our system directly quantify its technical success. To determine the *accuracy* of our algorithms, the ground truth —from the crowd ratings— can be compared to the predicted scores computed by the machine intelligence. Furthermore, mutual information between the $[A, V, C]$ vectors and $[S]$ vectors can be calculated to ensure that our algorithms use relevant features to the prediction outcomes. To determine the *latency* of our algorithms, a timing mechanism must be implemented, which measures the data-acquisition-related, network-related, pre-computation-related, and computation-related delays to determine the computation vs. network travel time of the data.

2) EDUCATIONAL SUCCESS

In Table 1, we define behavioral and affective cues, based on psychometric studies in education [160], [172], [173], to quantify the effectiveness of a presentation.

B. EVALUATION FRAMEWORK FOR REAL-TIME PERFORMANCE

Before we investigate the magnitude of necessary computational acceleration to achieve real-time performance, we will provide a reasonable definition for "*real-time*" in quantitative terms.

1) DEFINITION OF *REAL-TIME*

In Fig. 2, the delay, Presenter→ Cloudlet→ Cloud (Box I→Box III→Box IV) → Cloudlet→ Haptic Glove, is the *longest path delay* of our proposed system in *Presentation Mode* and includes Box II in *Training Mode*. This delay must be comfortably below how fast the presenter's brain processes the machine feedback. In a pioneering neuroscience study, Libet *et al.* [174] established a *delay* for when our brain receives sensory input vs. when an action becomes *conscious*; it involves the interplay between the two parts of the human brain [175]: the *motor* part that prepares an *action* in $\leq$200 ms and the *conscious part* that decides to either take the action or veto it in another $\approx$300 ms, for a total of $\approx$500 ms [176]. Conscious part is much slower due to its denser neural layers [177]. We define $t_{cognitive}$ as the rough estimate for how fast a machine feedback must be provided to the human brain to be useful and choose ($t_{cognitive}$ =150 ms) to guide our preliminary estimations, which is comfortably faster than the motor part of the brain.

2) THE *R* METRIC

Let us define $t_{feedback}$ as the time it takes for the machine to provide a feedback to the presenter by traversing the

TABLE 2. Estimated $t_{preprocessing}$ and $t_{network}$ delay components. M=Mobile,CL=Cloudlet, DC=datacenter.

Description	Time
Capture from camera	33.3 ms
CL pre-processing [22]	35 ms
M $\longleftrightarrow$ **CL**$\longleftrightarrow$ **DC** propagation	25 ms
Audio, video data transfer	25 ms
$t_{preprocess} + t_{network}$	118.3 ms

Presenter $\rightarrow \ldots \rightarrow$ Haptic Glove path. Three scenarios emerge:

$$R = \frac{t_{feedback}}{t_{cognitive}} = \frac{t_{preprocess} + t_{network} + t_{computation}}{t_{cognitive}}$$

$$\Longrightarrow \begin{cases} R < 1, & \text{System is too slow} \\ R = 1, & \text{System is real-time} \\ R > 1, & \text{System is too fast} \end{cases}$$

where $t_{preprocess}$, $t_{computation}$ and $t_{network}$ are the *pre-processing time*, *computational time*, and *network delay* components of $t_{feedback}$, respectively; $t_{preprocess}$ involves the speed at which the cloudlet in the observation room performs its computations, while $t_{network}$ includes the total amount of time that the data travels from its initial acquisition point to the cloudlet, cloud, and back to the cloudlet and haptic glove. $t_{computation}$ is composed of the time that it takes to perform the computations in Boxes I–IV and is different for *Training Mode* vs. *Presentation Mode*, because most of the machine learning algorithms exhibit drastically different computational properties in training vs. test.

C. ESTIMATING THE DELAYS FOR BOX I–IV

To determine the magnitude of the computational acceleration necessary for *real-time presenter feedback* (i.e., $R \geq 1$), we assume a 30 frames/sec video capture (RGB, 24 b/pixel) and a 44 100 Hz audio sampling rate (16 b/sample). Therefore, each video frame is 33.3 ms, which must be transmitted to the cloudlet for pre-processing along with the audio data captured from the microphones. We will estimate the delays that contribute to the $t_{feedback}$.

1) PREPROCESSING AND NETWORK DELAYS

Table 2 lists the estimated delays to acquire a single image frame from the microphones, camera, and mobile devices (denoted as **M**), pre-compute it in the cloudlet (**CL**), and transfer it to the datacenter (**DC**), based on the software and studies generated during the study in [22] and [178]. The *cloudlet* is the vital part of the architecture proposed in [22] to achieve real-time mobile-cloud performance by profiling and using cloud resources intelligently [23], [130], [138], [141]. According to Table 2, $t_{preprocess} + t_{network} = 118.3$ ms, leaving only 31.7 ms for $t_{computation}$ to achieve real-time performance ($R = 1$). We note here that the delays shown in Table 2 are not amenable to parallelization; *network delays* involve data-size-dependent transfer delays, round trip network delays, which cannot be improved without new network infrastructure, while *pre-processing delays* in the cloudlet cannot be improved unless multiple cloudlets are deployed in the observation room, which will not necessarily help due to the communication delays introduced among multiple cloudlets. Alternatively, *computation delays* can be improved by parallelization.

2) COMPUTATION DELAYS

Table 3 provides our preliminary results using commonly-accepted software to provide the functionality in Boxes I–IV.

- *Box I:* requires the processing of the raw $[A']$ vector to compute the audio features vector $[A]$; we use the well-accepted PRAAT software [179], which computes the prosodic and spectral features of speech; our test run takes 0.99 s to compute 4 seconds of speech. This means that within each $t_{cognitive} = 150$ ms, 37.1 ms is required. Similarly, facial feature extraction using CERT [18] and Multimodal Human Pose Estimation (MODEC) [111] are suitable for the $[V]$ vector in Box I; test runs took 166.7 ms and 1501 ms, respectively (per frame). Computing the cognitive load, based on a study in [180] takes $\approx$25 ms, conservatively (per frame). Therefore, Box I computation time per $t_{cognitive}$ is $37.1 + 166.7 + 1501 + 25 \approx 1730$ ms.
- *Box II:* eliminates outliers and potentially malicious scores in the $[S]$ vector; test results from [14] indicate a runtime of <50 ms.
- *Box III:* uses a deep learning network to learn the $[A]$, $[V]$, $[S]$ associations; we use Deep Belief Networks (DBN) [116] with 2000 instances and 200 nodes, 100 for audio, 100 for visual data, 50 nodes for each Restricted Boltzmann machine (RBM) for the first layer and 30 nodes for the final hidden layer. As shown in Fig. 9, the DBN consists of two 100 (input) $-$ 50 (output) RBMs on the first layer for audio and video modalities, respectively; finally, a 100 (input) $-$ 30 (output) RBM is at the final layer. The inference of an RBM at the first layer takes 4.108 ms, while the training time is 2.53 s. Cumulative, the DBN takes 75.4 ms total.
- *Box IV:* requires a machine learning platform that takes into account the cognitive load C to personalize the feedback; a generic Support Vector Machine (SVM) [181] is suitable; a test run took $\approx$95 ms for training and $\approx$64 ms for testing.

