Scaling and Adopting a Multimodal Learning Analytics Application in an Institution-Wide Setting

Federico Domínguez, *Senior Member, IEEE*, Xavier Ochoa, Dick Zambrano, Katherine Camacho, and Jaime Castells

***Abstract*—Multimodal learning analytics, which is collection, analysis, and report of diverse learning traces to better understand and improve the learning process, has been producing a series of interesting prototypes to analyze learning activities that were previously hard to objectively evaluate. However, none of these prototypes have been taken out of the laboratory and integrated into real learning settings. This article is the first to propose, execute, and evaluate a process to scale and deploy one of these applications, an automated oral presentation feedback system, into an institution-wide setting. Technological, logistical, and pedagogical challenges and adaptations are discussed. An evaluation of the use and effectiveness of the deployment shows both successful adoption and moderate learning gains, especially for low-performing students. In addition, the recording and summarizing of the perception of both instructors and students point to a generally positive experience in spite of the common problems of a first-generation deployment of a complex learning technology.**

***Index Terms*—Automated feedback, multimodal, oral presentation skills.**

I. Introduction

Multimodal learning analytics (MmLA) focuses on the collection and analysis of diverse traces obtained from different aspects of learning processes for better understanding and improving those processes [1]. For example, in a traditional classroom setting, those traces could be the direction of the gaze and facial expression of the students, the posture and tone of voice of the instructor, the words in the question made by the instructor, or the words being written by the students in their notebooks. To be able to capture those traces, multimedia recordings should be made (e.g., audio, video, digital pens, and online logs). The previously mentioned traces or features are then extracted from these recordings. These features are in turn combined and fused to calculate objectively measurable indicators (e.g., turn-taking or joint visual attention). These indicators are used as pieces of evidence to estimate learning-related high-level constructs (e.g., collaboration quality, expertise, and engagement). Finally, these estimations are used to test, modify, or create learning theories or more importantly to provide feedback reports to students or instructors to help them reflect about the learning process. This methodology is not new; it has been the basis for observation-based quantitative educational research [2]. What is new is the affordances provided by low-cost high-definition sensors that enable the capture of the traces with a level of detail that was not feasible before and the analytic capabilities of machine learning and artificial intelligence (AI) algorithms that provide a level of automation that enable MmLA to scale to large amounts of raw data and to provide real-time feedback [3].

MmLA has been applied to study and to provide feedback in several learning contexts: collaboratively problem solving [4], [5], classroom interactions [6], [7], computer programming learning [8], [9], and even martial arts learning [10]. These studies concentrate on three different aspects: the use of affordable, synchronized, multimodal sensors to capture relevant individual and social signals during learning activities (see, e.g., [11]); the analysis and fusion of these signals to estimate learning-relevant constructs (see, e.g., [12] and [13]); and the use of natural and multimodal interfaces to provide useful feedback to the activity's participants based on the results of the analysis (see, e.g., [14] and [15]). These studies, however, have been conducted mainly with prototyping tools in a laboratory or controlled environment. To the best of the authors' knowledge, after an exhaustive literature search, there is no scientific publication that describes how an MmLA system can be successfully scaled and adopted in large real learning scenarios.

This article describes the design, execution, and evaluation of a process to scale and deploy a multimodal oral presentation automatic feedback system (RAP for its Spanish acronym) [15] in a mid-size polytechnic higher education institution in the coast of Ecuador. Additionally, it discusses how this system was successfully integrated and used into existing pedagogical

Manuscript received July 3, 2020; revised May 24, 2021; accepted July 18, 2021. Date of publication July 29, 2021; date of current version September 3, 2021. This work was supported in part by the Escuela Superior Politécnica del Litoral and in part by the Secretariat of Higher Education, Science, Technology, and Innovation through the ANDAMIOS project under Grant PF-14-ESPOL-ANDAMIOS-002. *(Corresponding author: Federico Domínguez.)*

Federico Domínguez is with the Information Technology Center, and the Faculty of Electrical and Computer Engineering, Escuela Superior Politécnica del Litoral, Guayaquil 090902, Ecuador (e-mail: fexadomi@espol.edu.ec).

Xavier Ochoa is with the Department of Administration, Leadership, and Technology, Steinhardt School of Culture, Education, and Human Development, New York University, New York, NY 10003 USA (e-mail: xavier.ochoa@nyu.edu).

Dick Zambrano and Katherine Camacho are with the Faculty of Natural Sciences and Mathematics, Escuela Superior Politécnica del Litoral, Guayaquil 090902, Ecuador (e-mail: dzambra@espol.edu.ec; katmicam@espol.edu.ec).

Jaime Castells is with the Information Technology Center, Escuela Superior Politécnica del Litoral, Guayaquil 090902, Ecuador (e-mail: jaime.castells@cti.espol.edu.ec).

Digital Object Identifier 10.1109/TLT.2021.3100778

TABLE I
SUMMARY OF SIMILAR SYSTEMS FOUND IN THE LITERATURE REVIEW

Name [Citation]	Sensors	Modalities	Deployment
Presentation Sensei [16]	Camera, microphone	Pitch, filled pauses, speech rate, gaze	No
Cicero [17]	MS Kinect, cameras, microphone	Gaze, gestures, speech rate, volume, filled pauses, pitch	No
Logue [18]	MS Kinect, microphone	Speech rate, energy, openness	Yes (3 presenters)
— [19]	MS Kinect	Gaze, posture, gestures, filled pauses, pitch	No
— [20]	MS Kinect	Gaze, postures, gestures, movement	No
PresentMate [21]	Smart phone camera, microphone	Timing, body motion, volume	No
Presentation Trainer [22]	MS Kinect	Posture, gestures, speech rate	Yes (12 presenters)
Rhema [23]	Google Glass	Volume, speech rate	No
Automanner [24]	MS Kinect	Postures, gestures	No
RAP [15]	Camera, microphone	Gaze, posture, volume, filled pauses, slide quality	No
RoboCOP [25]	MS Kinect	Speech rate, filled pauses, pitch, gaze	No
Presentation Trainer VR [26]	Kinect, VR set	Posture, facial express., volume, gaze, filled pauses, movement	No

practices, the learning effect that it has on students, and qualitative and quantitative analysis of students' and instructors' attitudes toward the system and its automated feedback. The main contribution of this article to the learning technology community is to serve as an example of all the technological, pedagogical, and logistical adaptations needed to effectively convert MmLA tools from laboratory prototypes into institution-wide learning tools, and an evaluation of the impact that these systems could have in real-world learning scenarios.

The rest of this article is organized as follows. Section II presents a literature research of similar systems and how they have been deployed. Section III summarizes the most important aspects of the original RAP system, the initial prototype used to automatically assess oral presentations. Section IV discusses the logistical, technological, and pedagogical adaptations needed to scale the system and integrate it into a real institution-wide learning activity. Section V describes the actual deployment and usage of the RAP system during a semester, the effects the system had on the students' presentation skills, and the overall experience of usage in instructors and students. Section VI closes with lessons learned, conclusions, and recommendations for further work.

II. RELATED SYSTEMS

Providing feedback to students about their oral presentation skills has been one of the most active areas of MmLA [1]. The first examples of models to extract relevant features from multimodal recordings of presentations can be seen in the analysis of audio signals to estimate the "liveliness" of oral presentations [27], [28]. Already, these works propose a visual feedback mechanism to help students practice and develop their skills. Another seminal moment in the development of automatic analysis techniques for oral presentations was the result of the publication of two open datasets.

1) The Oral Presentation Quality Dataset [29] that captured 111 student groups presenting various topics related to a college course with roughly 19 hours of multimodal data (video, audio, skeleton joints, and digital slides) together with instructor-generated evaluation with a rubric.
2) The NUS Multisensor Presentation Dataset [30] that captured 51 student presentations including point-of-view cameras for the presenter and two members of the audience, skeleton joints of the presenter and audio, together with an expert graded rubric.

These datasets were analyzed by several researchers and resulted in improved algorithms to extract speech, gaze, movement, gestures, and digital slides features from the recordings [18], [31]–[35].

A nonsystematic but comprehensive review of the literature reveals that at least 12 multimodal systems to provide automatic feedback for oral presentations have been built (see Table I). The first fully multimodal system that incorporates a feedback mechanism was "Presentation Sensei" created by Kurihara *et al.* [16]. This system used a camera, a microphone, and a fiducial marker to extract the speed, tone, and filled pauses of the speed, the gaze, and the timing of the presenter. This system was only used in a laboratory setting with three presenters. A considerably more complex system, "Cicero," was developed by Batrinca *et al.* [17]. Apart from using more sophisticated sensors (a Microsoft Kinect and two cameras), it included a virtual interactive audience that provided visual and auditory feedback to the presenter in a similar way that real audiences do. The reported use of the Cicero system only took place in a laboratory with a small number of presenters (14). In 2015 and 2016, several independent but similar systems have been reported in the literature. "Logue," a system developed by Li *et al.* [18], used a wearable sensor (Google Glass) to obtain additional egocentric views from the presenter and members of the audience and to present the feedback in real time to the presenter about speech speed, energy of movement, and body openness. Logue was used in a real learning activity but only with three presenters. Dermody and Sutherland [19] focused on using only the Kinect to provide a feedback dashboard about gaze direction, body posture, gestures, speech tone, and disfluencies (filled pauses). This system was never deployed, even for laboratory evaluation. Nguyen *et al.* [20] also designed a system built around a Kinect that provided feedback in real time with an interactive virtual audience and offline through a dashboard about gaze, postures, gestures, and movement. The Nguyen *et al.* system was only deployed for laboratory evaluation with 11 presenters.

