Multi-sensor Self-Quantification of Presentations

Tian Gan[1,2,3], Yongkang Wong[2], Bappaditya Mandal[3], Vijay Chandrasekhar[3],
Mohan S. Kankanhalli[1,2]

[1]School of Computing, National University of Singapore
[2]Interactive & Digital Media Institute, National University of Singapore
[3]Institute for Infocomm Research, Singapore

gantian@comp.nus.edu.sg, yongkang.wong@nus.edu.sg,
{bmandal,vijay}@i2r.a-star.edu.sg, mohan@comp.nus.edu.sg

ABSTRACT

Presentations have been an effective means of delivering information to groups for ages. Over the past few decades, technological advancements have revolutionized the way humans deliver presentations. Despite that, the quality of presentations can be varied and affected by a variety of reasons. Conventional presentation evaluation usually requires painstaking manual analysis by experts. Although the expert feedback can definitely assist users in improving their presentation skills, manual evaluation suffers from high cost and is often not accessible to most people. In this work, we propose a novel multi-sensor self-quantification framework for presentations. Utilizing conventional ambient sensors (i.e., static cameras, Kinect sensor) and the emerging wearable egocentric sensors (i.e., Google Glass), we first analyze the efficacy of each type of sensor with various nonverbal assessment rubrics, which is followed by our proposed multi-sensor presentation analytics framework. The proposed framework is evaluated on a new presentation dataset, namely NUS Multi-Sensor Presentation (NUSMSP) dataset, which consists of 51 presentations covering a diverse set of topics. The dataset was recorded with ambient static cameras, Kinect sensor, and Google Glass. In addition to multi-sensor analytics, we have conducted a user study with the speakers to verify the effectiveness of our system generated analytics, which has received positive and promising feedback.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis; I.5.4 [**Applications**]: Signal processing

Keywords

Quantified Self; Multi-modal Analysis; Presentations; Sensors; Egocentric Vision

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806252.

1. INTRODUCTION

Presentation is one of the most important methods to convey ideas to an audience, where the ideas have generally been researched, organized, outlined and practiced [41]. The circumstances of a presentation can range from a public speech to an academic seminar. Studies have shown that effective oral communication skills are important in a variety of areas, such as politics, business, and education [9]. Similarly, nonverbal communication, such as gesture, facial expression, posture, and interaction with the audience, also plays a predominant role in effective delivery [33]. Nowadays, presentation software (e.g., PowerPoint, Keynote, *etc.*) is widely used to create quality slides and content for a presentation. Nevertheless, the speaker's presentation skills are still critical to convey ideas.

A bad presentation could be a result of speech anxiety, lack of confidence, insufficient preparation, communication apprehension, lack of practice, etc. Studies from the clinical psychology show that a good presentation is "not a gift bestowed by providence on only a few rarely endowed individuals" but rather a skill to be taught and learned [12]. In order to improve presentation skills, many works in the communication literature have designed various scoring rubrics as guidance for presentation evaluation [9, 24, 25, 29, 32, 36]. Cognitive learning theory suggests that the feedback from an expert facilitates deliberate practice, and these trial-and-error attempts allow for the successful approximation of the target performance [23]. These assessments can be used for individual diagnostic purposes, where this feedback loop serves as effective information for training in making of effective presentations [2, 12]. In spite of that, the manual assessment process requires a human evaluator which is not always feasible in most real-world settings.

In recent years, the advancement of sensor technologies has enabled the development of automated presentation analytics algorithms. These algorithms are designed for various ambient type of sensors, such as microphone, RGB camera, depth sensor, *etc.*, and can be categorized into single modality analysis and multi-modal analysis. Examples of single modality analysis include speech fluency analysis [1] and speech rate detection [8], whereas works on multi-modal analysis include body language analysis with RGB camera and depth sensor [6, 7, 44]. Recently, wearable sensing devices have enabled both opportunities and challenges for user behavior analytics [13, 16, 20]. These devices are equipped with multiple sensors, which include First-Person-View (FPV) visual sensor, microphone, proximity sensor, ambient light sensor, accelerometer, and magnetometer. For example, wearable fitness devices are being increasingly deployed to record the physical activity of a user, where a comprehensive activity report (i.e., quantified self) is automatically generated [15]. In contrast, the use of wearable sensing device has not yet been explored for self-quantification of presentations. This is in spite of the fact that a wearable sensor will provide a constraint-free setting for the speaker's movement, which makes it an ideal device for self-quantification of presentations.

In this work, we propose a multi-sensor self-quantification framework for presentations, where the framework can work with only a wearable sensor or can be combined with existing ambient sensors for improved precision. To the best of our knowledge, this is the first time that the wearable sensor is used to quantify the performance of presentations. Our contributions are as follows:

- We review the past studies in communication, cognitive science, and psychology along with the speech analysis literature, and formalize an assessment rubric suitable for presentation self-quantification.

- We propose a multi-sensor analytics framework for presentations, which analyzes both the conventional ambient sensors (audio, visual, and depth sensor) and wearable sensors (audio, visual, and motion sensor). We quantitatively evaluated our proposed framework on the assessment rubric under single sensor and multi-sensor scenarios. These findings provide an insightful benchmark for multi-sensors based self-quantification research.

- We recorded a new multi-sensor presentation dataset, namely NUS Multi-Sensor Presentation (NUSMSP) dataset, which consists of web cameras, Kinect sensor, and multiple Google Glasses. It consists of 51 presentations of varied durations and topics. In addition, we manually annotated each presentation based on the proposed assessment rubric. The dataset is now publicly available for the research community.

- We have conducted a user study with the presenters in this dataset. For each presenter, we provided our system generated feedback and then the presenter was asked to verify the effectiveness of this feedback. The study shows positive results of our proposed system and provides several useful insights for future research.

The remainder of the paper is organized as follows. Section 2 provides an overview of the related literature for presentations. Section 3 provides the assessment rubrics for self-quantification of presentations. Section 4 elaborates on the sensor configuration and the proposed analytics framework. Section 5 contains the details of new dataset, experimental results and discussion, whereas Section 6 discusses the feedback from the user study. Section 7 concludes the paper.