To summarize, the computations in Boxes I–IV take $1730 + 50 + 2530 + 94.7 = 4405$ ms for *Training Mode* and $1730 + 0 + 75.4 + 63.9 = 1869$ ms for *Presentation Mode*.

D. COMPUTATIONAL ACCELERATION

We target real-time performance in training as well as presentation. Therefore, we use the maximum of the two, $max(t_{Training}, t_{Presentation}) = 4405$ ms, to reach our $t_{computation} = 31.7$ ms goal, implying a necessary computational acceleration of $4405/31.7 \approx 139\times$. Furthermore,

TABLE 3. Computation runtime results for Boxes I–IV in milliseconds for a single frame. The * symbol indicates numbers taken from another study, and the rest are from our actual test runs.

Methodology	Box	Computation	Time
PRAAT [179]	I	$[A]$ prosodic, spectral features	37.1
CERT [18]	I	$[V]$ Facial features	166.7
MODEC [111]	I	$[V]$ Human Gesture	1501
Pupillometry* [180]	I	$[C]$ Cognitive Load	25
Crowd-Sense* [14]	II	$[S]$ reputations, outliers	< 50
DBN [116]	III	Training Mode	2530
Deep Learn. Engine		Presentation Mode	75.4
SVM [181]	IV	Training Mode	94.7
Feedback Engine		Presentation Mode	63.9
$t_{Training}$	All	Training Mode	4405
$t_{Presentation}$		Presentation Mode	1869

considering that the datacenter should be designed to serve multiple smart classrooms (e.g., 5–10) simultaneously, an acceleration target of *three orders-of-magnitude* (1000×) is realistic. In this section, we present multiple research ideas to achieve this goal in the following categories: **(i)** *single-server acceleration*, which aims to realize acceleration using the resources of a single server with one or more CPUs and GPUs, **(ii)** *multi-server (cloud-based) acceleration*, which targets distributed computing algorithms that use resources spread over multiple servers in the cloud, **(iii)** *pre-computation acceleration*, which aims to accelerate the acquisition and pre-processing of the data, rather than the computation itself, and **(iv)** *online/offline computational decomposition*, which identifies the components of the computations that are performed and reformulates them to use pre-computed vs. real-time computations; while the offline computations (pre-computed) can be performed before the presentation, when the computational resources are not loaded, the online computations can be computed during a presentation (as described in Section V-A).

1) SINGLE SERVER ACCELERATION

The results provided in Section VII-C are obtained by using single-threaded programs. Using parallel (multi-threaded) CPU programming and Massively Parallel (GPU) Computing, or using a Xeon Phi MIC co-processor, significant acceleration can be achieved [137]. Based on the results reported in [137], a single server at the SUNY Albany CEASHPC cluster, which contains two 14-core Xeon E5-2680V4 CPUs [182] and a 4992-core Nvidia Tesla K80 GPU [183]), a ≈20–40× acceleration can be expected.

2) MULTI-SERVER (CLOUD-BASED) ACCELERATION

The SUNY CEAS HPC cluster has 12 servers, which can further improve the aggregate acceleration by up to 12× by proper task scheduling. Furthermore, resource-limited institutions can use public cloud resources (e.g., Amazon EC2 [143]), rather than their own datacenter to perform the intensive computations. These public cloud services can have multiple servers with heterogeneous resources, as described in Section V-B, which can be used to achieve an order-of-magnitude acceleration.

3) PRE-COMPUTATION ACCELERATION

From Table 2, we observe that the acquisition of a video frame (33.3 ms), its cloudlet pre-processing ($t_{preprocess} = 35$ ms), and the network delays involved in the transmission of this captured data (25 ms) amount to a total of 93.3 ms. Hypothetically, if this delay can be reduced to 48.3 ms (about half), this would save 45 ms in Table 2, allowing the computation side a wider delay of $45 + 31.7 = 76.7$ ms, rather than the 31.7 ms that we based our acceleration computations on. Effectively, this is equivalent to achieving $76.7/31.7 = 2.4\times$ before even we perform the actual computations.

VIII. OPEN ISSUES AND CHALLENGES

The success of the construction of the system that is proposed in this paper depends on an advancement of the state of the art on multiple fronts. In this section, we highlight the open issues and challenges in engineering and education research.

A. ENGINEERING RESEARCH

In future work, we will investigate how to handle the label noise in crowd score and emotion recognition. One approach is to use complexity measure of classifiers to automatically identify the noisiness of labels. We can then use a semi-supervised approach to remove labels for the noisy data but retain the feature information. This approach may open a new gateway to minimize human-related artifacts.

Also, we will investigate multi-person, face-to-face social interactions in smart classrooms in our future work. Automatic behavioral feedback loops during multi-person social interactions have been studied with an aim to help the users to perceive their speaking time for balanced group discussions [64]. They use wearable devices, such as Myo armbands, headphones, and Google Glass, to capture multimodal behaviors of multiple users and deliver tactile (Myo armband), auditory (headphones), visual head-mounted (Google Glass), and visual remote (common monitor) stimulation. The authors also found the Google Glass and vibro-tactile feedback delivery devices to be the most disturbing. The study found that users are skeptical towards the usefulness of such systems, which opens new research questions to our proposed project.

To achieve real-time performance for the required computations in our system, a thorough study of the nature of the computations is necessary in future work. As described in Section V, only the portions of the computations that are parallel in nature can benefit the vast computational power of GPUs [137]. By determining the serial/parallelizable portions of the computations in the system, cloud resources can be used efficiently. For schools that use public and inexpensive cloud resources for their system (e.g., Amazon EC2 [143]), this becomes a much more important challenge, because resource inefficiencies correspond to increased cloud costs.

