The RAP System: Automatic Feedback of Oral Presentation Skills Using Multimodal Analysis and Low-Cost Sensors

Xavier Ochoa
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador

Federico Domínguez
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador

Bruno Guamán
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador

Ricardo Maya
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador

Gabriel Falcones
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador

Jaime Castells
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador

ABSTRACT

Developing communication skills in higher education students could be a challenge to professors due to the time needed to provide formative feedback. This work presents RAP, a scalable system to provide automatic feedback to entry-level students to develop basic oral presentation skills. The system improves the state-of-the-art by analyzing posture, gaze, volume, filled pauses and the slides of the presenters through data captured by very low-cost sensors. The system also provides an off-line feedback report with multimodal recordings of their performance. An initial evaluation of the system indicates that the system's feedback highly agrees with human feedback and that students considered that feedback useful to develop their oral presentation skills.

CCS CONCEPTS

• **Applied computing → Computer-assisted instruction**;

KEYWORDS

Multimodal Learning Analytics, Posture, Gaze, Filled-Pauses

ACM Reference Format:
Xavier Ochoa, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. 2018. The RAP System: Automatic Feedback of Oral Presentation Skills Using Multimodal Analysis and Low-Cost Sensors. In *LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.1145/3170358.3170406

1 INTRODUCTION

Communication, together with creativity, critical thinking and collaboration, have been identified as the main "4 Cs" skills that any individual needs to succeed in the 21st century society [18]. Due to their importance, higher education institutions work heavily on providing opportunities for students to develop their communication skills, especially in programs where they were often overlooked (e.g. engineering) [16]. The preferred way to foster these skills is to integrate their practice through the curricula [10]. However, embedding communication skills transversally in the curriculum requires that professors provide feedback to their students not only in the technical content of their work, but in how it is communicated. This can be an overwhelming additional task to professors.

As a contribution to reduce the burden on professors, while providing continuous opportunities for students to develop their communication skills, this work focuses on the description and initial evaluation of an automated system to provide feedback on basic oral presentation skills for entry-level students. This system uses Multimodal Learning Analytics (MMLA) [13] techniques to capture student oral presentation behavior through different modalities. The system synchronizes, processes, and analyzes the different data streams and provides automatic feedback to the student. This system was created based on the experience obtained in the evaluation of similar systems. We call this system RAP as it is the Spanish acronym for "Automatic Presentation Feedback".

2 RELATED WORK

Due to the availability of inexpensive sensors, like the Microsoft Kinect depth camera, during the last 5 years, there have been several systems that use multimodal information to provide feedback to oral presentations. This section discusses the main examples and initiatives and contrast them with the RAP system. To facilitate the comparison, these systems are classified according to the type of sensors used, the modalities that they analyze and the type of feedback that is provided.

Previous systems can be classified into heavy, medium, and light sensor requirements. The system presented in Gan et al. [8] is an example of heavy sensor use, requiring several static cameras, a Microsoft Kinect sensor, and Google Glass' wearable camera and sensors. A medium sensor use is exemplified by Batrinca et al. [4] with only 2 fixed cameras, a Microsoft Kinect, and a lapel wireless microphone. The number and type of sensors have a direct impact on the variety of modes that can be captured, but also in the cost and the invasiveness of the system. The RAP system will only use a very light set of sensors: a camera (webcam quality) and

LAK'18, March 7–9, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170406

an ambient microphone, relying on software to provide similar modality extraction than medium, and even heavy, sensor systems.

Different systems also analyze different modalities. The most studied modality for oral presentation is audio [3, 4, 8]. From the audio signal, paralanguage features, such as volume and filled-pauses, have proven useful for assessing oral presentation skills [9, 12]. The second most studied modality is posture [4, 8, 14], extracted usually using the depth information from the Microsoft Kinect sensor. Other less frequently used modalities (usually extracted from RGB video) are gaze [4, 6], facial expression [19], and gestures [8]. One very relevant modality that has only been used in research settings [7, 12], but not in complete systems, is the digital file with the slides used as visual aid for the presentation. The RAP system proposed in this work will include the most important features for its objective (entry-level students): audio paralanguage features, posture, gaze, and presentation slides.

Previous systems provide feedback in two different ways. Some present their feedback in real-time through a series of messages on the screen or through a simulated audience [4, 14, 19]. Others provide off-line feedback through analytics reports [8]. While the real-time feedback has the advantage of making the user aware of the moment when they fail, it can be intrusive and could derail the presentation flow. The RAP system will use an off-line system paired with recordings of examples of good and bad behavior to obtain the best of both worlds.