"PresentMate" [21] used the sensors of a mobile phone to create an application that provided instant or delayed feedback on timing, body motion, and voice level to the presenter. The report of the use of PresentMate describes only a laboratory experiment with 20 presenters. Schneider *et al.* [22] created and improved upon "Presentation Trainer," a Kinect-based system that provides automatic feedback of the presenter's posture, gestures, and speech cadence. In its several iterations, Presentation Trainer has been deployed in the wild but only used by a small number of presenters (12). "Rhema," the system designed by Tanveer *et al.* [23], uses only a Google Glass to analyze and provide feedback on the presenter's speech volume and speed. Rhema was only evaluated in a laboratory with 30 presenters. Tanveer *et al.* [24] also proposed "Automanner," a similar system based on Microsoft Kinect to provide feedback on the presenter's postures and gestures. This system was also evaluated in the laboratory with a small number of presenters (27). In recent years, new types of systems have been proposed that use new advances in sensor and AI technologies. Our system, "RAP" [15], replaces the Microsoft Kinect with a simple webcam and deep-learning-based computer vision algorithms to provide offline feedback to early-year student presenters. Initially, RAP was only deployed in the laboratory and tested with 83 presenters. "RoboCOP" [25] uses a humanoid robot head to provide verbal and nonverbal feedback to the presenter replicating advice from human experts. RoboCOP was only used in a laboratory with 30 presenters. Finally, a new version of the "Presentation Trainer" [26] uses virtual reality (VR) to provide feedback on gestures, postures, facial expressions, filled pauses, volume, and excessive movement. This version has not been deployed outside the laboratory.

None of the systems described before, except for Logue [18] and Presentation Trainer [22], have been deployed beyond a prototype stage, where they were mainly used for experimental and evaluation purposes with few presenters (less than 15).

Apart from the systems described in scientific literature, several patents have been obtained for systems that provide automatic feedback to presenters [36]–[43]. Despite all these patents, the only commercial product that is similar to the ones described in the scientific literature is the "Presenter Coach," a Microsoft PowerPoint plugin that provides feedback only on speech features such as pacing, inclusive language, filled pauses, culturally insensitive phrases, and if the presenter is reading from the slides. However, at the time of writing, there is no study or report about the use of this system.

The present work is the first time that one of these systems is deployed and evaluated at an institution-wide scale to be used in an authentic learning experience involving a cross section of all the students in that institution.

III. Description of the Original Prototype

For the convenience of the reader, this section will summarize the technical characteristics and evaluation of our first experimental version of the RAP system. A more detailed technical description and evaluation has been previously published [15].

The design principles that guided the development of the original RAP system were as follows.

1) The system should focus on specific use. It is not a general presentation trainer, but it was created to help entry-level students at higher education institutions.
2) The system should provide an immersive and unobtrusive experience. The presentation setting should be as similar as possible to real presentations without the need to wear any special equipment.
3) The system should be "Plug and Play." Any student and instructor should be able to use the system without the support or presence of researchers or technicians.
4) The feedback presented by the system should be objective, easy-to-understand, and unambiguous. The feedback should be provided after the presentation session in a persistent form (web page) and can be reviewed as many times as needed.

The original prototype of the RAP system was evaluated and able to fulfill all of these principles [15].

A. Components

The main component of the RAP system is the presentation room—known also as the RAP room—where users perform an oral presentation in front of a virtual audience. The hardware in the room captures the presentation through three modalities: audio, video, and slides. The room's layout is straightforward: the presentation slides and the virtual audience are shown in two screens opposite to one another. The video and audio of the presentation are captured using a camera and a microphone. The camera is camouflaged within the virtual audience screen, and the microphone is placed outside the field of view of the presenter. In this room, the presenter is being recorded for a fixed time (usually 5 min) or until their presentation slides reach the end.

B. Recording

The presentation's video is captured at 5 frames/s with a 1280 × 960 resolution using an 8-MP USB camera. Audio is captured at 44 kHz, 16 bits in one single channel using a directional microphone. The third modality, the slides, is captured before starting the presentation when the user inserts a USB drive that contains their Microsoft PowerPoint file. To extract the presentation features and generate the automatic feedback report, audio and video were captured by a workstation behind the audience screen. The slides are stored and processed after the presentation by the same workstation.

C. Feature Extraction

The RAP system extracts presentation features from the three modalities: video, audio, and the slides file. The system extracts posture and gaze from video, filled pauses and voice volume from audio, and slide quality from the slides file.

Using the computer vision library OpenPose [44], the video is analyzed to extract the skeletal joints of the presenter for each video frame. This information is used to estimate the presenter's pose. A posture is classified as BAD when a person has their hands in their pockets or behind their back, or if their hands or arms are held together (close posture). A GOOD posture is defined as having the hands in their upper part of the body held in an expressive gesture (open posture). During tests, this classification reached an accuracy of 0.80 [15].

The skeletal joints extracted with OpenPose are also to estimate the gaze direction. A trained random forest model was used to classify the gaze as GOOD or BAD for each frame, using the angles and normalized distances between the keypoints obtained from the face, such as the coordinates of the eyes, nose, ears, etc. A GOOD gaze is defined as the presenter looking at the audience. During tests, this classification reached an accuracy of 0.85 [15].

The PRAAT speech analysis library [45] is used to extract two features from the audio streaming: 1) voice volume; and 2) filled pauses. An automatically filled pause (FP) detection system was implemented taking as a reference the formant-based technique reported by Audhkhasi *et al.* [46]. As part of the processing, the entire audio segment is analyzed in small windows of 10 ms. For every window, the first and second formant values are extracted; their stability is analyzed over time and finally automatically tagged as filled pause or normal audio. During tests, this classification reached an accuracy of 0.87 [15].

For voice volume, volunteers were asked to classify the perceived volume of previously recorded audio segments as HIGH, LOW, or SILENCE; therefore, a threshold was established for automatic classification of volume. Similar to Filled Pause (FP) detection, the audio segments were divided into 10-ms windows, and the average volume level of each segment was obtained . The more time, i.e., the more segments classified as LOW led to an overall negative volume score, whereas the more segments classified as HIGH led to a positive volume score. During tests, this classification reached an accuracy of 0.87 [15].

For slide quality, three features are considered for this analysis: font size, slide contrast, and text quantity in each slide. These features were analyzed based on the findings of [33] and [47], which indicate that better presentations are achieved when there is less text, bigger font sizes are used, and the value of the slide contrast is close to 21. These features are combined into a single slide quality value. During tests, this classification reached an accuracy of 0.86 [15].

D. Feedback Report

After a presentation recording, the system calculates a five-level score for each one of the five features (posture, gaze, volume, filled pauses, and slide quality) depending on the percentage of correct instances. Using this information, the system then composes a feedback report that can be viewed by the presenter shortly after the presentation. Fig. 1 depicts an actual example of the generated feedback. The feedback contains a small evaluation of the presentation with a complete recording, an overall score, and individual scores for each one of the dimensions. Additionally, for each dimension, the report includes examples (pictures or audio clips) of the presenter's correct (e.g., looking at the audience) and incorrect (e.g., looking at the slides) moments during the presentation.

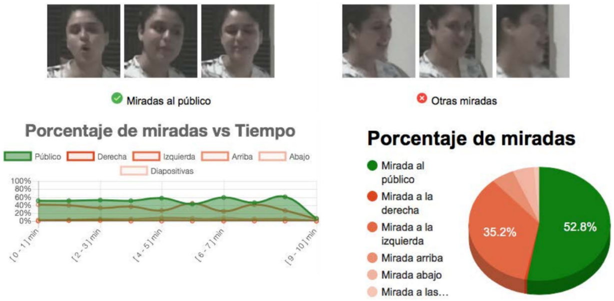


Fig. 1. Gaze direction section of the original feedback report [15].

E. Initial Evaluation

Our initial roll-out of the RAP system in 2017, as reported in [15], started with the participation of 83 computer science students from three courses using one RAP room for one semester. In this experiment, the technical capabilities of the system were validated. All of the features were automatically extracted with an accuracy of around 85% compared to ground truth coded by humans. Also, students reported an overall positive experience with the system.

The system described [15], at that point in time, was in a similar state than the dozens of similar systems presented in the related works section: a prototype with promising results, but with no real integration into existing formal learning activities. The following sections describe the steps taken to scale the system for institution-wide use and how instructors have incorporated it inside their learning designs.

IV. Scaling for Institution-Wide Use

The prototype RAP system and its results were presented to both professors and administrators at Escuela Superior Politécnica del Litoral, a mid-size polytechnic university in Guayaquil, Ecuador, where it was designed and initially tested. The interest expressed by the instructors of several courses and the expectations set by higher management made it clear that for the system to be useful, it should be scaled to serve thousands of students in a single semester, and it should be tightly and seamlessly integrated into existing pedagogical practices of different courses. This motivated a redesign to evolve the RAP system from a research prototype into an actual learning tool that could be used outside the laboratory at scale. The new design guides for the *scaled RAP system* were the following.

1) There should be several RAP rooms, geographically distributed among the campus, to manage the demand of thousands of recordings during the semester.
2) All RAP rooms should be able to operate concurrently.

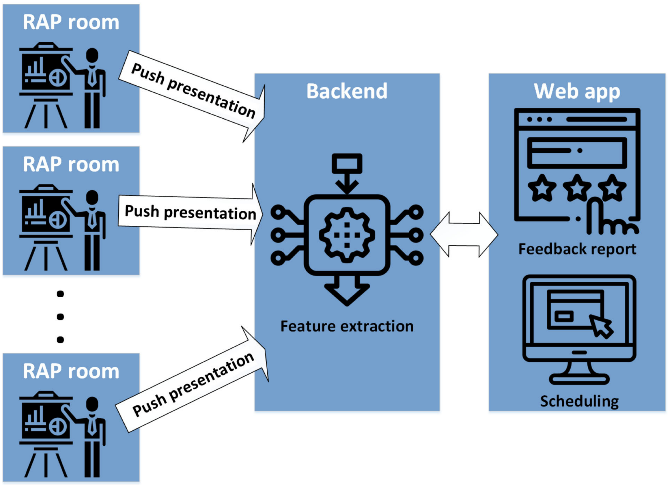


Fig. 2. Scaled RAP system components.