For large classrooms or schools with multiple smart classrooms, usage of multiple cloud servers is necessary; in such a scenario, being able to perform part of all of the pre-processing in the cloudlet becomes important. For schools that do not have an observation room (thus no way to utilize a cloudlet), one interesting future research challenge is to use the distributed network of the smartphones of the listeners as a cloudlet. In this scenario, the input data is received by the microphones and cameras of the listeners in the crowd and a distributed algorithm decides how to outsource the pre-computations among the existing members of the crowd; the resulting pre-computed multimodal data is then outsources into the cloud.

In addition to performing the *pre-computations* by the listeners in the crowd, another future research opportunity is to use the listeners' mobile devices to perform the *actual* computations, as well as the pre-computations. This will, however, the existence of fairly computationally-capable mobile devices, such as laptops or high-powered tablets. The computational strain on the mobile devices can be reduced by using online/offline decomposition, as described in Section V-B. Equation 1, which is performed in Box III, can benefit from online/offline computation decomposition as follows: this equation contains a significant number of multiplications that can be computed based on some reasonable intervals of $[A]$, $[V]$, and $[S]$ scores offline. The results can, then, be saved as a *look up table*, which can be used to simply gather the parameters during the presentation (real-time). Because the expected ranges of these vectors is not totally arbitrary, this can create a look-up table that contains many results under certain assumed conditions. For example, the gestures of presenters are restricted to certain values, because nobody is expected to present *upside down*. If we assume that the range of expected ''reasonable'' values for each vector is 50 for audio, 50 for video, and 50 for crowd scores, this quantitatively implies storing pre-computations for $50^3 = 125\,000$ possibilities, which significantly increases the memory usage, while turning hundreds of thousands of multiplications into a mere single look-up. Furthermore, as the machine intelligence learns what constitutes ''reasonable assumptions'' for audio, video, and crowd vectors, it can learn to store only the meaningful pre-computations to achieve storage efficiency. The expected acceleration can be 5–10$\times$ or even more.

B. EDUCATIONAL RESEARCH

In terms of research in education, the modality, time, and amount of real-time adaptive feedback can be further catered to individual presenters' based on their moment-to-moment cognitive load level and their preferred feedback modality. There is still debate in the literature about whether one particular modality would be the best vehicle to transmit feedback or if student modality preferences need to be heeded in presentation of the feedback [53], [66].

Future research could examine the use of eye-tracking glasses worn by the presenter to determine if he/she is paying attention to the feedback, where they look at, for how long, and how often. These valuable data can help in designing more appealing and meaningful feedback representations. Also, more targeted haptic prompts can be provided to the presenter to attract their attention to urgent and time-sensitive feedback.

Research in this area would also greatly benefit from the use of physiological sensors to measure the presenter's electro-dermal activation or galvanic skin response [184], which corresponds directly to their arousal and anxiety. These data can be aligned with emotions detected by cameras and the data coming from the crowd in order to provide remedial feedback to the presenter to help them in regulating their negative emotions or nervousness.

Last but not least, types of data collected from the crowd can be extended to include self-reports of their learning-related emotions, such as confusion, boredom, and frustration [185]. This information can serve as a proper indicator of presenters' effectiveness in delivering the ideas to the crowd. For instance, if confusion and frustration is detected from the crowd, feedback can be relayed to the presenter to reiterate or paraphrase what he/she said, or adopt strategies to remedy what caused confusion and frustration in the crowd. Self-reported crowd boredom coupled with presenter's vocal features analyzed by the machine can also be used as a proxy for ''energy'' or enthusiasm, and relevant feedback can be transmitted to the presenter to change his/her presentation mode, energy level, etc. on the fly.

IX. CONCLUSIONS

In this paper, we outlined the technological and operational components of a future emotionally-aware AI smart classroom that delivers automated real-time feedback through two modalities of an Open Learner Model to a presenter during a presentation in order to improve the effectiveness of the presentation, presenter's self-regulation and metacognitive awareness, and their verbal and non-verbal communication skills. The foundations of the proposed system are based on prominent developments, theories, and empirical studies in the fields of engineering and education. The system uses state-of-the-art algorithms, such as deep learning, high-performance GPU computing, multimodal sensing, and emotion recognition, which analyze multimodal presenter audio and visual information to extract the body language, intonation, and hand gesture information of the presenter. Simultaneously, the system receives scores from crowd listeners to determine the quality of a presentation.

We explored two main hypotheses in this paper: first, that input from presenter and listeners can be quantified, and second, that the relationship between presenter's behavior and how it is perceived by the audience can be learned and be used for shaping and providing real-time feedback by the system. Built on these hypotheses, the system works in two operational modes: In *Training Mode*, by making associations between the sensed multimodal data and the received human scores, the system learns how the crowd assesses a

presentation. In *Presentation Mode*, this learned data is used during a presentation to provide real-time feedback to a presenter, through either haptic gloves or simple emojis displayed on a computer screen. The eventual goal of the system is to improve the scores a presenter receives from the crowd.

Contributions of this paper are manifold. In terms of engineering, we bridge the gap in the existing literature on automated feedback systems by including quantified human input in machine analyses of a presenter's verbal and nonverbal behavior. We further the field of education at several fronts by integrating *real-time and adaptive* AI feedback in a classroom setting, advancing research on feedback modalities, integrating cognitive load and compliance in tailoring real-time feedback, and opening doors to human-machine tutoring of oral presentations, where anxiety is mitigated by removing human evaluators (in *presentation mode*) from the feedback equation. Furthermore, we proposed technical and educational evaluation metrics that will be used to assess the effectiveness and feasibility of the system. This project offers great applicability beyond the proposed education domain, in areas such as patient-doctor encounters in healthcare, or any training situation where humans interact.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] R. Hincks and J. Edlund, "Promoting increased pitch variation in oral presentations with transient visual feedback," *Lang. Learn. Technol.*, vol. 13, no. 3, pp. 32–50, 2009.

[2] A. Mehrabian, *Nonverbal Communication*. Piscataway, NJ, USA: Transaction Publishers, 1972.