2. RELATED WORK

In the psychology of learning, presentation in a small group or large public environment is one of the well-studied areas in the last few decades [5, 9, 12, 24, 25, 29, 32, 36]. Generally, the communication skills of a presentation are often assessed using certain rubrics [5, 9]. In the late 1970s, the National Communication Association (NCA) conducted a large scale study to identify the core competencies (including speaking and listening skills) for students. Quianthy [29] identified eight competencies: purpose determination, topic selection, organization, vocal variety, articulation, nonverbal behavior, language use, and use of supporting material. Following the study in [29], Morreale *et al.* [25] developed the "Competent Speaker Speech Evaluation Form", which evaluates eight items in a two-stage assessment process (i.e., *preparation and content* and *presentation and delivery*). Several other assessment rubrics have also been separately developed by different research groups [24, 32, 36]. Across these assessment rubrics, the core competencies only differ subtly where several items were adjusted to meet the respective analytic requirements [32].

In the computing literature, a variety of computational models have been proposed to analyze various types of competencies in presentation delivery, e.g., speech rate measurement [8], speech liveliness measurement [17], and social phobia analysis [34]. Kurihara *et al.* [19] proposed a presentation training system, which analyzes the speaking rate, eye contact with the audience, and timing during the presentation. The proposed system consists of only two sensors: "microphone" and "web camera". As the performance of the training system is mainly restricted by the analysis algorithms, the early prototype required the presenter to wear a special visual marker over the head to enhance the performance. Pfister and Robinson [28] proposed a system to analyze the speech emotion for the same application. The audio-based system focuses on the analysis of the various types of speech emotions (i.e., competent, credible, dynamic, persuasive, and pleasant). Recently, more modalities have been included in the analysis, especially the depth channel from Kinect sensor due to its robustness in tracking human body's motion. Several researchers have exploited the multi-modality data from the visual data, audio data and depth information [6, 7, 10, 27]. Nguyen *et al.* [27] used the Kinect sensor to recognize the bodily expressions and provide feedback on a scale of five degrees (i.e., bad, not bad, neutral, good, and excellent). Similarly, Echeverría *et al.* [10] proposed the use of the same sensor to grade the presenters' performance using eye contact score and body posture language score. Chen *et al.* [6] presented their initial study on the development of an automated scoring model, where they predict a singular score based on the analysis of the multi-modal features. In comparison, their later work [7] provides scores for the *delivery skills* and *slides quality.*

The technological advancements in microelectronics and computer systems have enabled new sensors and mobile devices with unprecedented characteristics. One of the new categories is the wearable sensing device, which has reduced size, weight and power consumption, and is generally equipped with multiple sensors. Examples of wearable sensing device include Fitbit, smartwatch, GoPro, and Google Glass. In contrast to the aforementioned sensors, denoted as ambient sensors in this work, the wearable sensor enables high precision in tracking the user's motion, and allows for continuous usage for daily activities [16]. For example, the skeleton data extracted with depth sensor is unreliable if the profile view of a user is given. But the Fitbit can provide activity data continuously. Another key difference resides in how the user interacts with the sensor [20]. The ambient sensors are pre-configured with a pre-determined region-of-interest, which restrict user interaction in a specific spatial location [14]. In contrast, the wearable sensor has no such constraints and user can perform the desired action in any location. There arise several new research problems with the wearable sensors. Ermes *et al.* [11] proposed the use of wearable sensors for detecting daily activities and sports under both controlled and uncontrolled conditions. Similarly, Hernandez *et al.* [16] estimates the physiological signals of the wearer using a head-mounted wearable device. Gan *et al.* [13] proposed a framework that used multiple egocentric visual sensors to recover the spatial structure of a social interaction. To the best of our knowledge, this is the first time that the wearable sensor has been used to quantify the performance of presentations.

3. THE ASSESSMENT RUBRIC

In this section, we detail the assessment rubric for multi-sensor self-quantification of presentations. Different from the assessment rubrics in the literature, the new rubric does not contain high level semantic concepts such as *topic selection* and *organization of ideas*, which makes it more suitable for data-driven analytics with sensors. This is motivated by the intention to make such self-quantification process automated, cheap yet useful. In the following sections, we first provide the overview of the proposed assessment rubric, followed by detailed discussion of each category.

3.1 Overview

In the psychology of learning and cognitive learning literature, the evaluation of presentation skills is always associated with the guidance of an assessment rubric [9, 24, 25, 29, 32, 36]. A rubric is a coherent set of criteria that includes descriptions of levels of performance quality on the criteria [5]. The human evaluator, based on speaker's behavior and the rubric, will decide the presentation quality and provide feedback to the speaker.

The computing literature follows a similar process and provides a *score* for each concept [6, 7, 10, 27]. However, these scores do not provide sufficient semantic cue to the speaker. For example, the system provides a speaking rate of 2 rather than a semantically meaningful label like "slow". Therefore, we have reviewed the prior work in the literature, and have proposed a new assessment rubric which is not only semantically meaningful, but is also more suitable for automated sensor-based analytics algorithms. The overview diagram of the proposed assessment rubric for multi-sensor self-quantification of presentations is shown in Figure 1.

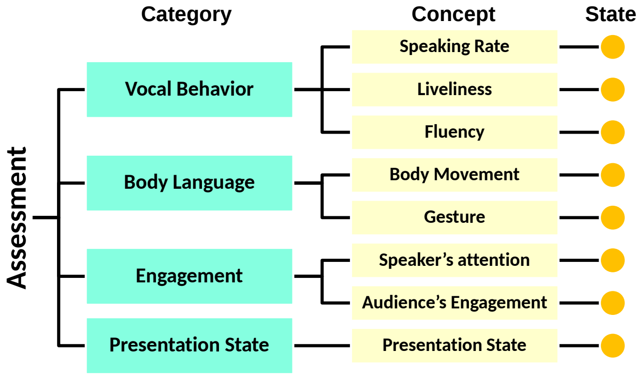

Figure 1: Proposed assessment rubric for multi-sensor self-quantification of presentations.

The proposed assessment rubric consists of a three layered hierarchical structure, namely **category**, **concept** and **state**. The category layer contains the high level separation of behavior type in presentation, which consists of *vocal behavior*, *body language*, *engagement*, and *presentation state*. The concept layer further segments each category into a more detailed behavior. For example, the vocal behavior category contains the *speaking rate*, *liveliness*, and *fluency* concepts. The state layer provides the semantically meaningful state/class for each concept. For example, the gesture concept can be divided into three states (i.e., *normal*, *excessive*, and *insufficient*). The detailed descriptions can be found in the next section.