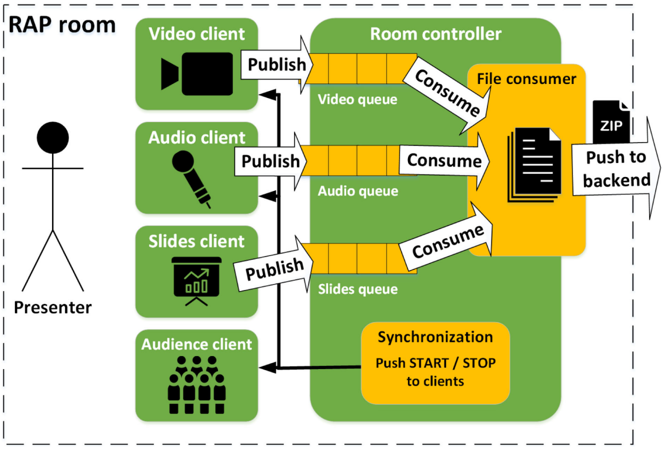


Fig. 3. RAP room components in the scaled RAP system.

3) The RAP room should be operated by the student without the need of a technician being present in the room.
4) The cost of the equipment and installation of a RAP room should be low.
5) The RAP report should go beyond just automatic feedback to accommodate also instructor-generated feedback.
6) The system should be robust enough to handle connectivity problems between the RAP rooms and the central processing without the need to stop their recording operations.
7) The data of the student recording should only be accessible to authorized users.

In this new *scaled RAP system*, the core idea of the system, the automatic feature extraction algorithms and feedback reports, remained unchanged; however, the actual data processing path and the logistics of physical resources were redesigned to respond to higher demand. Additionally, complementary systems were developed to facilitate better integration of the RAP system in course curricula. These adaptations could be classified into three types: technological, pedagogical, and logistical. This section describes them in detail.

A. *Technological Adaptations*

Laboratory prototypes are usually built with the purpose of pristine data acquisition and functional data processing, without much regard to features important for any system that will be deployed in the wild such as performance, robustness, adaptability, and cost. This subsection describes the technological adaptations made to implement the *scaled RAP system*.

1) Decouple Recording, Processing, and Reporting: The guidelines of the new design require that several RAP rooms work together in both a technically sound and economically viable way. It became necessary to decouple data recording that will still happen in each RAP room from the data processing and reporting that now will occur at a centralized backend. Also, given the processing-intensive task of extracting the features from video, audio, and presentations, this processing needs to be separated from more user-facing components, such as the report presentation application. Fig. 2 shows the high-level architecture of the scaled RAP system, where RAP rooms are removable modules that push their recordings asynchronously to the backend. The recordings are then processed by the feature extraction software that runs in servers optimized for heavy computational loads [high-performance central processor, supported by several graphic processing units (GPUs)]. Once the features are extracted, they are stored in a shared database to be consumed by the web applications that interact with students and instructors. These applications run in a server that is optimized for interactive queries (high-performance central processor, supported by large quantities of RAM memory and permanent storage). This design facilitates a more effective use of resources: processing power could be used for other tasks when not in use by the RAP system; RAP rooms only have the basic hardware needed for recording; each one of the components could be scaled independently according to demand.

Fig. 3 shows a more detailed view of the RAP recording room with its different data capture modules and their interconnections. The architecture uses the producer/consumer design pattern to decouple all modules, specifically employing the Message Queuing Telemetry Transport (MQTT) protocol, used commonly as a lightweight way to synchronize Internet of Things devices, to broker start/stop and status messages between modules. After a recording is finished, each capture module publishes its recording file to its respective message queue using the Advanced Message Queuing Protocol (AMQP). A File Consumer, subscribed to all queues, collects and packages all files from the finished recording, adds metadata, and pushes it as a single compressed presentation file to the backend using also a system-wide available MQTT message broker and an HTTP post message.

At the backend, a File processor module consumes all presentation files produced by all RAP rooms. When the File processor receives a presentation, it unpacks it and distributes its contents, i.e., the video, audio, and slides files, to their corresponding processing AMQP queues (see Fig. 4). This design enables load balancing of the data processing tasks by allowing several data processors to serve as consumers. Once a data

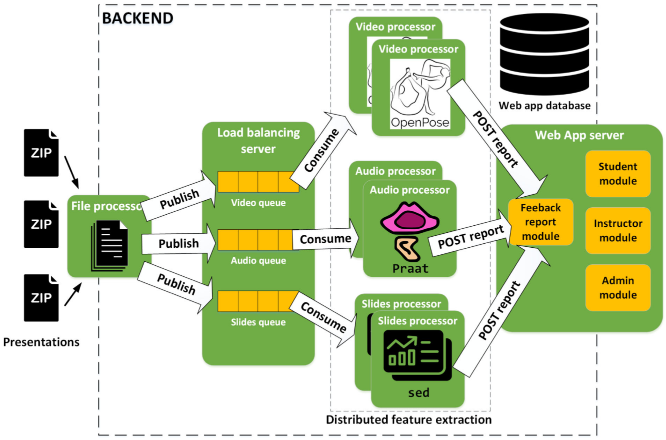


Fig. 4. Backend components in the scaled RAP system.

processor module finishes, it posts its results to the Web App server, using regular HTTP commands. The Feedback report module within the webserver generates the feedback report to be viewed by the student and the instructor.

It is important to note that the overall architecture of the system aims to decouple all submodules as much as possible to minimize single points of failure. For example, in the event of congestion/failure at the backend or a network blackout, the room is still able to perform recordings as it will simply enqueue presentations until the backend or the network recovers. This minimizes appointment cancellations during system failures.

2) Load Balancing: The feature extraction algorithms used in the RAP system are computation intensive. The OpenPose library uses convolutional deep neural networks to extract the skeletal joints of the presenter. This type of neural network uses the hardware of GPUs to run at an acceptable speed. The variable load generated by the nearly real-time processing of the recordings of several rooms could not be efficiently handled by just one processing server at peak demand. To handle the processing request without creating delays or installing expensive hardware at each recording room as in the RAP prototype, a load balancing scheme was introduced at the receptor side of the backend. As processing one frame of video, a window of audio or a slide in presentation does not require information about the previous frame, window, or slide, the processing of the input of several rooms could be efficiently parallelized with a simple "next available" queue approach, where the server will request the next available package to process, with the metadata in the package helping the feature extractor to determine the information needed to create a coherent report in the database.

3) Parameterization: In a laboratory setting, the equipment is tuned to produce the most accurate recordings, for example, setting thresholds for noise and light levels, which tend to remain constant through the course of the experimentation. However, as several rooms with different characteristics are now integral part of the recording system, the information about this type of parameters should be not only changed when the room is installed, but also reviewed frequently to adapt to changing conditions in the wild. Although adding an easy-to-change configuration interface to the recording software is a minor change, its importance could not be overstated to avoid unreliable data acquisition.

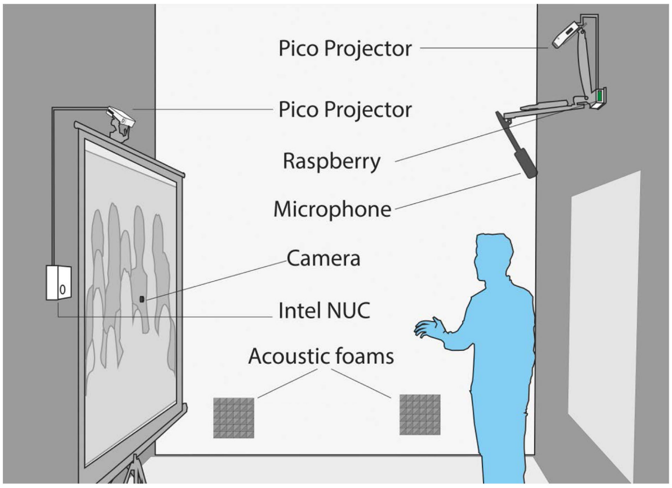
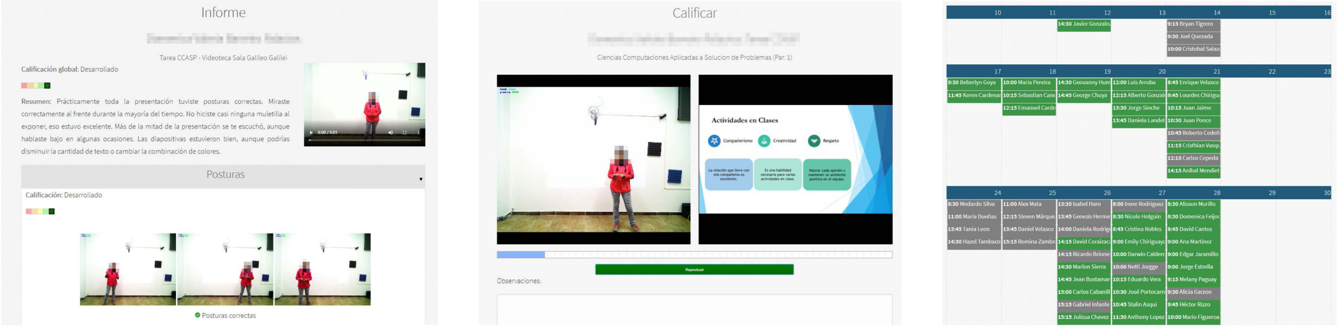


Fig. 5. Room layout in the scaled RAP system incorporates two sensors: a low-cost camera and microphone; two computers: a Raspberry Pi and an Intel NUC; and two pico projectors.

4) Recording Rooms Cost: As the RAP system scaled and more RAP rooms were needed in different parts of our campus, the cost of the rooms had to be optimized. Cost optimization meant finding the right compromises in using low-cost hardware components while maintaining a minimum quality in the recordings. The main changes to the original hardware used in the RAP prototype are: 1) pico projectors with screens instead of high-definition touch screens for displaying presentation slides and the virtual audience; 2) single-board computer (SBC), such as the Raspberry Pi, with a low-cost camera instead of a Kinect sensor plus a personal computer to capture video; 3) SBC with an omnidirectional microphone instead of a microphone array plus a personal computer to record the audio; and 4) acoustic foam to reduce the echo and reverberance of usually rectangular rooms. Fig. 5 shows the room layout of the scaled RAP system with this new components. The final cost of the equipment and basic installation of each room is approximately 2700 USD. The feature extraction performance obtained with this new hardware was at the same level of the one obtained in the prototype room, highlighting the robustness of the feature extraction algorithms.