[3] A. Mehrabian, *Silent Messages*, vol. 8. Belmont, CA, USA: Wadsworth, 1971.

[4] A. Mehrabian, *Silent Messages: Implicit Communication of Emotion and Attitude*. Belmont, CA, USA: Wadsworth, 1981.

[5] J. Elliot and L. Joyce. (2005). *Presentation Anxiety: A Challenge for Some Students and a Pit of Despair for Others*. [Online]. Available: http://www.isana.org.au/files/20051017165939_PresentationAnxiety.pdf

[6] P. Blikstein, "Multimodal lsearning analytics," in *Proc. 3rd Int. Conf. Learn. Anal. Knowl.-LAK*, Mar. 2013, pp. 102–106. [Online]. Available: http://dl.acm.org.myaccess.library.utoronto.ca/citation.cfm?id=2460296.2460316

[7] A. Andrade and J. A. Danish, "Using multimodal learning analytics to model student behaviour: A systematic analysis of behavioural framing," *J. Learn. Anal.*, vol. 3, no. 2, pp. 282–306, 2016.

[8] L. Chen *et al.*, "Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm," in *Proc. ACM*, Nov. 2016, pp. 161–168.

[9] L. Chen, G. Feng, and I. Bejar, "Towards assessing communicative competence using multimodal learning analytics," in *Proc. Mach. Learn. Digit. Edu. Assessment Syst.*, Oct. 2016, pp. 1–9. [Online]. Available: http://medianetlab.ee.ucla.edu/papers/ICML_chens.pdf

[10] T. Soyata, *Enabling Real-Time Mobile Cloud Computing Through Emerging Technologies*. Hershey, PA, USA: IGI Global, Aug. 2015.

[11] D. Kahneman, *Thinking, Fast and Slow*. New York, NY, USA: Macmillan, 2011.

[12] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Edu. Psychol.*, vol. 38, no. 1, pp. 53–61, 2003.

[13] M. Pouryazdan, C. Fiandrino, B. Kantarci, D. Kliazovich, T. Soyata, and P. Bouvry, "Game-theoretic recruitment of sensing service providers for trustworthy cloud-centric Internet-of-Things (IoT) applications," in *Proc. Globecom Workshops (GC Wkshps)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[14] M. Pouryazdan, B. Kantarci, T. Soyata, and H. Song, "Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing," *IEEE Access*, vol. 4, pp. 529–541, 2016.

[15] M. Pouryazdan, B. Kantarci, T. Soyata, L. Foschini, and H. Song, "Quantifying user reputation scores, data trustworthiness, and user incentives in mobile crowd-sensing," *IEEE Access*, vol. 5, pp. 1382–1397, Jan. 2017.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[17] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[18] G. Littlewort *et al.*, "The computer expression recognition toolbox (CERT)," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, Mar. 2011, pp. 298–305.

[19] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[20] Nvidia. *DIGITS Interactive Deep Learning GPU Training System*. Accessed: Jan. 17, 2018. [Online]. Available: https://developer.nvidia.com/digits

[21] Q. Liu, Z. Li, J. Lui, and J. Cheng, "Powerwalk: Scalable personalized pagerank via random walks with vertex-centric decomposition," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 195–204.

[22] T. Soyata, R. Muraleedharan, C. Funai, M. Kwon, and W. Heinzelman, "Cloud-vision: Real-time face recognition using a mobile-cloudlet-cloud acceleration architecture," in *Proc. 17th IEEE Symp. Comput. Commun. (ISCC)*, Cappadocia, Turkey, Jul. 2012, pp. 59–66.

[23] N. Powers and T. Soyata, "AXaaS (acceleration as a service): Can the telecom service provider rent a cloudlet?" in *Proc. 4th IEEE Int. Conf. Cloud Netw. (CNET)*, Niagara Falls, ON, Canada, Oct. 2015, pp. 232–238.

[24] J. A. Bellanca, *21st Century Skills: Rethinking How Students Learn*. Bloomington, IN, USA: Solution Tree, 2011.

[25] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture," *Instruct. Sci.*, vol. 32, nos. 1–2, pp. 1–8, 2004.

[26] D. Magin and P. Helmore, "Peer and teacher assessments of oral presentation skills: how reliable are they?" *Stud. Higher Edu.*, vol. 26, no. 3, pp. 287–298, 2001.

[27] T. Hovardas, O. E. Tsivitanidou, and Z. C. Zacharia, "Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students," *Comput. Edu.*, vol. 71, pp. 133–152, Feb. 2014.

[28] L. De Grez, M. Valcke, and I. Roozen, "How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments?" *Active Learn. Higher Edu.*, vol. 13, no. 2, pp. 129–142, 2012.

[29] D. L. Butler and P. H. Winne, "Feedback and self-regulated learning: A theoretical synthesis," *Rev. Edu. Res.*, vol. 65, no. 3, pp. 245–281, 1995.

[30] P. H. Winne and A. F. Hadwin, "nStudy: Tracing and supporting self-regulated learning in the Internet," in *International Handbook of Metacognition and Learning Technologies*. New York, NY, USA: Springer, 2013, pp. 293–308.

[31] R. Azevedo, "Multimedia learning of metacognitive strategies," *The Cambridge Handbook of Multimedia Learnings*. Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 647–672.

[32] B. J. Zimmerman, "Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects," *Amer. Edu. Res. J.*, vol. 45, no. 1, pp. 166–183, 2008.

[33] P. H. Winne and A. F. Hadwin, "The weave of motivation and self-regulated learning," *Motivation and Self-Regulated Learning: Theory, Research, and Applications*. New York, NY, USA: Lawrence Erlbaum Associates, 2008, pp. 297–314.

[34] R. Feyzi-Behnagh, R. Azevedo, F. Bouchet, and Y. Tian, "The role of an open learner model and immediate feedback on metacognitive calibration in metatutor," in *Proc. Annu. Meeting Amer. Edu. Res. Assoc.*, 2016.

[35] R. Feyzi-Behnagh, Z. Khezri, and R. Azevedo, "An investigation of accuracy of metacognitive judgments during learning with an intelligent multi-agent hypermedia environment," in *Proc. 33rd Annu. Conf. Cognit. Sci. Soc.*, 2011, pp. 96–101.

[36] R. Feyzi-Behnagh, R. Azevedo, E. Legowski, K. Reitmeyer, E. Tseytlin, and R. S. Crowley, "Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system," *Instruct. Sci.*, vol. 42, no. 2, pp. 159–181, 2014.