3.2 Assessment Categories

3.2.1 Vocal Behavior

Presentation skill is multifaceted in nature, including lexical usage, fluency, pronunciation, and prosody [7]. This work focuses on the nonverbal vocal behaviors, where the prosodic features (e.g., pitch, tempo, energy, *etc.*) correspond to the voice quality [37]. We have identified three concepts which are frequently used in the assessment rubric [9, 24, 25, 29, 32, 36]: *speaking rate*, *liveliness*, and *fluency*. The speaking rate is a good predictor of the subjective concept fluency and liveliness. Liveliness is defined as the variation in intonation, rhythm and loudness. Fluency is a speech language pathology term that means the smoothness or flow with which sounds, syllables, words and phrases are joined together when speaking quickly. These three cognitive concepts can be interpreted in the computational measurement such as the number of syllables per minute, variation in pitch, and the number of filled pauses per minute. In our work, we quantify these concepts into three states: *insufficient*, *normal* and *excessive*.

3.2.2 Body Language

Body language is a form of nonverbal delivery method to strengthen the messages during presentation [18], where the messages are expressed by physical behavior, such as facial expression, body posture, gesture, and eye contact. As the facial analysis techniques are still far from perfect for real-world applications [43], we deliberately exclude facial expression and eye contact in this work. In addition, the speaker is often far away from the audience, resulting in low facial image resolution in the video footage. Two concepts, namely *body movement* and *gestures*, are included in the proposed rubric. The body movement relates to the usage of space and posture of the body. On the other hand, gestures, which are movements of the head, hands, and arms, can be used to convey specific messages that have linguistic translations. In our work, we quantify these concepts into three states: *insufficient*, *normal* and *excessive*.

3.2.3 Engagement

Engagement with audience in training or educational presentations is the key factor for effective idea delivery [38].

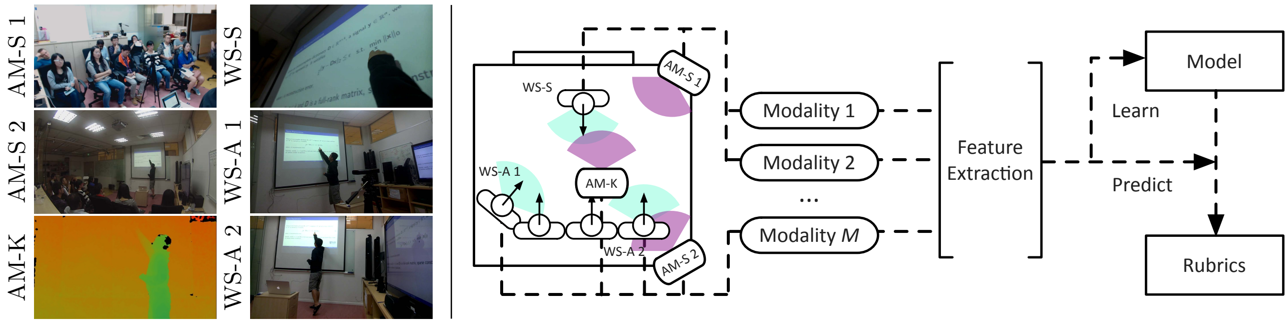

Figure 2: The sensor configuration and the proposed framework. AM-S and AM-K are the ambient static camera and ambient Kinect sensor. WS indicate the wearable sensor, where WS-S and WS-A represent WS from speaker and audience, respectively.

In this category, we evaluate both the speaker's and audience's attention, which are useful to characterize the engagement. During the presentation, the speaker may pay attention to the script, the audience, or the computer. Therefore, we list out the most common objects/scenes in a presentation and include an "others" category for completeness. Formally, the states for speaker's attention concept are *audience*, *screen*, *computer*, *script* and *others*. For the audience's engagement, we have formalized three states which are *no attention*, *attention without feedback*, and *attention with feedback*. The feedback can be reflected in the form of behavior like nodding of head to show acknowledgement, or involvement in the interaction between the audience and the speaker. For each state, the classifier will provide a binary decision for the presence of the state.

3.2.4 Presentation State

Question Answering (QA) is the interactive element of presentations. It provides speaker an opportunity to learn the current state of comprehension of the audience, and provides the audience a chance to convey their concerns. For this category, we have designed two states in the proposed assessment rubric, namely *presentation* and *QA*. The *presentation* characterizes the period where the speaker is the dominant person in the presentation, whereas *QA* characterizes the interval where the audience is asking questions or is having some discussion. The analytics on the presentation state mainly reflects the transition and flow of the presentation.

4. PROPOSED METHOD

In this section, we first provide an overview of the sensor configuration, followed by the proposed multi-sensor self-quantification framework and the details of the sensor analytics components.

4.1 Sensor Configuration

We set up our experiment environment in a meeting room with a Kinect sensor (denoted as AM-K) and two static RGB cameras with microphone (denoted as AM-S 1 and AM-S 2). AM-S 1 and AM-S 2 capture the speaker and audience from two different locations, whereas AM-K is configured to capture the behavior of the speaker with both RGB and depth channel. For each presentation, three Google Glasses are deployed, where one is worn by the speaker (denoted as WS-S) and the remaining two are worn by two randomly chosen audience members (denoted as WS-A 1 and WS-A 2). The overview of the sensor configuration and the approximate spatial location of the speaker and the audience are shown in Figure 2.

4.2 Multi-Sensor Analytics Framework

In the context of multi-sensor and multi-modal analysis, we assume that the data from each sensor can be modeled as a collection of feature set $\mathbb{X}$, where each member $\boldsymbol{X}^{m,i}$ is the feature extracted from the m modality and i feature type. In this work, we consider the *acoustic*, *visual*, *depth*, and *motion* modality, where the corresponding features of each modality are presented in the following sections. The feature $\boldsymbol{X}^{m,i}$ is further divided into N segments with an interval of T seconds: $\boldsymbol{X}^{m,i} = \{\boldsymbol{x}_1^{m,i}, \boldsymbol{x}_2^{m,i}, \dots, \boldsymbol{x}_N^{m,i}\}$. Given the extracted feature, the task is to learn the respective classifier $\mathcal{F}_c : \boldsymbol{x}_n^{m,i} \mapsto r_n^c$ to produce r_n^c, where r_n^c is the predicted state/class for the corresponding c concept. The overview of our proposed framework is shown in Figure 2.