B. Pedagogical Adaptations

The successful scaling of the RAP system required changes that went beyond the redesign of the system architecture. The scaled RAP system was also redesigned to be a pedagogical tool for instructors instead of a research-oriented prototype. This redesign involved discussions with interested instructors about what will be the role of the RAP system inside established learning practices. After these discussions, it became clear that the original RAP automated feedback was only one component of a more comprehensive feedback strategy that could involve several instructors and that instructors should be



Informe

[redacted]

Tarea CCASP - Videoteca Sala Galileo Galilei

Calificación global: Desarrollado

Resumen: Prácticamente toda la presentación tuviste posturas correctas. Miraste correctamente al frente durante la mayoría del tiempo. No hiciste casi ninguna muletilla al exponer, eso estuvo excelente. Más de la mitad de la presentación se te escuchó, aunque hablaste bajo en algunas ocasiones. Las diapositivas estuvieron bien, aunque podrías disminuir la cantidad de texto o cambiar la combinación de colores.

Posturas

Calificación: Desarrollado

✓ Posturas correctas

(a) Student module: Feedback report

Calificar

[redacted]

Ciencias Computaciones Aplicadas a Solucion de Problemas (Par. 1)

Actividades en Clases

Compañerismo — La relación que llevo con mis compañeros es excelente.

Creatividad — Es una habilidad necesaria para varias actividades en clase.

Respeto — Valorar cada opinión y mantener un ambiente positivo en el equipo.

Reproducir

Observaciones:

(b) Instructor module: Presentation review

10	11	12	13	14	15	16
		14:30 Javier Gonzale:		9:15 Bryan Tigrero 9:30 Joel Quezada 10:00 Cristobal Salaz		

17	18	19	20	21	22	23
9:30 Beberlyn Goya 11:45 Keren Cardenas	10:00 Maria Pereira 10:15 Sebastian Cane 12:15 Emanuel Carde	14:30 Geovanny Hum 14:45 George Chuya	12:00 Luis Arroba 12:15 Alberto Gonzal 13:30 Jorge Sinche 13:45 Daniela Landet	8:45 Enrique Velasco 9:45 Lourdes Chirigua 10:15 Juan Jaime 10:30 Juan Ponce 10:45 Roberto Cedeñ 11:15 Cristhian Vasqu 12:15 Carlos Cepeda 14:15 Anibal Mendiet		

24	25	26	27	28	29	30
8:30 Medardo Silva 11:00 Maria Dueñas 13:45 Tania Leon 14:30 Hazel Tambaco	11:00 Alex Mata 12:15 Steven Márque 13:45 Daniel Velasco 15:15 Romina Zambr	13:30 Isabel Haro 13:45 Genesis Herme 14:00 Daniela Rodrig 14:15 David Coraizac 14:15 Ricardo Brione 14:30 Marlon Sierra 14:45 Jean Bustamar 15:00 Carlos Cabanill 15:15 Gabriel Infante 15:15 Julissa Chavez	8:00 Irene Rodriguez 8:30 Nicole Holguin 8:45 Cristina Robles 9:00 Emily Chiriguaya 10:00 Darwin Calderc 10:00 Netti Jorgge 10:15 Eduardo Vera 10:30 José Portocarr 10:45 Stalin Asqui 11:30 Anthony Lopez	8:30 Alisson Murillo 8:30 Domenica Feijoc 8:45 David Cantos 9:00 Ana Martinez 9:00 Edgar Jaramillo 9:00 Jorge Estrella 9:15 Melany Paguay 9:30 Alicia Garzon 9:45 Héctor Rizzo 10:00 Mario Figueroa		

(c) Administrator module: Schedule monitor

Fig. 6. Scaling the RAP system required a web application with three modules. (a) Student module to schedule presentation appointments and review automatic feedback. (b) Instructor module to create presentation tasks and review students' presentations. (c) Administration module to monitor the usage of the system.

in full control of the system, without the need to interact with the researchers or technicians to evaluate their students and provide them with feedback. The main pedagogical changes and adaptations are described in this subsection.

1) Comprehensive Feedback: The instructors, based on their experience and recommended practice [48], valued the contribution of RAP to provide feedback on the mechanical aspects of presenting, such as posture and voice volume, but they also stated the importance of providing feedback on higher level presentation aspects, such as confidence and rhythm. Moreover, presentations that occur as part of other academic subjects require feedback on the actual content of the presentation. Under this guidance, the feedback report was expanded to be able to contain information from three different sources: the RAP automatic analysis of multimodal features, the evaluation of higher level presentation features provided by a communication expert, and the evaluation of the content of the presentation given by a disciplinary expert. These three sources of feedback are seamlessly integrated into one comprehensive report. A screen capture of the new presentation report can be seen in Fig. 6(a).

2) Instructor Interface: To easily embed the system as an educational tool in a course curriculum, an instructor module was developed for mainly three tasks: register students, assign presentation tasks, and evaluate presentations. Instructors need to register their students in the RAP system at the beginning of the semester to authorize their authentication credentials (e.g., to be able to reserve a slot in the scheduling interface). To facilitate this task, the instructor module allows batch registration of students using a spreadsheet file. After registration, the instructor can create a presentation task, defined by a subject name, start/end date, and course sections, before assigning it to their students. The instructor then can assign the presentation task using established communication channels such as email or the institution's learning management system. The extension of the feedback report required the creation of a review interface, where one or more instructors could evaluate the recording of the presentation and provide their feedback. In this interface, the instructor sees a video of the entire presentation synchronized with the student's slides, and a comment box is provided for the instructor's feedback. Fig. 6(b) shows a screen capture of a presentation review in the instructor interface.

C. Logistical Adaptations

The scaling of the RAP system also required changes in the user interaction process, the error handling, and even the reservation of the recording rooms. We have grouped these required changes under logistical adaptations, which deal with how the system is usually operated and maintained. While these changes seem minor, they can have a profound impact on the actual adoption of the system as the evaluation presented in the following section will show.

1) Recording Scheduling Application: The first small, but required, addition to scale the RAP system for thousands of recordings during the semester was the creation of an application to enable the students to book recording slots in the RAP rooms. This application was implemented as a Django module in the Web App layer that is also responsible for reporting and the instructor interface (see Fig. 4). This application used the institution's single sign-on service to authorize access to the module and to obtain information about the different pedagogical activities in which the student was involved. This system facilitated orderly access to RAP rooms with minimal intervention by administrators. Fig. 6(c) shows a screen capture of the scheduling application.

2) Easy to Use and Maintain: Laboratory prototypes are usually operated and maintained by researchers or their assistants, while end users are passive subjects. Some of the automatic oral presentation feedback systems, such as the Presentation Trainer [22] and the original RAP, were initially designed to be easy to operate, but they were not designed to be used without the close supervision of a technician. In order to be scaled efficiently, the RAP recording room must be operated by the end user, in this case the student/presenter, without the constant need of technical support. All the steps needed to produce a recording were redesigned to be intuitive and tolerant to failures. The overall procedure to use the scaled RAP system can be seen in Fig. 7. First, the student, using the scheduling application, reserves a convenient time slot in one of the available recording rooms. Second, on the appointment

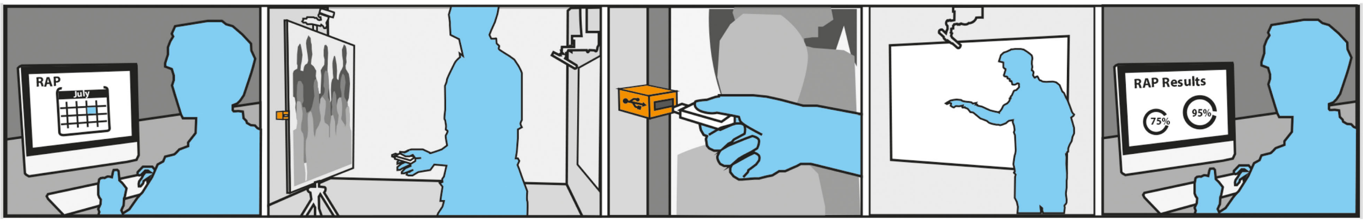


Fig. 7. Typical user experience with the scaled RAP system starts with: 1) schedule a presentation date online; 2) be on time 5 min before the scheduled time with PowerPoint slides in a USB drive; 3) load the presentation on the system; 4) present; and 5) check the feedback results later online.

date and time, the student shows up bringing their slides in the Microsoft PowerPoint format stored in a USB drive. Third, the student follows the instruction provided by the screens of the RAP system to start and stop their presentation. Finally, the student, at a computer at home or in campus, reviews the comprehensive feedback report with the automatic evaluation and, if available, the comments from the instructors. To reduce the need of constant technical support, the RAP system was redesigned to internally handle solvable common technological problems, for example, local recording for later retransmission in case of network failure or to detect more serious problems, such as sensor disconnection, and alert a nearby technical support team.

V. Scale-Test Study

The main objective of the scaled RAP system is to be offered as a service to any instructor that wishes to include it in their courses or to any student that needs to practice an oral presentation inside the institution. To evaluate if this objective is being met by the scaled version of the RAP system, we conducted an "in-the-wild" field study. This study took place during the first semester of 2018. The following methodology was used to evaluate the technological, pedagogical, and logistical adaptations, as well as the development of presentation skills and the general reception of both students and instructors.

1) The technical adaptations were evaluated through a real-world load with hundreds of users.
2) The pedagogical adaptations were evaluated through adoption and use by 40 sections, 16 instructors, and 1099 students.
3) The logistical adaptations were evaluated through the deployment of three RAP rooms across the campus, staffed with three nontechnical operators.
4) The effect on mechanical presentation skills caused by the combination of human and automated feedback was evaluated through the statistical comparison of the scores obtained by the students during two consecutive uses of the scaled RAP system.
5) The perceptions of the students were evaluated using a survey.
6) The perceptions of the instructors were evaluated with personal interviews.

The following subsections present the characteristics of the study, together with the results and main findings.

TABLE II
Total Number of Students and Instructors

	Sections	Instructors	Total Students
Physics I	15	8	532
Communications II	25	8	933
Total	**40**	**16**	**1099**

A. *Study Setting and System Usage*

The scale-test study consisted of adopting the RAP system in two courses: Communications II and Physics I. This courses were selected because they are mandatory in almost all study programs in the institution and together both courses have a range of 1000–2000 registered students in over 40 sections each semester. Also, the leading instructors of these courses were interested in piloting the use of the RAP system.