[37] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2nd Quart., 2010.
[38] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Newbury Park, CA, USA: Sage, 2013.
[39] M. Pouryazdan, C. Fiandrino, B. Kantarci, T. Soyata, D. Kliazovich, and P. Bouvry, "Intelligent gaming for mobile crowd-sensing participants to acquire trustworthy big data in the Internet of Things," *IEEE Access*, vol. 5, no. 1, pp. 22209–22223, Dec. 2017.
[40] M. Habibzadeh, Z. Qin, T. Soyata, and B. Kantarci, "Large scale distributed dedicated- and non-dedicated smart city sensing systems," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7649–7658, Dec. 2017.
[41] M. Habibzadeh, A. Boggio-Dandry, Z. Qin, T. Soyata, B. Kantarci, and H. and Mouftah, "Soft sensing in smart cities: Handling 3Vs using recommender systems, machine intelligence, and data analytics," *IEEE Commun. Mag.*, to be published.
[42] M. Hassanalieragh *et al.*, "Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, New York, NY, USA, Jun. 2015, pp. 285–292.
[43] A. Page, S. Hijazi, D. Askan, B. Kantarci, and T. Soyata, "Research directions in cloud-based decision support systems for health monitoring using Internet-of-Things driven data acquisition," *Int. J. Services Comput.*, vol. 4, no. 4, pp. 18–34, 2016.
[44] G. Honan, A. Page, O. Kocabas, T. Soyata, and B. Kantarci, "Internet-of-everything oriented implementation of secure digital health (D-health) systems," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Messina, Italy, Jun. 2016, pp. 718–725.
[45] M. C. Scheeler, J. K. McAfee, K. L. Ruhl, and D. L. Lee, "Effects of corrective feedback delivered via wireless technology on preservice teacher performance and student behavior," *Teacher Edu. Special Edu., J. Teacher Edu. Division Council Exceptional Children*, vol. 29, no. 1, pp. 12–25, 2006.
[46] M. C. Scheeler, M. Congdon, and S. Stansbery, "Providing immediate feedback to co-teachers through bug-in-ear technology: An effective method of peer coaching in inclusion classrooms," *Teacher Edu. Special Edu., J. Teacher Edu. Division Council Exceptional Children*, vol. 33, no. 1, pp. 83–96, 2010.
[47] I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. André, "Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 565–574.
[48] M. C. Scheeler, M. Macluckie, and K. Albright, "Effects of immediate feedback delivered by peer tutors on the oral presentation skills of adolescents with learning disabilities," *Remedial Special Edu.*, vol. 31, no. 2, pp. 77–86, 2010.
[49] F. Dermody, "Multimodal positive computing system for public speaking with real-time feedback," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, s pp. 541–545.
[50] J. Schneider, D. Börner, P. van Rosmalen, and M. Specht, "Can you help me with my pitch? Studying a tool for real-time automated feedback," *IEEE Trans. Learn. Technol.*, vol. 9, no. 4, pp. 318–327, Oct. 2016.
[51] S. Bull and J. Kay, "SMILI: A framework for interfaces to learning data in open learner models, learning analytics and related fields," *Int. J. Artif. Intell. Edu.*, vol. 26, no. 1, pp. 293–331, 2016.
[52] S. Bull and J. Kay, "Open learner models as drivers for metacognitive processes," in *International Handbook of Metacognition and Learning Technologies*. New York, NY, USA: Springer, 2013, pp. 349–365.
[53] S. Bull, "Supporting learning with open learner models," *Planning*, vol. 29, no. 14, p. 1, 2004.
[54] F. Lazarinis and S. Retalis, "Analyze me: Open learner model in an adaptive Web testing system," *Int. J. Artif. Intell. Edu.*, vol. 17, no. 3, pp. 255–271, 2007.
[55] D. Zapata-Rivera, E. Hansen, V. J. Shute, J. S. Underwood, and M. Bauer, "Evidence-based approach to interacting with open student models," *Int. J. Artif. Intell. Edu.*, vol. 17, no. 3, pp. 273–303, 2007.
[56] S. Bull and H. G. Pain, "Did I say what i think i said, and do you agree with me?': Inspecting and questioning the student model," Ph.D. dissertation, Dept. Artif. Intell., Univ. Edinburgh, Edinburgh, U.K., 1995.
[57] V. Dimitrova, J. Self, and P. Brna, *Involving the Learner in Diagnosis: Potentials and Problems*. Leeds, U.K.: Citeseer, 2001.
[58] J. Kay, Z. Halin, T. Ottomann, and Z. Razak, "Learner know thyself: Student models to give learner control and responsibility," in *Proc. Int. Conf. Comput. Edu.*, 1997, pp. 17–24.
[59] A. Mitrovic and B. Martin, "Evaluating the effects of open student models on learning," in *Proc. Int. Conf. Adapt. Hypermedia Adapt. Web-Based Syst.*, 2002, pp. 296–305.
[60] Y. Long and V. Aleven, "Supporting students' self-regulated learning with an open learner model in a linear equation tutor," in *Proc. Int. Conf. Artif. Intell. Edu.*, 2013, pp. 219–228.
[61] R. M. Maldonado, J. Kay, K. Yacef, and B. Schwendimann, "An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2012, pp. 482–492.
[62] A. T. Corbett, K. R. Koedinger, and J. R. Anderson, "Intelligent tutoring systems," in *Handbook of Human-Computer Interaction*, vol. 5. New York, NY, USA: Elsevier Science Publishers, 1997, pp. 849–874.
[63] S. Bull, M. Mangat, A. Mabbott, A. S. A. Issa, and J. Marsh, "Reactions to inspectable learner models: Seven year olds to university students," in *Proc. Workshop Learner Modelling Reflection, Int. Conf. Artif. Intell. Edu.*, 2005, pp. 1–10.
[64] I. Damian, T. Baur, and E. André, "Measuring the impact of multimodal behavioural feedback loops on social interactions," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 201–208.
[65] S. Tanimoto, "Dimensions of transparency in open learner models," in *Proc. 12th Int. Conf. Artif. Intell. Edu.*, 2005, pp. 100–106.
[66] C.-Y. Law, J. Grundy, R. Vasa, and A. Cain, "An empirical study of user perceived usefulness and preference of open learner model visualisations," in *Proc. IEEE Symp. Vis. Lang. Hum.-Centric Comput. (VL/HCC)*, Sep. 2016, pp. 49–53.
[67] M. Fung, Y. Jin, R. Zhao, and M. E. Hoque, "ROC speak: Semi-automated personalized feedback on nonverbal behavior from recorded videos," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 1167–1178.
[68] J. J. Van Merrienboer and J. Sweller, "Cognitive load theory and complex learning: Recent developments and future directions," *Edu. Psychol. Rev.*, vol. 17, no. 2, pp. 147–177, 2005.
[69] S. Kalyuga, P. Chandler, and J. Sweller, "Managing split-attention and redundancy in multimedia instruction," *Appl. Cognit. Psychol.*, vol. 13, no. 4, pp. 351–371, 1999.
[70] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Edu. Psychol.*, vol. 38, no. 1, pp. 43–52, 2003.
[71] S. Y. Mousavi, R. Low, and J. Sweller, "Reducing cognitive load by mixing auditory and visual presentation modes," *J. Edu. Psychol.*, vol. 87, no. 2, p. 319, 1995.
[72] S. Kalyuga, "Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective," *Edu. Psychol. Rev.*, vol. 19, no. 3, pp. 387–399, 2007.
[73] M. Jalali, A. Bouyer, B. Arasteh, and M. Moloudi, "The effect of cloud computing technology in personalization and education improvements and its challenges," *Procedia-Social Behav. Sci.*, vol. 83, pp. 655–658, Jul. 2013.
[74] J. R. Anderson, C. F. Boyle, and B. J. Reiser, "Intelligent tutoring systems," *Sci. (Washington)*, vol. 228, no. 4698, pp. 456–462, 1985.
[75] V. J. Shute and J. Psotka, "Intelligent tutoring systems: Past, present, and future," Armstrong Lab Brooks AFB TX Human Resources Directorate, Dublin, OH, USA, Tech. Rep., 1994.
[76] J. Leppink, F. Paas, C. P. M. Van der Vleuten, T. Van Gog, and J. J. G. Van Merriënboer, "Development of an instrument for measuring different types of cognitive load," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1058–1072, 2013.
[77] J. Beatty and B. Lucero-Wagoner, "The pupillary system," in *Handbook of Psychophysiology*, vol. 2. Cambridge, U.K.: Cambridge Univ. Press, 2000, pp. 142–162.
[78] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Edu. Psychol.*, vol. 38, no. 1, pp. 63–71, 2003.
[79] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychol. Bull.*, vol. 91, no. 2, p. 276, 1982.
[80] J. Hyönä, J. Tommola, and A.-M. Alaja, "Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks," *Quart. J. Experim. Psychol.*, vol. 48, no. 3, pp. 598–612, 1995.
[81] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.
[82] J. Klingner, B. Tversky, and P. Hanrahan, "Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks," *Psychophysiology*, vol. 48, no. 3, pp. 323–332, 2011.