4.3 Feature Representation and Classification

4.3.1 Acoustic Feature

Four types of acoustic feature accounting for the speaking style are extracted from raw audio data using the Praat script [3]. The selected acoustic features are *pitch*, *intensity*, *formants*, and *syllables*.

Pitch is an auditory sensation which allows the ordering of sounds on a frequency-related scale. It is defined in the music literature as a stretch of sound whose frequency is clear and stable enough to be heard as not noise [30]. Following the work in [17], we compute the Pitch Variation Quotient (PVQ) feature as follows:

$$\boldsymbol{x}_n^{A,pitch} = \mu(\boldsymbol{p})_n / \sigma(\boldsymbol{p})_n \tag{1}$$

where $\boldsymbol{p}$ is the pitch feature computed from a T seconds interval.

The second feature is called the acoustic intensity and is defined as the sound power per unit area. The sound intensity is highly related to the subjective measure of loudness. The n acoustic intensity feature, $\boldsymbol{x}_n^{A,intensity}$, is defined as the mean and standard deviation of intensity over T seconds. These two features $\boldsymbol{x}^{A,pitch}$ and $\boldsymbol{x}^{A,intensity}$ are used to predict the liveliness concept. In this work, T is set to 10 seconds as it guarantees the inclusion of a fair amount of speech at the normal pausing rates [17].

Formant is a concentration of acoustic energy around a particular frequency in the speech wave. The fluency of speech is related to the presence of filled pauses, which are the hesitation sounds that speakers employ to indicate uncertainty or to maintain control of a conversation while thinking of what to say next. The formant information is used to characterize the fluency of a speech due to its ability of detecting filled pauses [1]. Based on the work in [1], we first compute the formant at a frame rate of 10 ms. Then the standard deviation of the first and second formant, $\sigma(F1)$ and $\sigma(F2)$, are computed over a window of T seconds centered on the current frame. The distribution of the $\sigma(F1)$ and $\sigma(F2)$ are discretized into 51 bins over the range from 0 to 200, which leads to a 51-point probability mass function $F_1 = f_{\sigma_{F1}}(i)$ and $F_2 = f_{\sigma_{F2}}(i)$ for $i = 1, 2, \ldots, 51$. We follow the work [21] to extract the feature $r(F)$, which is the ratio of the sum of the frequencies at the left side of a given frequency point. This feature is extracted for both F1 and F2 when the frequency point is 100Hz, $r_1(F)$, and 40Hz, $r_2(F)$, and is used to predict the fluency concept:

$$\boldsymbol{x}^{A,fluency} = [r_1(F_1), r_1(F_2), r_2(F_1), r_2(F_2)]. \tag{2}$$

The last feature, syllables, is used to predict the state of the speaking rate concept, which is a critical component of speech delivery. The speaking rate can be measured as the total number of words or syllables per minute. Generally, speech with a slower speaking rate is more intelligible than faster one. However, the variation in speaking rate may also contribute to the liveliness of speech [8, 35]. Using the method in [8], we extracted the location of the syllables for audio segment. For each n-th segment, we further count the number of syllables $||syl^n||$ within every one second, and discretize this distribution into 4 bins, which results in:

$$\boldsymbol{x}^{A,speakingrate} = \text{hist}(||syl^n||). \tag{3}$$

In order to detect the presentation state, it is important to understand the dominant role in a conversation (i.e., who is currently speaking?), as well as the identity of the speaker (i.e., who is the presenter and who are the audience?). In this work, we first utilize the open-source speaker diarization toolkit LIUM [31] to learn the identity (i.e., speaker or audience) from the audio data of the whole presentation. Conceptually, we perform an audio clustering where each cluster represents a unique person. Then, under the assumption that the presenter is the person who speaks for the longest duration, we label the audio data from this person as the presenter and the rest are labeled as audience. Given this information, the diarization feature is computed as a binary T-dimensional vector $\boldsymbol{x}^{A,speakerID}$ that indicates the speaker is currently speaking, where the t dimension corresponds to t seconds of the segmented data. This feature is used to predict the presentation state concept.

4.3.2 Visual Feature

The class of scene (e.g., audience, computer, script, *etc.*) from the speaker's FPV image is a good indicator of where the speaker is paying attention to. Given an image set $\mathbb{I}_n$ from the n-th segment, we first extract the pyramid histograms of visual words (PHOW) features [4] from each image and represent it as a Bag-of-Words (BoW) descriptor with vector quantization. Due to the scalability to perform scene classification in each frame, we pool all the frames from each second into a BoW descriptor $\boldsymbol{d}_n^t$. Inspired by the multi-instance learning problem, the n-th video segment is represented as $\boldsymbol{D}_n = [\boldsymbol{d}_n^1, \boldsymbol{d}_n^2, \ldots, \boldsymbol{d}_n^T]$, where each $\boldsymbol{d}_n^t$ is considered as an instance. Adopting a recent multi-instance learning approach [40], the BOW descriptor $\boldsymbol{D}_n$ is mapped into Fisher Vector to represent the n video segment:

$$\boldsymbol{x}_n^{V,scene} = M_f(\boldsymbol{D}_n, p), \tag{4}$$

where M_f is the mapping function proposed in [40], and p is the Gaussian Mixture Model (GMM).

4.3.3 Depth Feature

The recent advancements in depth sensing have created many opportunities for human behavior analysis [44]. By using Kinect sensor and the provided API, the Kinect skeletal tracking can represent the human body by a number of joints such as head, neck, shoulders. The skeleton information is useful in characterizing the behavior of the speaker, for example the body language. To differentiate the body movement and gesture, we use two kinds of features. First, the movement of the body, which is calculated by the average and standard deviation of the displacement of 12 upper body joints[1] *disUpper* over T seconds, are used to describe the body movement; second, an average and standard deviation of relative displacement of the selected joints[2] with respect to the human body center over T seconds are used to describe the gesture. In addition, the facial features like the landmarks of the face and orientation of the face can be obtained more efficiently and robustly with depth images when compared to extracting using only RGB images. Specifically, the head direction vector ori^{XYZ} are used:

$$\begin{aligned}
\boldsymbol{x}^{D,body} &= [\mu(disUpper), \sigma(disUpper)] \\
\boldsymbol{x}^{D,gesture} &= [\mu(disUpper^{relative}), \sigma(disUpper^{relative})] \\
\boldsymbol{x}^{D,head} &= [\mu(ori^{XYZ}), \sigma(ori^{XYZ})]
\end{aligned} \tag{5}$$