A total of 1099 students from 15 Physics I sections and 25 Communications II sections participated in the scale-test study during the first semester of 2018. Eight instructors from the Physics I courses and eight instructors from the Communications II courses embedded the RAP system in their curriculum during the study; this included assigning two presentation tasks to their students during the semester and reviewing these presentations while providing feedback.

By including the RAP system in the class curriculum, students obtained comprehensive feedback on their oral presentations within the system. This feedback came from three sources: a communication expert (Communications II instructor), a disciplinary content expert (Physics I instructor), and the RAP system.

Out of the 1099 students that used the RAP system, 933 were enrolled in the Communications II course and 532 in the Physics I course; 366 students were enrolled in both courses (see Table II). This last group was allowed to use the same presentation assignment for both courses and, therefore, received feedback from both instructors.

Instructors assigned two oral presentation tasks to their students: one in the middle of the semester and the other toward the end of the semester with three weeks of separation between them. These assignments were mandatory, and all students had to use the RAP web application to schedule their presentations. Students received a grade for the content of the presentation; presentation skills or the RAP feedback was not used to grade the students.

Three RAP rooms were constructed to accommodate the demand for Tasks 1 and 2. All rooms were available for

TABLE III
USAGE RATE OF RAP ROOMS

	Recordings Task 1	Usage	Recordings Task 2	Usage
Room Alpha	129	29%	303	67%
Room Beta	107	24%	332	74%
Room Physics	309	69%	369	82%
Total	545	40%	1004	74%

TABLE IV
PARTICIPATION RATE OF STUDENTS IN PRESENTATION TASKS

	# Students	Participation rate	Start date
Task 1	536	49%	July 5, 2018
Task 2	1001	91%	August 1, 2018
Tasks 1 and 2	441	40%	

reservation from 8:00 till 16:30 in 15-min slots from Monday to Friday for 14 days for each task. This accounted for a total capacity of 476 recordings per room per task. Table III presents the actual usage of the RAP rooms for both tasks. Students preferred to book their recordings in Room Physics because it was placed significantly closer to them; an internal bus was needed to reach rooms Alpha and Beta.

The student's response was lukewarm at first, with only 49% participation during the first task, despite being mandatory. Participation increased to 91% for the second task, with 40% of students completing both tasks (see Table IV). Fig. 8 shows how the usage of the RAP system rooms progressed over time during both tasks; congestion is observed during the last three days of the first task and almost every day during the second task.

In total, 1549 presentations were made with an average length of 4 min. Fig. 9 shows the daily information upload to the backend per task. On average, the RAP system generated 1.77 GB of data per day with peaks around 3 GB during congestion.

B. *Validation of Technological Adaptations*

The four technological adaptations described in Section IV: decoupling, load balancing, parameterization, and room cost were implemented and evaluated in this scale-test study.

1) Decouple Recording, Processing, and Reporting: The nonscaled version of the RAP system required the constant presence of technical staff in the RAP room during all presentations as failures often required physical access to the hardware on the rooms. The scaled version of the RAP system was redesigned as seen in Figs. 2–4 to decouple all software modules using MQTT for command signaling and AMQP for file queuing. During the scale-test study, most failures requiring technical assistance were isolated at the backend, where they were remotely attended. Failures in the RAP rooms were solved by nontechnical staff by simply restarting the hardware. Consequently, no technical staff was needed at the RAP rooms during recordings.

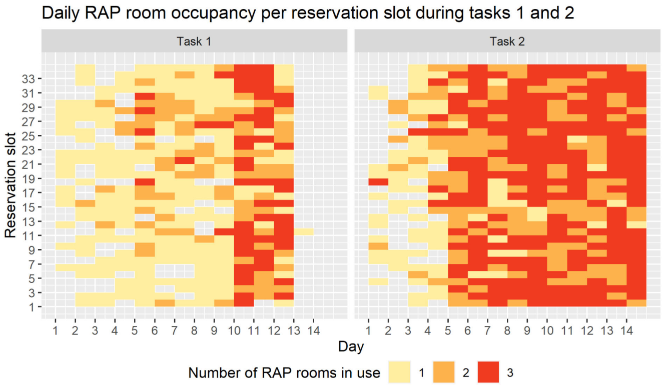


Fig. 8. Reservation of the three RAP rooms was subdivided into 34 15-min slots for each day in both tasks. Room occupancy congestion was observed in the last few days of task 1 and almost in entirety in task 2.

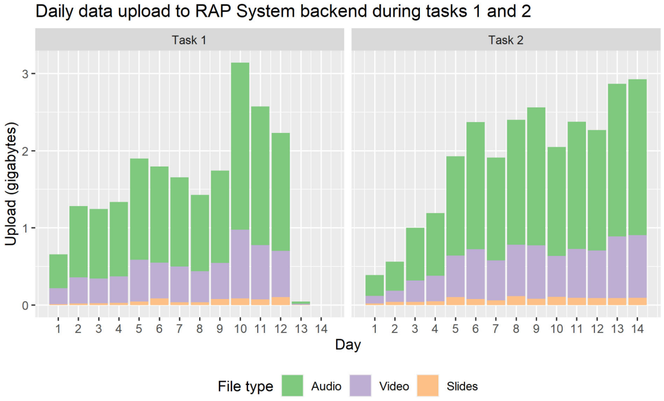


Fig. 9. Daily data upload to the RAP system backend, including audio, video, and slides for each presentation, peaked at around 3 GB for both tasks.

2) Load Balancing: In the scale-test study, three GPU-equipped servers were tasked at the backend to handle the load of processing the presentations. The main computational bottleneck was the extraction of skeletal joints with OpenPose; therefore, videos were downsampled to 5 frames/s, and a NVidia Tesla K20m GPU was used at each server to process each presentation. On average, a feedback report was ready for each presenter 5 min after finishing their presentation, even during congestion.

3) Parameterization: During the scale-test study, parameterization allowed the independent calibration of all RAP rooms. For example, for voice volume, it is necessary to measure the room's acoustic noise floor, and for video features (posture and gaze), it is necessary to measure the room's lighting level. These measurements were different for each room and tended to change over time (e.g., recalibration of room lighting after a spent fluorescent lamp). As a result, data across rooms were comparable, and no statistically significant difference was observed between RAP room measurements.

4) Recording Rooms Cost: The average price of equipping each room was around 2730 USD (see Table V for details); this price range facilitated the procurement of the three RAP rooms used in this study.

TABLE V
COST OF ONE RAP ROOM IN US DOLLARS

Quantity	Description	Cost
1	Projection screen	130
2	pico projectors	800
1	Fish eye camera	100
1	SBC (Raspberry Pi + accesories)	150
1	SBC (Inter NUC + accesories)	500
1	Tripod	80
10	Acoustic foam sets	700
1	WiFi router	50
1	Shotgun microphone + XLE adapter	200
1	Laser pointer	20
	Total	2730

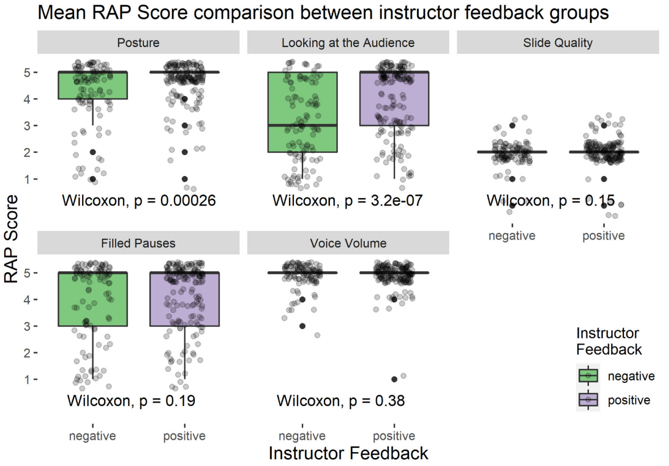


Fig. 10. Some agreement was observed between instructor feedback, classified as negative or positive, and RAP posture and looking at the audience feedback.

C. Validation of Pedagogical Adaptations

Instructors provided feedback to their students through a text box in the instructor interface of the RAP system. This feedback was later automatically appended to the student's presentation report. Overall, 74% of presentations had feedback from an instructor that addressed the following: 75% disciplinary content of the presentation, 19% higher level presentation aspects (speed, graphics, tone, etc.), and 4% mechanical presentations aspects covered by the RAP system.

In order to evaluate if agreement existed between instructor and RAP feedback, we classified the instructor feedback using keywords such as "excellent," "very well," or "well done" as positive and "lack," "improve," or "bad" as negative. This method labeled 25% of instructor feedback as positive and 15% as negative. Using this classification, statistically significant agreement was observed between the instructor and the RAP feedback for posture (effect size 10%) and gaze (effect size 22%), that is, positive feedback was generally given to students with high scores on these features (see Fig. 10).

D. Validation of Logistical Adaptations

The scheduling application used in this scale-test study presented two important advantages: it was well integrated with the rest of the RAP system and it scaled easily to more than 1500 presentations. Students and instructors reported no issues regarding availability and use of this tool.

Troubleshooting the RAP rooms during recordings required mostly the intervention of three nontechnical staff hired to be present on-site during recordings in each RAP room. They reported failures caused mainly by defective slide changers and Internet connection blackouts. One technical staff was placed on call to attend system failures either remotely or on-site if needed. On average, technical staff was required on-site thrice per week during recordings.

E. Analysis of Learning Gains

The main objective of using an oral presentation feedback system such as RAP is to help students improve their oral presentation skills. One way to measure if the use of the system has a positive effect on the development of these skills is to compare the score that the students obtained during their first and second interaction with the system. This methodology was used to base the comparison in an objective measurement of the oral presentation skills being trained through the RAP system and to follow the same methodology used in most laboratory-based analyses to assess learning gains [22], [49], but this time on the wild. Given that the scaled RAP system uses both human and automated feedback, this analysis evaluates the combined effect of these two components. For a more detailed evaluation of just the automated feedback, refer to the controlled experiment described in [50].