The main contribution of the RAP system to the state-of-the-art in automatic feedback of oral presentation is 1) the use of non-intrusive very low-cost sensors that could provide the same level of modalities than more complex systems with a similar level of quality, 2) the analysis and feedback of the presentation slides and 3) an off-line report system with referenceable recordings that incorporate the benefits of real-time feedback without its disadvantages.

3 DESIGN

The main objective of the RAP System is to provide an effective, sustainable and scalable learning environment to train oral presentation skills for entry-level students. The system is low cost and open source, and was not designed to be a research experiment, but to be a system viable for its wide-spread use in academic institutions. To accomplish this, the RAP system is guided by the following design principles:

(1) **Focus on intended use**: The system provides simple and appropriate feedback to entry-level students. The focus will be on the most common and basic errors made by novices (e.g. not looking at the audience, using too small font-size).
(2) **Immersive experience and unobtrusive**: The system must simulate a real oral presentation setting and environment. Users shouldn't have to use, wear, or made aware of any device, sensor or contraption used to capture data.
(3) **Plug and Play**: Users must be able to walk-in, present and obtain feedback. Except for the presenter, no other humans should be needed to operate the system.
(4) **Relatable feedback**: The feedback provided by the system should be objective, easy-to-understand and unambiguous. All feedback should be paired with recordings of the desired/undesired behavior from the actual presentation.
(5) **Low-cost and scalable**: Each institution should be able to provide several RAP environments for their students. One of the objectives is to make the system as low-cost and easy to deploy as technically possible.

Figure 1 depicts the latest design of the RAP system recording environment. This environment is embedded into a small room providing the presenter with 16 m^2 of space to move freely, a presentation display with an associated computer where the students upload their presentation and provide contact information to deliver the automated feedback. Another projection screen in front of the student displays a pre-recorded audience to provide a sense of immersiveness. The only two sensors used, a camera and a microphone, are camouflaged within the pre-recorded audience screen. The sensors are connected to a second computer that conducts the analysis of the different streams and builds the automatic report.

3.1 Sensor Hardware

The presentations are recorded using two media streams: audio and video. Audio is recorded using an omnidirectional microphone located in the upper part of the audience screen (visible on Figure 1c), pointing at the presenter. To capture video, we placed a Raspberry Pi 3 in the center of the audience screen, with a Raspicamera v2.1 of 8MP attached. The camera covers the entire presentation area at a resolution of 1080x720. Figure 1b shows the positioning of the sensors.

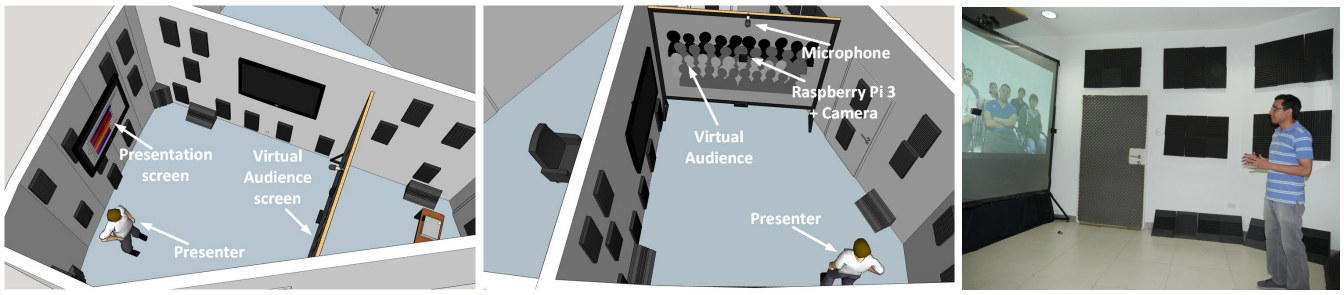
3.2 Usage

The presentation screen displays an application where the participants can control the system on their own. It allows them to view the state of the recording devices, and start/stop a presentation. To start a presentation, the user has to write its name and email on the application, next he has to plug in a USB drive to select the slides to load, and once he is ready, press the "Start Presentation" button. The system will load the slides on the presentation screen and will start the recording. The audience will appear in front of the presenter, which signals that the presenter can start the presentation. When the slides are over, the system automatically stops the presentation and the audience disappears. The URL address of the feedback report will then be emailed to the presenter.

4 FEATURE EXTRACTION

The RAP system extracts features intended to be used in the feedback report. The features were selected to address the main errors that entry-level students made during presentations: not looking at the audience, having a close posture (e.g. hands-in-pockets), overuse of filled pauses, speaking too low, and using too much text or very small font on their slides [11]. All these features were extracted from the multimodal streams, video captured with a low-cost camera, audio captured with a directional microphone, and slides used by the presenter. An overview of all of these features is detailed in the subsections below.