4.3.4 Motion Feature

In addition to the visual RGB sensor, Google Glass is also equipped with sensors like gyroscope, magnetometer, and accelerometer. These inertial motion sensors provide information about the device wearer's viewing direction and motion pattern. In contrast with the use of visual information to track human's head, which will fail in the non-frontal view, the sensor data is more accurate and robust. Two features, the mean and standard deviation of the camera viewing direction vector ori^{XYZ}, and the linear acceleration of the camera acc^{XYZ} over T seconds are used:

$$\boldsymbol{x}^M = [\mu(ori^{\text{XYZ}}), \sigma(ori^{\text{XYZ}}), \mu(acc^{\text{XYZ}}), \sigma(acc^{\text{XYZ}})] \tag{6}$$

4.4 Multi-Modal Analytics

For every single feature, we employ machine learning approach to predict the state of the concept. The multi-class SVM using a polynomial kernel was utilized as the machine learning tool. For multiple features, for example x^A and x^B, where A and B can be cross-sensors with the same type of feature, or cross-modalities from the same sensor, decision level fusion is done by training-based super-kernel fusion [42].

[1] The upper joints are HipCenter, Spine, ShoulderCenter, Head, ShoulderLeft, ElbowLeft, WristLeft, HandLeft, ShoulderRight, ElbowRight, WristRight, HandRight, and HipLeft.

[2] ElbowLeft, WristLeft, HandLeft, ElbowRight, WristRight, and HandRight.

5. EXPERIMENTS

In this section, we first describe the new multi-sensor presentation dataset, followed by the quantitative evaluation and discussion of the sensor performance using the assessment rubric.

5.1 Multi-Sensor Presentation Dataset

We have collected a new presentation dataset, namely NUS Multi-Sensor Presentation (NUSMSP) dataset[3], which is designed for experiments in multi-sensor self-quantification of presentations under real-world conditions. The dataset is recorded in a meeting room with two static cameras (with built-in microphone), one Kinect sensor, and three Google Glasses. Figure 2 shows the sensor configuration. The NUSMSP dataset consists of 51 unique speakers (32 male and 19 female), where each speaker was asked to prepare and deliver a 15 minutes presentation with no restriction on the topic and content. The number of audience ranged from 4 to 8. For each presentation, the speaker and two randomly chosen audience were asked to wear Google Glass. In total, we have about 10 hours of presentation data. Due to unforeseen technical issues in the recording devices, a small portion of presentation has not been recorded.

We manually annotated the NUSMSP dataset based on the proposed assessment rubric (see Figure 1 for details). The video data are segmented into multiple clips of 10 seconds duration. To remove the subjective bias in the annotation, each clip was annotated with a minimum of five human annotators and the final groundtruth is determined by majority voting.

5.2 Result and Discussion

In this section, we present the single modality and multi-modality based evaluation on *body language*, *engagement*, and *presentation state* category. The analytics are evaluated with depth information (D), motion information (M), visual information (V), and audio data. As the analytics of vocal behavior category are achieved with existing methods [1, 8, 17], it is not reported in this section. The classification performance is reported with 5-fold cross-validation. In the remaining section, we denote ambient static camera and ambient Kinect sensor as Am-S and Am-K, where the wearable sensor on speaker and audience are denoted as WS-S and WS-A, respectively. An overview of the configuration of sensor type, data modality, and the respective concept are listed in Table 1.

Table 2 shows the classification result on the body language category. The "Body Movement" and "Gesture" concept have three states: "In", *insufficient*; "Norm", *normal*; and "Exc", *excessive*. Am-K's depth sensor data and WS-S's motion sensor data are used for the classification. The depth modality performs the best for the *insufficient* state in both body movement and gesture concept. However, it does not perform well on the *normal* and *excessive* states. The reason is because the *insufficient* state mostly has little amount of joint location displacement, and is comparatively easier to differentiate from *normal* or *excessive* states. In comparison, the distinction between *normal* and *excessive* is less obvious, even in the human annotated ground truth. The motion data performs average on the three states, in which the performance on the gesture concept is comparatively worse than body movement. This is because the motion information from the head-mounted wearable sensors have less relation with the gesture as compared to the body movement. The fusion of depth data and motion data indeed helps in the classification of the *insufficient* and *excessive* state in body movement concept. The poor performance on the *normal* state further validates the ambiguous boundary between two consecutive states.

Table 1: The configuration of sensor type, data modality, and the respective concept to be analyzed. The red, blue, and green columns represent concepts from the body language, engagement, and presentation state, respectively. Am-S and Am-K denote the ambient static camera and ambient Kinect sensor, where WS-S and WS-A represent the wearable sensor on Speaker and Audience, respectively.

Sensor Type	Modality	Body Movement	Gesture	Speaker's Attention	Audience Engagement	Presentation State
Am-S	Visual					
	Audio					✓
Am-K	Depth	✓	✓	✓		
WS-S	Motion	✓	✓	✓		
	Visual			✓		
	Audio					✓
WS-A	Motion				✓	
	Visual				✓	
	Audio					✓

Table 2: Average classification accuracy on the body language category. D and M represent Am-K-Depth and WS-S-Motion, respectively.

Modality	Body Movement			Gesture		
	In	Norm	Exc	In	Norm	Exc
D	81.28	26.73	36.28	80.13	38.59	30.65
M	52.07	42.04	65.13	48.58	53.77	34.18
D+M	84.33	10.10	63.46	67.69	29.73	51.27

The results on speaker's attention are shown in Table 3. The classification on four states *script*, *audience*, *computer*, and *screen* are evaluated. The head tracking information extracted from Am-K's depth sensor and the visual and motion information from WS-S are used for the classification. In general, the visual modality performs best among the three modalities. The superiority is less prominent for the *audience* state. This is because compared to *script*, *computer*, and *screen* state, the *audience* state has more visual variations. The depth information performs the worst among these three modalities. This is because the depth head tracking which provides the head orientation works well only when the human shows the frontal view. The performance on the motion sensor data which also provide the orientation information is much more accurate than on the depth data. The fusion of these modalities

[3] http://mmas.comp.nus.edu.sg/NUSMSP.html

Table 3: Average classification accuracy on the speaker's attention concept. D, V, M represent Am-K-Depth, WS-S-Visual and WS-S-Motion, respectively.