Due to the highly skewed and nonnormal distribution of scores in the different dimensions, usual parametric tests, such as a paired t-test, are not recommended to determine the difference between the scores in the consecutive measurements. The Matched Samples Sign Test [51] is used to determine if there is a difference in the median of two paired populations with higher statistical power. Through this test, it was found that the students that performed poorly during the first task measurement had a statistically significant improvement of one or two points in most dimensions during the second task measurement, hinting to a possible positive learning effect by the use of the system. The system, however, does not seem to affect students with already high scores, as measured by the system. The following subsections will describe in detail the analysis and results for each one of the different presentation dimensions measured by the RAP system.

1) Posture: During the RAP measurement of the first task, most students (89%) already obtained a high grade (4 or 5 over 5). During the second measurement, this percentage increased (93%). Fig. 11 shows that this increase was felt by all students, especially those with lower scores, as the new median for students that originally obtained a 1 was later a 4, while for the rest of the students, the new median is 5. A more formal signed-test to detect a difference between the distributions indicates that there is a small (less than 1 in the median value) but statistically significant ($p < 0.01$) improvement for all students. This improvement is larger (two in the median value) for low-performing students (scores 3 or less), and this

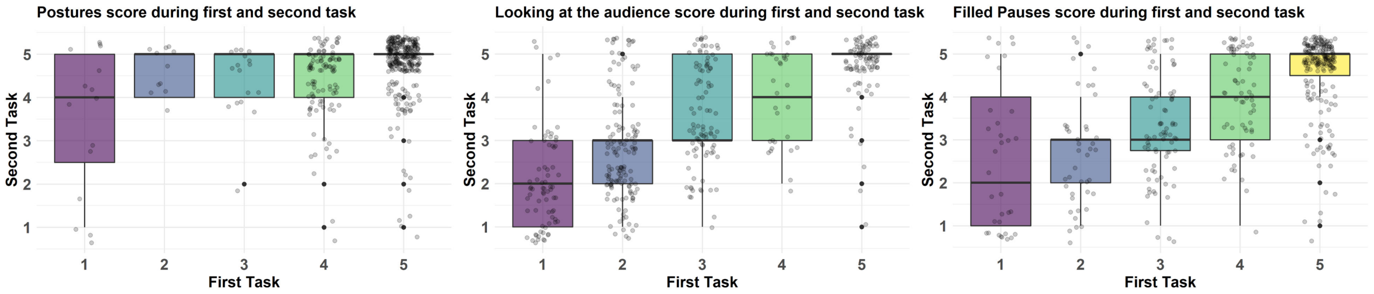


Fig. 11. RAP score gains in Posture, looking at the audience (gaze), and filled pauses from first task to second task.

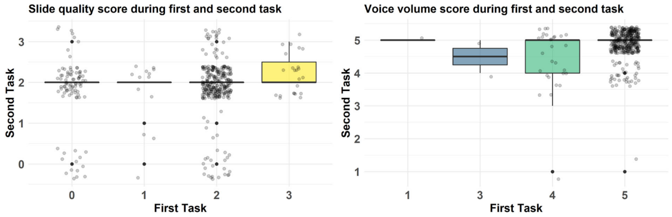


Fig. 12. RAP score gains in slide quality and voice volume from first task to second task. In this version of the RAP system, the slide quality score ranged from 0 to 3.

improvement is statistically significant ($p < 0.001$). The difference for higher performers (score four or higher) is not statistically significant.

2) Looking at the Audience: In the first measurement, there was a wide distribution of scores, with a minority (29%) obtaining a four or higher. During the second measurement, this percentage increased considerably (41%). Fig. 11 provides a more general view of the change between measurements. Low-scoring students (one and two in the first measurement) increased their medians (to 2 and 3, respectively); the signed-test confirms the visual interpretation. There is a small, but significant, increment in the score ($p < 0.001$) in general. For originally low-performing students (score 3 or lower), the difference is larger (one point in the median) and significant ($p < 0.001$). Again, no significant difference was found in the higher performing group (score 4 or higher).

3) Filled Pauses: Similarly to Looking at the Audience, the scores of Filled Pauses are widely distributed in the scale, with only 64% of the students obtaining a high score during the first measurement (four or higher). This percentage remained similar during the second measurement (63%), not indicating a large change in the performance of students. A visual inspection of Fig. 11 reveals that there is change only for low initial scores, where students with 1 or 2 obtained a median of 2 and 3, respectively, in the second measurement. This result is again corroborated by the statistical test, where there is only a one point significant ($p < 0.001$) difference for students with scores three or lower.

4) Voice Volume: The measurement of Voice Volume was saturated during the first measurement. As can be seen in Fig. 12, practically, all of the students (99%) obtained a high score (4 or higher) during the first and second measurements. Due to this saturation, the effect of the system on this oral presentation skill cannot be established (no statistically significant difference was found at any performance level).

5) Slide Quality: For Slide Quality, the percentage of high performers (two or three over three) in the first measurement was 76%, while in the second measurement, it increased to 89%. Fig. 12 suggests that lower performers (zero or one) mostly increased their performance in the second measurement, while higher performers obtained similar scores in both measurements. This visual interpretation coincides with the statistical test, which found that there was a small (median difference equal to zero) but significant ($p < 0.001$) increase for the general population, and a large (median difference equal to two) and significant ($p < 0.001$) increase for low performers. There is no significant difference between high performers (two or higher).

F. Analysis of the Student and Instructor Experience

After 1549 RAP recordings from 1099 students in 40 course sections with 16 instructors in one academic semester, feedback from instructors and students was collected after the end of the academic semester. Student experiences were collected using an online survey, which contained open questions about the student's perception of technological, logistical, and educational aspects of the RAP system. Instructors were interviewed individually.

1) Instructors: The qualitative experience of participating instructors was obtained by interviewing four of the 16 participating instructors: two from the Physics sections and two from the Communication sections. During the interviews, the instructors were asked about their experiences with the usage of the RAP system and to highlight its positive and negative aspects.

The main task of instructors was to review the recordings of their students and provide timely feedback using the instructor module in the online web app. Instructors reported that this workload represented 8–10 min per student's recording during office hours. This was equivalent to grading a short evaluation, and in courses with 40 or more students, this was a significant addition to their office hours workload. Instructors, therefore, emphasized the need to consider this while planning the curriculum of the course.

According to instructors from the Communication courses, the biggest advantage of the RAP system was that they were now able to assess the presentation skills of every student

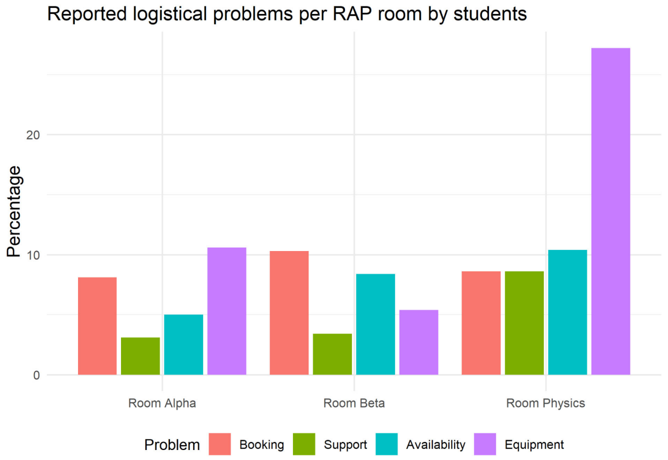


Fig. 13. Percentage of reported problems in the students' survey: technical problems and booking overload in Room Physics was the source of most complaints.

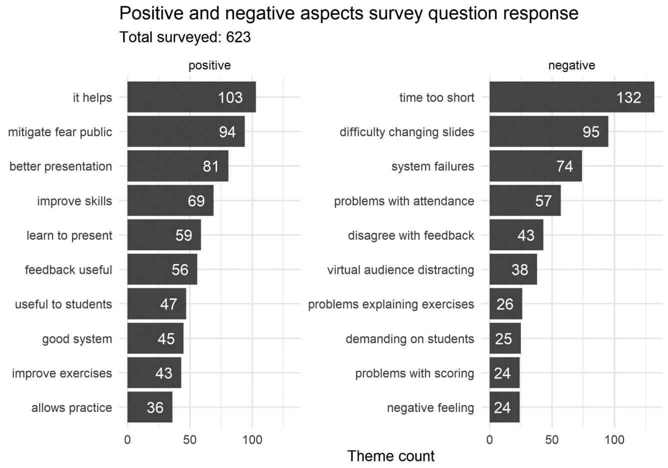


Fig. 14. Themes found in answers to the question: What are the positive and negative aspects that you would highlight in the development and dissertation of the RAP task, as well as the feedback of the system?

several times during the semester. This was simply not possible before as there was no time for every student to present during class sessions. However, they noted that the system should be able to summarize the progress of the student's skills over time to easily detect students who needed additional help to develop their skills.

Instructors from the Physics courses remarked that the RAP system allowed them to evaluate the student's understanding of the course contents in a different context. For example, they quickly detected incorrect usage of physics terms in the RAP recordings and were able to reinforce misunderstood concepts during class.

2) Students: The student's experience with the RAP system was obtained through a survey at the end of the semester. The survey was optional and offered to all registered students in the participating Physics I and Communications II sections through an online tool. It contained seven free-form text entry questions about their experiences with the RAP system in general, the RAP room, the RAP feedback, and the instructor's feedback. In total, 623 students answered the survey; a summary of their responses is presented in the following.

Logistical problems during the semester, especially regarding online booking, availability of slots, and equipment failure during presentations, were the source of most complaints by students (see Fig. 13). The most common themes in the students' free-form response to the positive and negative aspects of the RAP system are presented in Fig. 14. The negative aspects themes *time too short*, *difficulty changing slides*, and *problems with attendance* are the most salient because students struggled mostly with the reservation system and the slide changer device in the room. The slide changer failed often, causing frustration to students and in some cases cancellations. Also, the fact that the RAP system cuts the presentation after 5 min annoyed most students. As for scheduling issues, students were given three weeks to book an appointment in one of the RAP rooms for each presentation task. However, most students booked their presentation in the third week, resulting in last-minute congestion. This was evident during the first task, where the third week was completely booked, while the first and second weeks were barely used resulting in a misleading 40% occupancy rate. To avoid this problem in the second task, instructors had to vehemently warn students to book in a timely manner; this resulted in a higher occupancy rate of 76% as the time distribution of reservations was more evenly spread.