A random sample of real recordings used to evaluate the feedback of the system (see description in section 6) were also used to evaluate the precision and recall of the feature extraction process.



(a) Environment distribution (b) Positioning of Sensors (c) Implementation test

Figure 1: The recording environment has a touch screen to allow the user to display and control the presentation (a), a set of sensors to record the presentation (b), and a large screen to project a pre-recorded live audience (c).

4.1 From Video

Video from the presentation is captured at 5fps and 1080x720 resolution, using a Raspicamera. Each frame is streamed to a server, which uses the OpenPose C++ Library[5] to extract coordinates of face keypoints and body joints of the person during the presentation. The use of this library enables the use of a simple RGB camera instead of a more costly depth sensor, such as the Microsoft Kinect to extract the student's skeletal joints.

4.1.1 Body posture. The position in which the presenter holds her body plays an important role during a presentation. Learning to have an open body posture is a necessary skill for an oral presentation [15]. Using the location of the joints from the body of the presenter, such as the location of wrists, elbow, shoulder, the RAP system classifies the body posture of the presenter as **GOOD** or **BAD**. 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) according to [11]. To evaluate the technical quality of the extraction of the body posture from the video signal, an independent human evaluator reviewed 6 random selected frames from each recording. Three of them classified by the system as good and three as bad. The classification reaches a precision of **0.80** (ratio of frames that the system classified as "Bad" and the human classified also as "Bad") and a recall of **0.93** (ratio of frames that the human classified as "Bad" and the system also classified as "Bad").

4.1.2 Gaze. The gaze is defined as the aim of the stare while presenting, which is an important signal of attention. Some researchers have interpreted that gaze directed at the listener is related with the rapport or the closeness between each other [1]. A purposeful eye contact can have a great impact in engaging the listener attention and consequently the correct reception of the knowledge given by the presenter. The gaze detection system detects the head pose orientation as a proxy for gaze. In the distances involved, gaze and head orientation only differ in a minor way [17]. 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. Any gaze to the audience in front of the presenter is considered as a good gaze, while a gaze to any other direction is considered as bad. To evaluate the technical quality of the extraction of the gaze from the video signal, an independent human evaluator reviewed 6 random selected frames from each recording. Giving that the camera is located in the middle of the simulated audience, it is easy to determine when the presenter has its gaze on the audience. The system achieved a precision of **0.85** and a recall of **0.94** in its classification of a "Bad" gaze.

4.2 From Audio

The audio of the presentation is recorded in 10 seconds audio segments to reduce processing times. A Python script was developed for recording the direct input from the microphone and Praat[1] scripts were developed to measure the volume levels and detect filled pauses.

4.2.1 Filled pauses. Being one the most common speech disfluencies, the Filled Pause (FP) detection has been an important step in order to evaluate the fluency skills of a presenter [2]. An automatically FP detection system was implemented taking as a reference the formant-based technique reported by Audhkhasi et.al [2]. In the aforementioned technique, it is demonstrated that vocal tract resonances, i.e the formants, remain stable when a filled pause occurs. As part of the processing, the entire audio segment is analyzed in smaller windows of 10 milliseconds. For every window, the first and second formant values (F1 and F2) are extracted, their stability is analyzed over time and finally automatically tagged as filled pause or normal audio. Additionally, to evaluate the technical extraction quality, the audio of every presentation was tagged manually by an independent human evaluator and contrasted with the filled pauses detected by the system. As a result, a precision of **0.87** was found, which indicates that most of the FPs tagged were correct. However, the recall value was **0.27**, indicating that a high number of filled pauses were not detected by the system. *A posteriori* analysis indicates that low voice levels had a direct impact on the value of the recall. While the recall can be improved, the high precision of FP extraction enables the use this feature in the feedback report.

[1]Praat speech processing software http://www.fon.hum.uva.nl/praat/

4.2.2 Volume. A good voice volume during a speech is an important feature when presenting because it allows everybody in the audience to hear and understand what is being said. Even though it is not necessary to shout at the audience, it is still important to perform an adequate volume level [11]. Before any processing, a volume test was performed in the room. 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 FP detection, the audio segments were divided into 10 milliseconds 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. To evaluate the accuracy of this feature estimation, the automatic volume classification of every presentation was contrasted with an independent human-based volume classification. Five random segments of 10 seconds per presentation were rated by a human evaluator. The total accuracy of the system was **0.87** and the total recall was **0.81** (between HIGH, LOW and SILENCE).