Modality	Script	Audience	Computer	Screen
D	65.83	56.35	62.67	66.15
V	96.25	78.94	86.21	85.03
M	77.36	71.97	66.30	79.44
V + M	95.97	82.71	84.26	83.36
D + V	95.97	82.71	84.96	84.76
D + M	77.36	73.08	67.13	80.28
D + V + M	95.97	82.71	84.54	83.50

Table 4: Average classification accuracy on the audiences' engagement concept. V and M represent WS-A-Visual and WS-A-Motion, respectively.

Modality	No	Yes without Feedback	Yes with Feedback
V	59.28	63.73	65.07
M	42.94	68.95	65.80
V+M	68.37	62.85	64.48

performs similar with their corresponding single modality on *script*, *computer*, *screen* state, in which V is the dominating modality. In comparison, fusion on *audience* state helps improve the performance.

Table 4 reports the result of the audience's engagement. The three states are *no*, *yes without feedback*, and *yes with feedback*. The performance on the *no* state is worse than the other two states. This is because the audience mostly focus on the other audience, screen, or are engaged in the discussion with speaker and other audience mentioned in the *yes* states. In comparison, the *no* state has more variations. The fusion of V and M helps improve the performance on *no* state, with slight decrement on the two other states.

The presentation state is detected based on the speaker diarization information using the audio modality: the audio from the ambient static sensor, the speaker's wearable sensor, and the audience's wearable sensors. As shown in Table 5, the ambient sensor's performance on the "QA" state is significantly worse than the performance using wearable sensors' data. By analyzing the result we found out that the audio from the speaker's and audience's FPV sensor achieves better performance on the speaker diarization than the ambient sensor. The speaker's and audience's FPV sensor helps diarize the speaker and audience better.

To summarize, in terms of the cross-sensor analysis, based on the sensor configuration in Table 1, the speaker's wearable sensor WS-S can classify the speaker's attention and the presentation state well. The performance on the body language is not satisfying on our selected head-mounted wearable sensor. This may be improved by other new types of sensors such as smart watches. The conventional ambient sensors can also classify the same set of concepts, with better performance on analyzing human body language during presentations. However, the performance of the ambient sensors on tracking the speaker's attention is poor, due to its speaker frontal view restriction. Also, the performance on the presentation state is lower when using the ambient audio data when compared to the wearable sensors' audio data. Due the low facial image resolution in the speaker view camera, conventional methods that detects and tracks audience's faces to predict audience's engagement are unreliable in the unconstrained environment [22, 26, 39]. In contrast, wearable sensors on the audience can provide more accurate and detailed information about the audience's engagement state and the presentation state. However, the audience's engagement is based on the selected audience who wear the sensor. It is difficult to equip every audience with a wearable sensor and to expect them to share their data.

Table 5: Average classification accuracy on presentation state.

Modality	Presentation	QA
Am-S-Audio	87.31	58.16
WS-S-Audio	81.19	80.91
WS-A-Audio	80.81	80.75

6. USER STUDY

In this section, we elaborate the details of user study with the speakers in our dataset. We first describe the system generated self-quantification report, followed by discussion of the presenters' feedback.

6.1 Analytics

Using the analytic model for each proposed assessment rubric, a system generated analytics report is generated for each presentation. The report consists of three main components: **(1)** footage from all video capturing devices, **(2)** time-series analytics, and **(3)** cumulative sum of the analytic result. Snapshots of two system-generated reports are shown in Figure 3.

The top section of the generated feedback shows the video footage from each device. The current frames correspond to the moment highlighted in the time-series analytics section. The bottom left section shows the time-series result, in which the first five concepts, i.e., speaking rate, fluency, liveliness, body movement, and gesture, are shown with the *insufficient*, *normal*, and *excessive* state quantified into three regions (bottom, middle, and top portion for each horizontal bar). The speaker's attention concept is shown as the red occupation mark at each time point. For example, if the speaker is paying attention on *audience* in the t-th segment, the corresponding location on the *audience*'s row will be marked as red. It is common that multiple attention states can co-exist in the same segment, e.g., the audience state and the screen state are marked with attention when the speaker switches between audience and screen. If none of the state in *audience*, *screen*, *computer*, or *script*, the attention will be shown as *others*. The last two concepts are audience's engagement, in which the corresponding states are *no attention*, *attention without feedback* and *attention with feedback* from bottom to top. These concepts (except for the speaker's attention) are highlighted in red, purple, and grey, representing *presentation* state, *QA* state, and unavailable state (this may due to the error of the sensor). The number of the audience's engagement concepts can vary depend on the available audience's FPV devices. The bottom right section shows the cumulative sum of the analytic result.

The analytics feedback allows the user to quickly examine their performance. From Figure 3, we can see that Speaker A and Speaker B have very different behaviors during the presentation. From the time-series results, we notice that