[83] P. W. Van Gerven, F. Paas, J. J. Van Merriënboer, and H. G. Schmidt, "Memory load and the cognitive pupillary response in aging," *Psychophysiology*, vol. 41, no. 2, pp. 167–174, 2004.

[84] E. H. Hess and J. M. Polt, "Pupil size as related to interest value of visual stimuli," *Science*, vol. 132, no. 3423, pp. 349–350, 1960.

[85] D. C. Aboyoun and J. M. Dabbs, "The Hess pupil dilation findings: Sex or novelty?," *Social Behav. Personality, Int. J.*, vol. 26, no. 4, pp. 415–419, 1998.

[86] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1, pp. 185–198, 2003.

[87] T. van Gog and H. Jarodzka, "Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments," in *International Handbook of Metacognition and Learning Technologies*. New York, NY, USA: Springer, 2013, pp. 143–156.

[88] K. Rayner, "Eye movements and attention in reading, scene perception, and visual search," *Quart. J. Experim. Psychol.*, vol. 62, no. 8, pp. 1457–1506, 2009.

[89] Y. Kim and E. M. Provost, "Emotion recognition during speech using dynamics of multiple regions of the face," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 1, pp. 25:1–25:23, Oct. 2015. [Online]. Available: http://doi.acm.org/10.1145/2808204

[90] S. Gaulin and D. McBurney, *Psychology: An Evolutionary Approach*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[91] A. Ortony, G. L. Clore, and A. M. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[92] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, 2011, pp. 827–834.

[93] D. Grandjean, D. Sander, and K. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness Cognit.*, vol. 17, no. 2, p. 484, 2008.

[94] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.

[95] H. Schlosberg, "Three dimensions of emotion," *Psychol. Rev.*, vol. 61, no. 2, p. 81, 1954.

[96] R. P. Abelson and V. Sermat, "Multidimensional scaling of facial expressions," *J. Experim. Psychol.*, vol. 63, no. 6, p. 546, 1962.

[97] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[98] B. Schuller *et al.*, "Paralinguistics in speech and language—State-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4–39, 2013.

[99] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[100] E. Mower, M. J. Mataric, and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 843–855, Aug. 2009.

[101] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9, pp. 1062–1087, 2011.

[102] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image Vis. Comput.*, vol. 31, no. 2, pp. 137–152, 2013.

[103] S. Wan and J. K. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," *Pattern Recognit.*, vol. 47, no. 5, pp. 1859–1868, 2014.

[104] M. Pantic and M. S. Bartlett, *Machine Analysis of Facial Expressions*. London, U.K.: InTech, 2007.

[105] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 15–33, Jan. 2013.

[106] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.

[107] M. S. Hussain, S. K. D'Mello, and R. A. Calvo, "25 research and development tools in affective computing," in *The Oxford Handbook of Affective Computing*. Oxford, U.K.: Oxford Univ. Press, 2014, p. 349.

[108] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar. 2013.

[109] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[110] B. Schuller *et al.* (2013). *The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism*. [Online]. Available: http://eprints.gla.ac.uk/93665/

[111] B. Sapp and B. Taskar, "Multimodal decomposable models for human pose estimation," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3674–3681.

[112] S. Hijazi, A. Page, B. Kantarci, and T. Soyata, "Machine learning in cardiac health monitoring and decision support," *IEEE Comput. Mag.*, vol. 49, no. 11, pp. 38–48, Nov. 2016.