4.3 From Slides

Visual aids are an integral part of academic presentations by entry-level students. The RAP system also analyzes the digital files of the slides used during the presentations. 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 [12] and [7], that 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. The digital file is sent to a remote server where it is unpacked and every slide analyzed while the user is performing the presentation. After the analysis, it outputs a score of 0 (bad), 1 (regular), and 2 (good) for each feature. The threshold was calibrated using the work of [12]. The scores presented by the system for 20 of the 72 randomly selected slide presentations were compared with scores obtained by an independent human evaluator. To facilitate the evaluation the score of 1 was considered BAD and 2 and 3 as GOOD. As a result, the achieved values were: precision **1.0** and recall **0.84** for font size, precision **0.78** and recall **0.78** for text quantity, and precision **0.75** and recall **0.92** for slide contrast, all of them measured for the "Bad" score.

5 FEEDBACK REPORT

After the students finished their presentations, they received an email with a link to a website that displayed the automatic feedback on their presentations. The objective of this report is to confront the student with evidence (in the form of multimodal recordings) of the eventual errors that they had.

The report starts with an automatically generated summary of the overall quality of the presentation. A global score is provided, by averaging all the individual scores of each modality. The full video of the presentation can be seen in this section. This introduction also contains pre-generated advice to improve the perceived weaknesses. The report continues with a detailed score for each analyzed mode.

The report of each modality presents a small recording (video, images, or audio) of good and bad examples taken from the student

Figure 2: The gaze section of the feedback presented to the student.

presentation and slides. Each modality is scored on a 5-point scale between "Very Good" and "Very Bad". The thresholds for each range in the scale for each modality were calibrated with the help of an oral communication expert. For gaze and volume, also some graphs are presented to provide feedback about the moments when the good or bad behavior happened and the persistence of that behavior (See figure 2)

6 SYSTEM INITIAL EVALUATION

To evaluate the capacity of the RAP system to provide meaningful and reliable feedback, 83 entry-level students of a Computer Engineering program performed a training presentation session in the system and obtained its automated feedback report. The language of the presentation was Spanish. Their ages oscillate between 18 and 22 years. From the total population, 22% are female (reflecting a very similar female/male ratio as the program). All of them filled a user experience survey and 9 of them were later interviewed about their experience.

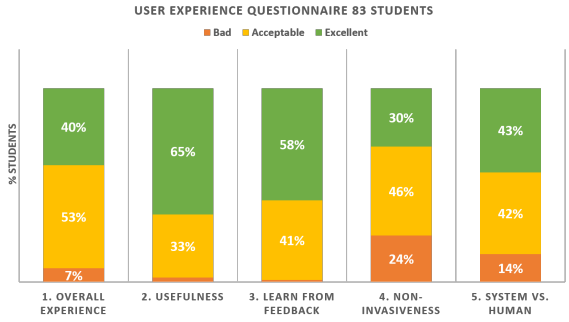


Figure 3: Summary of user experience questionnaire results: All 5 questions were answered on a scale from 1 to 10. An answer from 1 to 5 is considered "Bad", 6 to 8 "Acceptable", and 9 to 10 "Excellent".

6.1 Feedback Accuracy

The feedback provided by the system was compared against the consensus of 3 human reviewers, two of them which were students of a Master Degree in pedagogy, and one professor of Computer

Science. The human reviewers graded the presentation using a scoring rubric aligned with the same features extracted by the RAP system. The grades of the human reviewers and the RAP system were mapped into three categories "Bad", "Regular" and "Good". The agreement percentage between the system and human evaluations were as follows: posture 65%, gaze 78%, filled pauses 75%, voice level 71%, font size 46%, text quantity 44%, and slide contrast 59% (against a 30% of by-chance agreement). The video and audio extracted features present a high level of agreement with human evaluators. The poor results of the slides analysis can be explained by the fact that it was common for students to use screenshot images of their work in their slides; the RAP system does not evaluate the content of images while humans considered them into their evaluation.

6.2 Student Perception

The user experience of the RAP system was measured by using a questionnaire and in-depth interviews. The questionnaire was administered to all 83 participants and the individual interviews to nine randomly selected participants. Figure 3 presents the results of the questionnaire. It revealed an overwhelmingly positive perception of the system especially in the dimensions of perceived usefulness and feedback which were rated as excellent by 65% and 58% of the students respectively. The qualitative analysis helped discover specific issues, on the positive side, students commented on the potential of the system to quickly learn some basic presentation skills: *"I would like to see this system used in our Communications class"*. On the negative side, students commented that they sometimes were aware that they were being recorded and that the environment was too small. Also, some students felt uncomfortable with a pre-recorded audience because it didn't seem to react to their presentation: *"the audience had always the same expressions"*. Overall, the students agreed that the system was useful and that they learned about their own presentation skills while using it.