Overall, the student's perception of the system and the technology was positive. As for the positive aspects in Fig. 14, the themes *it helps*, *mitigate fear public*, and *improve skills* stand out because most students reported that the system helped them to prepare and mitigate their fear of public speaking. "*It allowed me to see my mistakes during an oral presentation*," "*It helped me to improve my posture and self-confidence*," "*It helped me to control my presentation time and my posture*," "*You get feedback right away*" are a few representative comments from the students regarding the positive aspects of the system and the impact it had on their communication skills.

VI. Lessons Learned and Conclusion

After analyzing the results obtained from the deployment of the scaled RAP system, there are several lessons learned that could guide the transitions from other multimodal learning analytic applications from laboratory to institution.

1) *Instructor Buy-In*: Most of the success of the deployment, and even the opportunity of executing the deployment, is based on the active interest and participation of the instructors. Having them, instead of the researchers behind the tool, as main advocates for the system was a key element in convincing the institutional management to invest the economical, political, and logistical resources needed to execute a project at institutional level. Ultimately, the acceptance of the system by instructors becomes critical for its long-term sustainability.
2) *Simplifying the System*: The system has to be reliable, easy to maintain, and easy to use for all stakeholders to favor acceptance and adoption. This is best accomplished

by deploying only the most mature and well-tested features of the system.

3) *Integration With Existing Practices*: It is of critical importance that the system is perceived as a useful educational tool and not as another imposed burden. To avoid this, apart from accounting for the expected surge in the demand for physical and computational resources, it is necessary to consider how the application can integrate into the day-to-day educational activities of instructors with minimal management overhead and how can it be accessed by students with minimal effort.
4) *Role of Logistics*: Even if the system works as desired, small but important logistical aspects can get in the way of its adoption.
5) *Providing Evidence*: The effectiveness of the system should be monitored continuously to keep the interest in its use. Providing clear metrics for the learning gains obtained by the students (using the same capabilities of the multimodal system) is an easy way to report back to the main stakeholders in the institution.

Out of the development of this scaled RAP system, the authors have envision several possible improvements on the RAP system and oral presentation feedback systems in general.

1) *Exploring New Modalities:* The diversity of modalities employed to assess oral presentation quality should be improved in new versions of the RAP system. The existence of mechanical presentation feedback given by instructors, such as cadence, nervousness, or tone, indicate that the space could be better explored. For example, biosignals can be captured from the speaker to detect arousal and nervousness. Visual aids information analysis can be expanded to include content and organization. Speech content can be exploited to assess coherence and difficulty. In the same way that the access to low-cost sensors started the first wave of oral presentation feedback systems, the availability of new and improved AI tools could lead to a new generation of systems with the capability of not only examining nonverbal behavior but also semantically analyzing the verbal components in the contexts of those behaviors.
2) *Use or Development of a Multimodal Framework:* The use of a foundational framework (in the same line as Social Signal Interpretation [52] or the Platform for Situated Intelligence [53]), which takes care of basic functions such as synchronization, buffering, multimodal feature extraction, and fusion, is not only important to reduce the effort needed to create and maintain multimodal systems, but also to more readily share existing solutions. Instead of having to share raw code, researchers could share predefined pipelines and plugins inside these frameworks. This will lead to the establishment of best practices and an incremental progression of both the effectiveness and efficiency of new oral presentation feedback systems, while also contributing to the whole field of MmLA.
3) *Focus on Adoption Features:* The era of proofs of concepts for oral presentation systems is well past. New systems should be designed from the beginning with adoption and scalability in mind. The minimization of cost, intrusiveness, and difficulty of use should be early requirements on par with the extraction of features and generation of feedback reports. Only systems that have these features have a chance to be released from the laboratory and have an impact on the acquisition of presentation skills.
4) *Connect With Pedagogical Practice:* The ultimate goal of most oral presentation systems is to be used in formal educational settings. As such, these systems should integrate with the way that oral presentation is currently taught. Also, these systems should connect with existing learning management systems, where practice sessions can be assigned and the evaluations can be stored.
5) *Improved Privacy*: As the system records students, privacy is an important concern. Future versions of the system should include privacy options such as a "delete/forget" button for any recording and tokenization of sensitive stored data.

This article has proven that systems that can automatically evaluate oral presentation and are able to provide a detailed report back to the speaker are not only possible but abundant. It is now the challenge of the MmLA community to prove also that these systems can be improved, made useful, and open the door to a new class of systems that can augment our capabilities to learn communication and other 21st century skills.

The authors hope that this article serves as inspiration to other researchers in the MmLA field to take the next step with their prototypes. For this, it is critical that we bring instructors, students, and even institutional managers into the design table, not only to help us to create a better product, but also to set the real objectives of the system beyond research.

Acknowledgment

The authors would like to thank all the instructors and students involved in the scale study. They would also like to thank the academic management of Escuela Superior Politécnica del Litoral for their support in the scaling of the RAP application.

References

[1] X. Ochoa, "Multimodal learning analytics," *Soc. Learn. Analytics Res.*, vol. 1, pp. 129–141, 2017.

[2] T. Janík, T. Seidel, and P. Najvar, *Introduction: On the Power of Video Studies in Investigating Teaching and Learning*. New York, NY, USA: Waxmann, 2009, pp. 7–19.

[3] X. Ochoa and M. Worsley, "Augmenting learning analytics with multimodal sensory data," *J. Learn. Analytics*, vol. 3, no. 2, pp. 213–219, Sep. 2016, doi: 10.18608/jla.2016.32.10.

[4] R. Martínez-Maldonado, J. Kay, S. B. Shum, and K. Yacef, "Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data," *Human–Comput. Interact.*, vol. 34, no. 1, pp. 1–50, 2019, doi: 10.1080/07370024.2017.1338956.

[5] E. Starr, J. Reilly, and B. Schneider, "Toward using multi-modal learning analytics to support and measure collaboration in co-located dyads," in *Proc. 13th Int. Conf. Learn. Sci.*, London, U.K., Jun. 2018, pp. 448–455, doi: 10.22318/cscl2018.448.

[6] L. Prieto, K. Sharma, L. Kidzinski, M. J. Rodríguez-Triana, and P. Dillenbourg, “Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data,” *J. Comput. Assisted Learn.*, vol. 34, no. 2, pp. 193–203, 2018, doi: 10.1111/jcal.12232.

[7] S. Howard, K. Thompson, J. Yang, J. Ma, A. Pardo, and H. Kanasa, “Capturing and visualizing: Classroom analytics for physical and digital collaborative learning processes,” in *Proc. 12th Int. Conf. Comput. Supported Collaborative Learn.*, Philadelphia, PA, USA, Jun. 2017, pp. 801–802.

[8] H. Cornide-Reyes *et al.*, “Introducing low-cost sensors into the classroom settings: Improving the assessment in agile practices with multimodal learning analytics,” *Sensors*, vol. 19, no. 15, 2019, Art. no. 3291, doi: 10.3390/s19153291.

[9] D. Azcona, I.-H. Hsiao, and A. F. Smeaton, “Personalizing computer science education by leveraging multimodal learning analytics,” in *Proc. IEEE Front. Educ. Conf.*, San Jose, CA, USA, Jun. 2018, pp. 1–9, doi: 10.1109/FIE.2018.8658596.

[10] A. Corbi, O. Santos, and D. Burgos, “Intelligent framework for learning physics with aikido (martial art) and registered sensors,” *Sensors*, vol. 19, no. 17, 2019, Art no. 3681, doi: 10.3390/s19173681.

[11] F. Domínguez, V. Echeverría, K. Chiluiza, and X. Ochoa, “Multimodal selfies: Designing a multimodal recording device for students in traditional classrooms,” in *Proc. 17th Int. Conf. Multimodal Interact.*, Seattle, WA, USA, Nov. 2015, pp. 567–574, doi: 10.1145/2818346.2830606.

[12] D. Di Mitri, J. Schneider, M. Specht, and H. Drachsler, “From signals to knowledge: A conceptual model for multimodal learning analytics,” *J. Comput. Assisted Learn.*, vol. 34, no. 4, pp. 338–349, 2018, doi: 10.1111/jcal.12288.

[13] M. Worsley and P. Blikstein, “Deciphering the practices and affordances of different reasoning strategies through multimodal learning analytics,” in *Proc. Multimodal Learn. Analytics Workshop Grand Challenge*, Istanbul, Turkey, Nov. 2014, pp. 21–27, doi: 10.1145/2666633.2666637.

[14] S. Praharaj, M. Scheffel, H. Drachsler, and M. Specht, “Multimodal analytics for real-time feedback in co-located collaboration,” in *Proc. Eur. Conf. Technol. Enhanced Learn.*, Leeds, U.K., Sep. 2018, pp. 187–201.

[15] X. Ochoa, F. Domínguez, B. Guamán, R. Maya, G. Falcones, and J. Castells, “The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors,” in *Proc. 8th Int. Conf. Learn. Analytics Knowl.*, Sydney, Australia, Mar. 2018, pp. 360–364, doi: 10.1145/3170358.3170406.

[16] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi, “Presentation sensei: A presentation training system using speech and image processing,” in *Proc. 9th Int. Conf. Multimodal Interfaces*, Nagoya, Japan, Nov. 2007, pp. 358–365, doi: 10.1145/1322192.1322256.

[17] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, “Cicero—Towards a multimodal virtual audience platform for public speaking training,” in *Proc. Intell. Virtual Agents*, Edinburgh, U.K., Aug. 2013, pp. 116–128, doi: 10.1007/978-3-642-40415-3_10.

[18] J. Li, Y. Wong, and M. S. Kankanhalli, “Multi-stream deep learning framework for automated presentation assessment,” in *Proc. IEEE Int. Symp. Multimedia*, San Jose, CA, USA, Dec. 2016, pp. 222–225, doi: 10.1109/ISM.2016.0051.