[113] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," *Depression*, vol. 1, p. 1, 2014.

[114] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proc. ACM Int. Conf. Multimodal Interact.*, Santa Monica, CA, USA, Oct. 2012, pp. 485–492.

[115] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1. Aug. 2006, pp. 1136–1139.

[116] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 3687–3691.

[117] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, 2015, pp. 73–80.

[118] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 7–13, Jan. 2012.

[119] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[120] G. S. V. S. Sivaram and H. Hermansky, "Sparse multilayer perceptron for phoneme recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 23–29, Jan. 2012.

[121] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems*, vol. 19. Cambridge, MA, USA: MIT Press, 2007, p. 153.

[122] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[123] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20. 2008, pp. 873–880.

[124] Y. Kim and E. M. Provost, "Leveraging inter-rater agreement for audio-visual emotion recognition," in *Proc. Affect. Comput. Intell. Interact. (ACII)*, Xi'an, China, Sep. 2015, pp. 553–559.

[125] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5688–5691.

[126] R. Brückner and B. Schuller, "Likability classification—A not so deep neural network approach," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, Sep. 2012, p. 4.

[127] B. Schuller *et al.*, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, Sep. 2012, p. 4.

[128] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[129] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact. (ACII)*, 2013, pp. 511–516.

[130] T. Soyata *et al.*, "COMBAT: Mobile Cloud-based cOmpute/coMmunications infrastructure for BATtlefield applications," *Proc. SPIE*, vol. 8403, p. 84030K, May 2012.

[131] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proc. Spring Joint Comput. Conf.*, Apr. 1967, pp. 483–485.
[132] J. L. Gustafson, "Reevaluating Amdahl's law," *Commun. ACM*, vol. 31, no. 5, pp. 532–533, 1988.
[133] M. D. Hill and M. R. Marty, "Amdahl's law in the multicore era," *Computer*, vol. 41, no. 7, pp. 33–38, Jul. 2008.
[134] Stanford University. *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*. Accessed: Jan. 17, 2018. [Online]. Available: http://swww.image-net.org/
[135] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[136] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[137] T. Soyata, *GPU Parallel Program Development Using CUDA*. New York, NY, USA: Taylor & Francis, Feb. 2018.
[138] T. Soyata, H. Ba, W. Heinzelman, M. Kwon, and J. Shi, "Accelerating mobile cloud computing: A survey," in *Communication Infrastructures for Cloud Computing*, H. T. Mouftah and B. Kantarci, Eds. Hershey, PA, USA: IGI Global, Sep. 2013, ch. 8, pp. 175–197.
[139] E. Cuervo *et al.*, "MAUI: Making smartphones last longer with code offload," in *Proc. 8th Int. Conf. Mobile Syst., Appl., Services*, 2010, pp. 49–62.
[140] B. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: Elastic execution between mobile device and cloud," in *Proc. 6th Conf. Comput. Syst.*, 2011, pp. 301–314.
[141] N. Powers *et al.*, "The cloudlet accelerator: Bringing mobile-cloud face recognition into real-time," in *Proc. Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–7.
[142] Y. Song, H. Wang, and T. Soyata, "Theoretical foundation and GPU implementation of face recognition," in *Enabling Real-Time Mobile Cloud Computing Through Emerging Technologies*. Hershey, PA, USA: IGI Global, 2015, ch. 11, pp. 322–341.
[143] Amazon Web Services. *Elastic Compute Cloud (Amazon EC2)*. Accessed: Jan. 17, 2018. [Online]. Available: https://aws.amazon.com/ec2
[144] N. Powers, A. Alling, R. Gyampoh-Vidogah, and T. Soyata, "AXaaS: Case for acceleration as a service," in *Proc. Globecom Workshops (GC Wkshps)*, Austin, TX, USA, Dec. 2014, pp. 117–121.
[145] N. Powers and T. Soyata, "Selling FLOPs: Telecom service providers can rent a cloudlet via acceleration as a service (AXaaS)," in *Enabling Real-Time Mobile Cloud Computing Through Emerging Technologies*. Hershey, PA, USA: IGI Global, 2015, ch. 6, pp. 182–212.
[146] *Verizon Terremark*. [Online]. Available: http://www.terremark.com/
[147] H. Ba, W. Heinzelman, C.-A. Janssen, and J. Shi, "Mobile computing—A green computing resource," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 4451–4456.
[148] C. Funai, C. Tapparello, H. Ba, B. Karaoglu, and W. Heinzelman, "Extending volunteer computing through mobile ad hoc networking," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 32–38.
[149] O. Kocabas, T. Soyata, and M. K. Aktas, "Emerging security mechanisms for medical cyber physical systems," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 3, pp. 401–416, Jun. 2016.
[150] O. Kocabas and T. Soyata, "Utilizing homomorphic encryption to implement secure and private medical cloud computing," in *Proc. IEEE 8th Int. Conf. Cloud Comput. (CLOUD)*, New York, NY, USA, Jun. 2015, pp. 540–547.
[151] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. STOC*, 2009, pp. 169–178.
[152] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in *Proc. EUROCRYPT*, 2011, pp. 129–148.
[153] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. EUROCRYPT*, 1999, pp. 223–238.
[154] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proc. 3rd ACM Workshop Cloud Comput. Secur. Workshop*, 2011, pp. 113–124.
[155] O. Kocabas, T. Soyata, J.-P. Couderc, M. Aktas, J. Xia, and M. Huang, "Assessment of cloud-based health monitoring using homomorphic encryption," in *Proc. IEEE 31st Int. Conf. Comput. Design (ICCD)*, Oct. 2013, pp. 443–446.
[156] O. Kocabas, R. Gyampoh-Vidogah, and T. Soyata, "Operational cost of running real-time mobile cloud applications," in *Enabling Real-Time Mobile Cloud Computing Through Emerging Technologies*. Hershey, PA, USA: IGI Global, 2015, ch. 10, pp. 294–321.
[157] NIST, "Advanced encryption standard (AES)," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. fIPS-197, Nov. 2001.
[158] S. van Ginkel, J. Gulikers, H. Biemans, and M. Mulder, "Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education," *Edu. Res. Rev.*, vol. 14, pp. 62–80, Feb. 2015.
[159] P. Patel, "Using formative assessment to improve presentation skills," *Voices From Middle*, vol. 22, no. 1, p. 22, 2014.
[160] M. C. Scheeler, K. L. Ruhl, and J. K. McAfee, "Providing performance feedback to teachers: A review," *Teacher Edu. Special Edu., J. Teacher Edu. Division Council Exceptional Children*, vol. 27, no. 4, pp. 396–407, 2004.
[161] L. De Grez, M. Valcke, and I. Roozen, "The impact of an innovative instructional intervention on the acquisition of oral presentation skills in higher education," *Comput. Edu.*, vol. 53, no. 1, pp. 112–120, 2009.
[162] M. Patri, "The influence of peer feedback on self-and peer-assessment of oral skills," *Lang. Test.*, vol. 19, no. 2, pp. 109–131, 2002.
[163] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[164] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4277–4280.
[165] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.
[166] T. Baur *et al.*, "Context-aware automated analysis and annotation of social human–agent interactions," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 2, p. 11, 2015.
[167] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learn. Instruct.*, vol. 4, no. 4, pp. 295–312, 1994.
[168] R. S. Frackowiak, *Human Brain Function*. San Diego, CA, USA: Academic, 2004.
[169] A. Page, M. K. Aktas, T. Soyata, W. Zareba, and J. Couderc, "QT clock to improve detection of QT prolongation in long QT syndrome patients," *Heart Rhythm*, vol. 13, no. 1, pp. 190–198, Jan. 2016.
[170] A. Page, T. Soyata, J. Couderc, and M. K. Aktas, "An open source ECG clock generator for visualization of long-term cardiac monitoring data," *IEEE Access*, vol. 3, pp. 2704–2714, Dec. 2015.
[171] WIRED. *Apple is Bringing the AI Revolution to Your iPhone*. Accessed: Jan. 17, 2018. [Online]. Available: https://www.wired.com/2016/06/apple-bringing-ai-revolution-iphone/
[172] C. A. Reitmeier and D. A. Vrchota, "Self-assessment of oral communication presentations in food science and nutrition," *J. Food Sci. Edu.*, vol. 8, no. 4, pp. 88–92, 2009.
[173] V.-W. Mitchell and C. Bakewell, "Learning without doing: Enhancing oral presentation skills through peer review," *Manage. Learn.*, vol. 26, no. 3, pp. 353–366, 1995.
[174] B. Libet, C. A. Gleason, E. W. Wright, and D. K. Pearl, "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential)," *Brain*, vol. 106, no. 3, pp. 623–642, 1983.
[175] S. Blackmore, *Consciousness: An Introduction*. Evanston, IL, USA: Routledge, 2013.
[176] B. Libet and S. M. Kosslyn, *Mind Time: The Temporal Factor in Consciousness*. Cambridge, MA, USA: Harvard Univ. Press, 2009.
[177] J. Konorski, "Learning, perception, and the brain. (Book reviews: Integrative activity of the brain. An interdisciplinary approach)," *Science*, vol. 160, no. 3823, pp. 652–653, 1968.
[178] M. Kwon, Z. Dou, W. Heinzelman, T. Soyata, H. Ba, and J. Shi, "Use of network latency profiling and redundancy for cloud server selection," in *Proc. IEEE 7th Int. Conf. Cloud Comput. (CLOUD)*, Jun. 2014, pp. 826–832.
[179] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, pp. 341–345, Sep. 2002.
[180] J. G. Milton and A. Longtin, "Evaluation of pupil constriction and dilation from cycling measurements," *Vis. Res.*, vol. 30, no. 4, pp. 515–525, 1990.
[181] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27-1–27-27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[182] INTEL. *Xeon E5-2680v4 Processor*. Accessed: Jan. 17, 2018. [Online]. Available: https://ark.intel.com/products/91754/Intel-Xeon-Processor-E5-2680-v4-35M-Cache-2_40-GHz