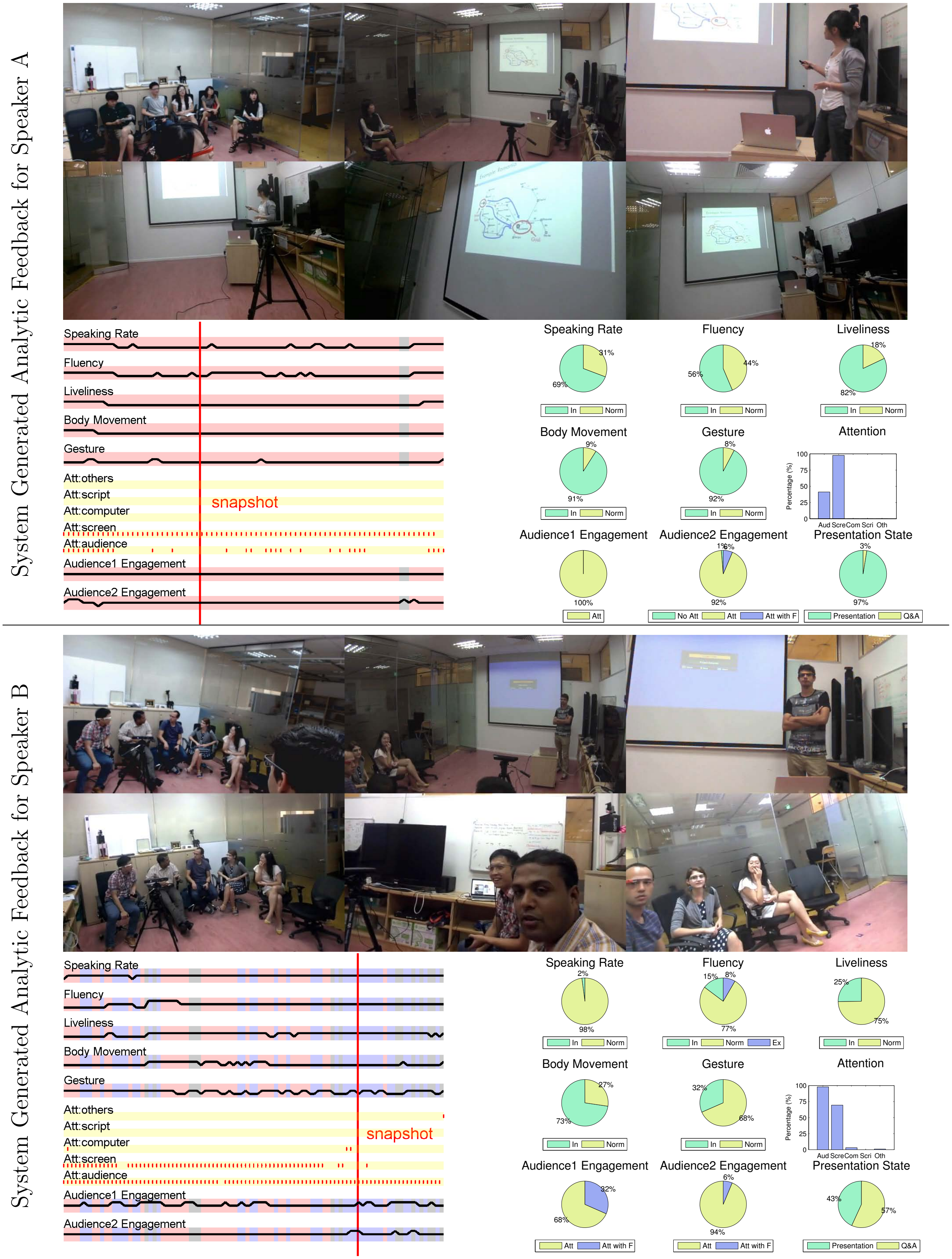

Figure 3: Example of two system generated analytic feedbacks. For each feedback, the six snapshots are taken from the ambient sensors (top row) and wearable first-person-view sensors (bottom row).

Speaker A spent a significant of time on the screen, without interacting or even looking at the audience. In contrast, although Speaker B spent around 70% of the time looking at the screen, we can see that he kept switching between the screen and the audience. Also, Speaker B's presentation involves significant time in the *QA* state. For example, the snapshot in this example is in the state where a question triggers the audience members to discuss among themselves.

6.2 Feedback from the Speakers

In order to further validate the efficacy of our system generated analytics report, we presented the results to the speakers and requested them to complete the following survey:

1. Are the analytics results surprising to you?
2. Are the analytics results useful to you?
3. Can these results help you improve your presentation skills? If so, could you please precisely explain how they can do so?
4. Do you find it intrusive to use the wearable sensor during the presentation? Please elaborate on your experience.
5. If the technology is mature and non-intrusive enough to provide such feedback with good accuracy, would you like to use it?
6. Do you have any other suggestions for improving this analytics feedback?

The first five questions are to be answered using a score from 1 to 5, ranging from "strongly disagree" to "strongly agree". Question 3, 4 and 6 require speaker to elaborate their opinions. In total, we received 21 responses from the speakers, with an average score of 3.59, 4.27, 4.36, 2.79, 4.58 for the first five questions.

Generally speaking, we found that the speakers held a positive attitude towards our summarized feedback. One of the speakers who was "surprised" by the results responded with "I was not aware that I have spent so many time looking at the screen." Most of the speakers found the presentation behavior patterns identified by the system to be accurate, and agreed that the feedback is useful (4.27) as well as helpful for improvement (4.36). Their feedback for the results are: "Avoid excessive gestures and have more eye contact with the audience." By the synchronization with the presentation state, one speaker responded that "Also, maybe more gestures during the presentation (not only during QA session) will be beneficial.", which implies the actual usage of the combination of presentation concept and gesture concept. The opinions about the "intrusiveness" of the Google Glass are mostly neutral or positive, except that three of the speakers criticized the design of the Google Glass: " I just could not help looking at the small screen of Google Glass. It's not comfortable since I have to roll my eyes to one side to see it clearly.", " Very intrusive, the Google glass blocks my view and is very heavy.", "Not convenient especially for people who wear glasses". The answers for trying this type of system are positive (4.58), with all the responses from ranging from agree to strongly agree.

The suggestions from the speakers can be summarized as follows: (1) comparison with the average performance or good performance will be helpful as one speaker mentioned that "A reference frame would be useful, e.g., the average score of all participants or the best score for a good presentation. If I don't know what I am compared with, I won't know how I could improve."; (2) timely feedback after the presentation: "If I am given the results right after the presentation, I can remember what I did to lose my audience's attention, or why I had a bad body expression, etc."; (3) showing the examples of the typical behavior: "Maybe examples of 'insufficient', 'normal', and 'excessive' behaviors will be useful, in form of short video clips?" We can see that the speakers all agree that this kind of information is useful, however, timely feedback immediately after the presentation and reference statistics can help them better interpret the data. Also, they are interested in observing their actual behavior for each concept.

7. CONCLUSIONS AND FUTURE WORKS

In this work, we have proposed a multi-sensors analytics framework to provide the analysis of a presentation. The proposed framework consists of ambient sensors (e.g., visual sensors and depth sensors), First-Person-View (FPV) sensor on speaker, and multiple FPV sensors by the audience. Based on the literature from both social science and computing, we have developed a novel assessment rubric for presentations. We evaluate the performance of existing computational models and analyze the efficacy by combining multiple sensors. To complement the research, we have recorded a new dataset, *NUSMSP*, which consists of 51 subjects giving presentations on diverse topics. We have conducted user studies on the speakers with our generated analytics report, which provides good insights for future research directions.