[19] F. Dermody and A. Sutherland, “A multimodal system for public speaking with real time feedback,” in *Proc. 17th Int. Conf. Multimodal Interact.*, Seattle, WA, USA, Nov. 2015, pp. 369–370, doi: 10.1145/2993148.2998536.

[20] A. Nguyen, W. Chen, and M. Rauterberg, “Intelligent presentation skills trainer analyses body movement,” in *Proc. Int. Work Conf. Artif. Neural Netw.*, Palma de Mallorca, Spain, Jun. 2015, pp. 320–332, doi: 10.1007/978-3-319-19222-2_27.

[21] A. Lui, S. Ng, and W. Wong, “A novel mobile application for training oral presentation delivery skills,” in *Proc. Int. Conf. Technol. Educ.*, Hong Kong, China, Jul. 2015, pp. 79–89, doi: 10.1007/978-3-662-48978-9_8.

[22] J. Schneider, D. Börner, P. Van Rosmalen, and M. Specht, “Presentation trainer, your public speaking multimodal coach,” in *Proc. 17th Int. Conf. Multimodal Interact.*, Seattle, WA, USA, Nov. 2015, pp. 539–546, doi: 10.1145/2818346.2830603.

[23] I. Tanveer, E. Lin, and M. E. Hoque, “Rhema: A real-time in-situ intelligent interface to help people with public speaking,” in *Proc. 20th Int. Conf. Intell. User Interfaces*, Atlanta, GA, USA, Mar. 2015, pp. 286–295, doi: 10.1145/2678025.2701386.

[24] I. Tanveer, R. Zhao, K. Chen, Z. Tiet, and M. E. Hoque, “Automanner: An automated interface for making public speakers aware of their mannerisms,” in *Proc. 21th Int. Conf. Intell. User Interfaces*, Sonoma, CA, USA, Mar. 2016, pp. 385–396, doi: 10.1145/2856767.2856785.

[25] H. Trinh, R. Asadi, D. Edge, and T. Bickmore, “RoboCOP: A robotic coach for oral presentations,” *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 27, Jun. 2017, Art. no. 27, doi: 10.1145/3090092.

[26] J. Schneider, G. Romano, and H. Drachsler, “Beyond reality—Extending a presentation trainer with an immersive VR module,” *Sensors*, vol. 19, no. 16, 2019, Art no. 3457, doi: 10.3390/s19163457.

[27] R. Hincks, “Processing the prosody of oral presentations,” in *Proc. Symp. Comput. Assist. Learn.*, Venice, Italy, Jun. 2004, pp. 1–4.

[28] R. Hincks, “Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism,” *System*, vol. 33, no. 4, pp. 575–591, 2005, doi: 10.1016/j.system.2005.04.002.

[29] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz, “MLA’14: Third multimodal learning analytics workshop and grand challenges,” in *Proc. 16th Int. Conf. Multimodal Interact.*, Istanbul, Turkey, Nov. 2014, pp. 531–532.

[30] T. Gan, Y. Wong, B. Mandal, V. Chandrasekhar, and M. S. Kankanhalli, “Multi-sensor self-quantification of presentations,” in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, Oct. 2015, pp. 601–610, doi: 10.1145/2733373.2806252.

[31] L. Chen, C. W. Leong, G. Feng, and C. M. Lee, “Using multimodal cues to analyze MLA’14 oral presentation quality corpus: Presentation delivery and slides quality,” in *Proc. Multimodal Learn. Analytics Workshop Grand Challenge*, Istanbul, Turkey, Nov. 2014, pp. 45–52, doi: 10.1145/2666633.2666640.

[32] V. Echeverría, A. Avendaño, K. Chiluiza, A. Vásquez, and X. Ochoa, “Presentation skills estimation based on video and kinect data analysis,” in *Proc. Multimodal Learn. Analytics Workshop Grand Challenge*, Istanbul, Turkey, Nov. 2014, pp. 53–60.

[33] G. Luzardo, B. Guamán, K. Chiluiza, J. Castells, and X. Ochoa, “Estimation of presentations skills based on slides and audio features,” in *Proc. Multimodal Learn. Analytics Workshop Grand Challenge*, Istanbul, Turkey, Nov. 2014, pp. 37–44.

[34] F. Haider, L. Cerrato, N. Campbell, and S. Luz, “Presentation quality assessment using acoustic information and hand movements,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 2812–2816, doi: 10.1109/ICASSP.2016.7472190.

[35] A. Hanani, M. Al-Amleh, W. Bazbus, and S. Salameh, “Automatic estimation of presentation skills using speech, slides, and gestures,” in *Proc. Speech and Comput.*, Hatfield, U.K., Sep. 2017, pp. 182–191, doi: 10.1007/978-3-319-66429-3_17.

[36] D. A. Silverstein and T. Zhang, “System and method of providing evaluation feedback to a speaker while giving a real-time oral presentation,” U.S. Patent US7 050 978B2, May 2006.

[37] S. Lewis, “Interactive speech preparation,” U.S. Patent US20 110 231 194A1, Sep. 2011.

[38] L. Chen, G. Feng, C. W. Leong, C. Kitchen, and C. M. Lee, “Systems and methods for providing a multi-modal evaluation of a presentation,” U.S. Patent US20 140 302 469A1, Oct. 2014.

[39] S. M. Miller and A. R. Sand, “System and method using feedback speech analysis for improving speaking ability,” U.S. Patent US9 230 562B2 Oct. 2014.

[40] J. Pasquero, D. R. Walker, and S. H. Fyke, “Methods and devices for facilitating presentation feedback,” U.S. Patent US9 264 245B2, Feb. 2016.

[41] S. M. Brand, E. M. Dow, and T. D. Fitzsimmons, “Cognitive presentation advisor,” U.S. Patent US9 633 008B1, Apr. 25, 2017.

[42] G. Perez, I. P. Tudela, and M. Castro, “Automated speech coaching systems and methods,” Spain Patent WO2 019 017 922A1, Jan. 2019.

[43] V. Vangala and R. Gunda, “Intelligent assistance in presentations,” U.S. Patent US20 180 122 371A1, May 2018.

[44] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 1302–1310, doi: 10.1109/CVPR.2017.143.

[45] N. H. De Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–390, 2009, doi: 10.3758/BRM.41.2.385.

[46] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken english," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4857–4860, doi: 10.1109/ICASSP.2009.4960719.

[47] V. Echeverría, B. Guamán, and K. Chiluiza, "Mirroring teachers' assessment of novice students' presentations through an intelligent tutor system," in *Proc. Conf. Comput. Aided Syst. Eng.*, Quito, Ecuador, Jul. 2015, pp. 264–269.

[48] 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," *Educ. Res. Rev.*, vol. 14, pp. 62–80, 2015, doi: 10.1016/j.edurev.2015.02.002.

[49] 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 ACM Conf. Human Factors Comput. Syst.*, Seoul, South Korea, Apr. 2015, pp. 565–574, doi: 10.1145/2702123.2702314.

[50] X. Ochoa and F. Domínguez, "Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting," *Brit. J. Educ. Technol.*, vol. 51, no. 5, pp. 1615–1630, 2020, doi: 10.1111/bjet.12987.

[51] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference: Revised and Expanded*. Boca Raton, FL, USA: CRC Press, 2014.

[52] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The Social Signal Interpretation (SSI) framework: Multimodal signal processing and recognition in real-time," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, Oct. 2013, pp. 831–834, doi: 10.1145/2502081.2502223.

[53] D. Bohus, S. Andrist, and M. Jalobeanu, "Rapid development of multimodal interactive systems: A demonstration of platform for situated intelligence," in *Proc. 19th Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 493–494, doi: 10.1145/3136755.3143021.

Federico Domínguez (Senior Member, IEEE) received the Ph.D. degree in sensor networks for environmental monitoring from Vrije Universiteit Brussel, Ixelles, Belgium, in 2014.

Since 2014, he has been a Professor with the Faculty of Electrical and Computer Engineering, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador, and the Head of the Smart Environments Laboratory, Information Technology Center, Escuela Superior Politécnica del Litoral. His main research interests include Internet of Things technologies applications in learning and working environments, intersecting with learning analytics, smart buildings, homes, and cities.

Xavier Ochoa received the bachelor's degree in computer science from Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador, in 2000, the M.Sc. degree in applied computer science from Vrije Universiteit Brussels (VUB), Ixelles, Belgium, in 2002, and the Ph.D. degree in engineering from Katholieke Universiteit Leuven (KULeuven), Leuven, Belgium, in 2008.

He is an Assistant Professor of Learning Analytics in the Department of Administration, Leadership, and Technology, Steinhardt School of Culture, Education, and Human Development, New York University (NYU), New York, NY, USA. He has been an active researcher in the field of learning analytics. His main research interest includes the automatic measurement and feedback of 21st century skills through multimodal analysis.

Dr. Ochoa is the Editor-in-Chief for the *Journal of Learning Analytics*.

Dick Zambrano received the master's degree in physics teaching and the engineering degree in electronics and telecommunications from the Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador, in 2010 and 1989, respectively.

He is currently a Professor with the Faculty of Natural Sciences and Mathematics and the Director of the Action-Research Program (PIA) with the Escuela Superior Politécnica del Litoral, where he is focused on educational innovation that promotes active learning. His research interests include active, flexible, and personalized learning, as well as formative assessment and feedback.

Katherine Camacho received the engineering degree in electronics and telecommunications from the Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador, in 2016.

She is currently a Teacher Assistant with the Physics Laboratory Section, Faculty of Science and Mathematics, Escuela Superior Politécnica del Litoral. She directs and organizes the student's laboratory sessions and is interested in the research of digital tools that contribute to teaching–learning techniques.

Ms. Camacho is a Member of the IEEE Ecuador Section.

Jaime Castells received the engineering degree in computer science from the Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador, in 2015.

Since 2012, he has been a Researcher with the Information Technology Center, Escuela Superior Politécnica del Litoral. He has participated in several research projects as an expert Web developer, graphical user interface creator, and experiment designer. His main research interest includes human–computer interaction in educational applications.