[183] Nvidia. *Tesla K80 GPU*. Accessed: Jan. 17, 2018. [Online]. Available: http://www.nvidia.com/object/tesla-k80.html
[184] W. Prokasy, *Electrodermal Activity in Psychological Research*. Amsterdam, The Netherlands: Elsevier, 2012.
[185] R. Pekrun, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," *Edu. Psychol. Rev.*, vol. 18, no. 4, pp. 315–341, 2006.

YELIN KIM (S'12–M'16) received the B.S. degree from Seoul National University, South Korea, and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, all in electrical and computer engineering. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University at Albany, SUNY. Her research is in the area of audio-visual emotion recognition and affective computing. Her research builds upon techniques from diverse areas, including machine learning, multimodal (speech and video) signal processing, computer vision, and behavioral science. She was a recipient of several awards and scholarships, including the SUNY-A Faculty Research Award, Korean Government Scholarship for Study Abroad, the Qualcomm Scholarship, the Korea National Science Scholarship, and the Temasek Foundation Scholarship. During her Ph.D. studies, she led a project that awarded the Best Student Paper at ACM Multimedia, 2014.

TOLGA SOYATA (M'08–SM'16) received the B.S. degree from Istanbul Technical University in 1988, the M.S. degree from Johns Hopkins University in 1992, and the Ph.D. degree from the University of Rochester in 2000, all in electrical and communications engineering. He joined the ECE Department, University of Rochester, in 2008. He was an Assistant Professor (Research) with the ECE Department, University of Rochester, when he left to join with the Department of ECE, University at Albany, SUNY, Albany, as an Associate Professor, in 2016. His teaching interests include CMOS VLSI ASIC Design, FPGA-based High Performance Data Processing System Design, and GPU Parallel Programming. His research interests include Cyber Physical Systems, Digital Health, and GPU-based high-performance computing. He is a Senior Member of ACM.

REZA FEYZI BEHNAGH received the Ph.D. degree in educational psychology, learning sciences from McGill University, Montreal, Canada. He joined University at Albany, SUNY, in 2014, where he is currently an Assistant Professor with the Department of Educational Theory and Practice. In terms of research, he is interested in how students learn and self-regulate their learning while learning about complex science topics with computer-based learning environments; how they plan and metacognitively monitor their learning and use effective learning strategies; improving learners' metacognitive judgments; learning from multimedia content; using process and product data in studying self-regulated learning (SRL); using eye-tracking data and log-file data in researching SRL; and studying the effectiveness of Open Learner Models in computerbased learning environments.