For the future work, we plan to conduct user studies with our system to measure the individual improvement over multiple presentation sessions. This process will include presentation coaches for validation purposes. In addition, we would like to generalize the current framework with an online learning approach so that the system can adapt to new speakers and new environments efficiently. Last, we aim to improve the visualization of our system generated analytics feedback with richer information.

8. ACKNOWLEDGMENT

This research was carried out at the SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

9. REFERENCES

[1] K. Audhkhasi, K. Kandhway, O. Deshmukh, and A. Verma. Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4857–4860, 2009.

[2] T. W. Banta. *Assessing Student Achievement in General Education: Assessment Update Collections*, volume 5. John Wiley & Sons, 2007.

[3] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International*, pages 341–345, 2002.

[4] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *International Conference on Computer Vision*, pages 1–8, 2007.

[5] S. M. Brookhart and F. Chen. The quality and effectiveness of descriptive rubrics. *Educational Review*, pages 1–26, 2014.

[6] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the International Conference on Multimodal Interaction*, pages 200–203, 2014.

[7] 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 *Proceedings of the ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 45–52, 2014.

[8] N. De Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009.

[9] N. E. Dunbar, C. F. Brooks, and T. Kubicka-Miller. Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, 31(2):115–128, 2006.

[10] V. Echeverría, A. Avendaño, K. Chiluiza, A. Vásquez, and X. Ochoa. Presentation skills estimation based on video and kinect data analysis. In *Proceedings of the ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 53–60, 2014.

[11] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):20–26, 2008.

[12] S. B. Fawcett and L. K. Miller. Training Public-Speaking Behavior: An Experimental Analysis and Social Validation. *Journal of Applied Behavior Analysis*, (2):125–135, 1975.

[13] T. Gan, Y. Wong, B. Mandal, V. Chandrasekhar, L. Li, J.-H. Lim, and M. S. Kankanhalli. Recovering social interaction spatial structure from multiple first-person views. In *Proceedings of International Workshop on Socially-Aware Multimedia*, pages 7–12, 2014.

[14] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli. Temporal encoded F-formation system for social interaction detection. In *Proceedings of ACM International Conference on Multimedia*, pages 937–946, 2013.

[15] F. Guo, Y. Li, M. S. Kankanhalli, and M. S. Brown. An evaluation of wearable activity monitoring devices. In *Proceedings of ACM International Workshop on Personal Data Meets Distributed Multimedia*, pages 31–34, 2013.

[16] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard. BioGlass: Physiological parameter estimation using a head-mounted wearable device. In *International Conference on Wireless Mobile Communication and Healthcare: "Transforming healthcare through innovations in mobile and wireless technologies"*, pages 55–58, 2014.

[17] R. Hincks. Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, 33(4):575 – 591, 2005.

[18] E. S. Klima. *The signs of language*. Harvard University Press, 1979.

[19] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi. Presentation sensei: a presentation training system using speech and image processing. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 358–365, 2007.

[20] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013.

[21] G. Luzardo, B. Guamán, K. Chiluiza, J. Castells, and X. Ochoa. Estimation of presentations skills based on slides and audio features. In *Proceedings of the ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 37–44, 2014.

[22] W. Mansell, D. M. Clark, A. Ehlers, and Y.-P. Chen. Social anxiety and attention away from emotional faces. *Cognition and Emotion*, 13(6):673–690, 1999.

[23] R. E. Mayer. *Learning and instruction*. Prentice Hall, 2003.

[24] S. Morreale and P. Backlund. Large-scale assessment in oral communication: K-12 and higher education, washington. *National Communication Association*, 2007.

[25] S. P. Morreale, M. R. Moore, K. P. Taylor, D. Surges-Tatum, and R. Hulbert-Johnson. *The competent speaker speech evaluation form*. National Communication Association, 1993.

[26] J. Müller, J. Exeler, M. Buzeck, and A. Krüger. Reflectivesigns: Digital signs that adapt to audience attention. In *International Conference on Pervasive Computing*, pages 17–24, 2009.

[27] A.-t. Nguyen, W. Chen, and M. Rauterberg. Online feedback system for public speakers. In *IEEE Symposium on E-Learning, E-Management and E-Services*, pages 1–5, 2012.

[28] T. Pfister and P. Robinson. Speech emotion classification and public speaking skill assessment. In *Human Behavior Understanding Workshop*, pages 151–162. Springer, 2010.

[29] R. L. Quianthy. *Communication is life: Essential college sophomore speaking and listening competencies*. Speech Communication Association, 1990.

[30] D. M. Randel. *The Harvard dictionary of music*, volume 16. Harvard University Press, 2003.

[31] M. Rouvier, G. Dupuy, P. Gay, E. el Khoury, T. Merlin, and S. Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. In *INTERSPEECH*, pages 1477–1481, 2013.

[32] L. M. Schreiber, G. D. Paul, and L. R. Shibley. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233, 2012.

[33] A. W. Siegman and S. Feldstein. *Nonverbal behavior and communication*. Psychology Press, 2014.

[34] M. Slater, D. Pertaub, C. Barker, and D. M. Clark. An experimental study on fear of public speaking using a virtual environment. *Cyberpsychology, Behavior, and Social Networking*, 9(5):627–633, 2006.

[35] S. M. Tasko and K. Greilick. Acoustic and articulatory features of diphthong production: A speech clarity study. *Journal of Speech, Language, and Hearing Research*, 53(1):84–99, 2010.

[36] S. Thomson and M. L. Rucker. The development of a specialized public speaking competency scale: Test of reliability. *Communication Research Reports*, 19(1):18–28, 2002.

[37] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[38] J. Webster and H. Ho. Audience engagement in multimedia presentations. *ACM SIGMIS Database*, 28(2):63–77, 1997.

[39] X. Wei and Z. Yang. Mining in-class social networks for large-scale pedagogical analysis. In *Proceedings of ACM International Conference on Multimedia*, pages 639–648, 2012.

[40] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable multi-instance learning. In *IEEE International Conference on Data Mining*, pages 1037–1042, 2014.

[41] J. S. Wrench, A. Goding, D. I. Johnson, and B. A. *Stand Up, Speak Out: The Practice and Ethics of Public Speaking*. 2011.

[42] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of ACM International Conference on Multimedia*, pages 572–579, 2004.

[43] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[44] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, 2012.