7 CONCLUSION AND FURTHER WORK

The main conclusion of this work is that an affordable system, using only a camera and a microphone, is able to provide feedback to avoid common errors in oral presentations to entry-level higher education students, and that all of the technical components function properly. The initial evaluation of the feature extraction and the automatic feedback is similar to what is reported as a result of more complex systems [4, 8]. However, there still some technical issues to solve, for example, increasing the recall of filled pauses and analyzing images containing text embedded into slides. The qualitative analysis of the student perception is highly encouraging, being the quality of the feedback one of the most positive aspects of the system. Using a bigger room and reactive audience (like in [4]) could help to improve the main concerns of the students. Due to its goal to improve oral presentation skills, the main further activity for this work-in-progress is the evaluation of the impact that the provided feedback has in the actual learning of the skills. This evaluation will consist of a series of practice sessions of the same students, interlaced with human evaluations during the course of one semester. The analysis of the number of errors, the human-assigned grades and the perception of professors should be a full evaluation of the effectiveness of the system.

REFERENCES

[1] Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).

[2] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma. 2009. Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4857–4860. https://doi.org/10.1109/ICASSP.2009.4960719

[3] Lucas Azaïs, Adrien Payan, Tianjiao Sun, Guillaume Vidal, Tina Zhang, Eduardo Coutinho, Florian Eyben, and Björn Schuller. 2015. Does my Speech Rock? Automatic Assessment of Public Speaking Skills. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[4] Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2013. Cicero-towards a multimodal virtual audience platform for public speaking training. In *International Workshop on Intelligent Virtual Agents*. Springer, 116–128.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[6] Lei Chen, Chee Wee Leong, Gary Feng, Chong Min Lee, and Swapna Somasundaran. 2015. Utilizing multimodal cues to automatically evaluate public speaking performance. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 394–400.

[7] Vanessa Echeverria, Bruno Guaman, and Katherine Chiluiza. 2015. Mirroring Teachers' Assessment of Novice Students' Presentations through an Intelligent Tutor System. In *Computer Aided System Engineering (APCASE), 2015 Asia-Pacific Conference on*. IEEE, 264–269.

[8] Tian Gan, Yongkang Wong, Bappaditya Mandal, Vijay Chandrasekhar, and Mohan S Kankanhalli. 2015. Multi-sensor self-quantification of presentations. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 601–610.

[9] Fasih Haider, Loredana Cerrato, Nick Campbell, and Saturnino Luz. 2016. Presentation quality assessment using acoustic information and hand movements. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2812–2816.

[10] Steve Johnson, Sarah Veitch, and Silvia Dewiyanti. 2015. A Framework to Embed Communication Skills across the Curriculum: A Design-Based Research Approach. *Journal of University Teaching and Learning Practice* 12, 4 (2015), 6.

[11] Stephen E Lucas. 1999. Teaching public speaking. *Teaching communication: Theory, research and methods* (1999), 75–84.

[12] Gonzalo Luzardo, Bruno Guamán, Katherine Chiluiza, Jaime Castells, and Xavier Ochoa. 2014. Estimation of Presentations Skills Based on Slides and Audio Features. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics*. ACM, Istanbul, Turkey, 37–44.

[13] Xavier Ochoa. 2017. *Handbook of Learning Analytics*. Society for Learning Analytics Research (SoLAR), Chapter Multimodal Learning Analytics, 129–141. https://doi.org/10.18608/hla17 978-0-9952408-0-3.

[14] Jan Schneider, Dirk Börner, P Rosmalen, and Marcus Specht. 2017. Presentation Trainer: what experts and computers can tell about your nonverbal communication. *Journal of Computer Assisted Learning* 33, 2 (2017), 164–177.

[15] Aaron W Siegman. 1987. *Nonverbal behavior and communication*. Psychology Press. 37–64 pages.

[16] Sarah Stawiski, Amy Germuth, Preston Yarborough, Vernal Alford, and Leotis Parrish. 2017. Infusing Twenty-First-Century Skills into Engineering Education. *Journal of Business and Psychology* 32, 3 (2017), 335–346.

[17] Rainer Stiefelhagen and Jie Zhu. 2002. Head Orientation and Gaze Direction in Meetings. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, USA, 858–859. https://doi.org/10.1145/506443.506634

[18] Bernie Trilling and Charles Fadel. 2009. *21st century skills: Learning for life in our times*. John Wiley & Sons.

[19] Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. 2015. Multimodal public speaking performance assessment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 43–50.