Utilizing Multimodal Cues to Automatically Evaluate Public Speaking Performance

Lei Chen, Chee Wee Leong, Gary Feng, Chong Min Lee, Swapna Somasundaran
Educational Testing Service (ETS)
660 Rosedale Rd
Princeton, New Jersey 08541
Email: {LChen, CLeong, GFeng, CLee001, SSwapna}@ets.org

***Abstract*—Public speaking, an important type of oral communication, is critical to success in both learning and career development. However, there is a lack of tools to efficiently and economically evaluate presenters' verbal and nonverbal behaviors. The recent advancements in automated scoring and multimodal sensing technologies may address this issue. We report a study on the development of an automated scoring model for public speaking performance using multimodal cues. A multimodal presentation corpus containing 14 subjects' 56 presentations has been recorded using a Microsoft Kinect depth camera. Task design, rubric development, and human rating were conducted according to standards in educational assessment. A rich set of multimodal features has been extracted from head poses, eye gazes, facial expressions, motion traces, speech signal, and transcripts. The model building experiment shows that jointly using both lexical/speech and visual features achieves more accurate scoring, which suggests the feasibility of using multimodal technologies in the assessment of public speaking skills.**

I. Introduction

Oral communication skills are critical to success in both learning and career development. For example, it has been consistently rated as one of the most valued workforce skills in large scale surveys [18], and its importance is also reflected in the newly developed national K-12 education standards [26]. There is a strong need for valid, reliable, and cost-efficient public speaking assessments.

Public speaking involves a range of constructs, including but not limited to content, organization, language use, vocal quality, and the effective use of non-verbal cues. Evidence for these qualities is distributed across multiple modalities, from linguistic (e.g., coherence of the message and word choice), vocal (e.g., intonation and disfluencies), facial expressions, to hand and body gestures. A competent public speaker coordinates all aspects of these modalities to achieve an engaging and effective performance. It is not surprising that most existing rubrics for public speaking skills [20], [27], [29] evaluate both verbal and nonverbal aspects of communication.

Public speaking performances are traditionally scored by humans. Human scoring has several inherent limitations, particularly in the context of large assessment programs [3]. One issue is reliability, i.e., the same performance may receive different scores from different human raters, or even from the same raters at different times. The cost of training and maintaining the scorers make it difficult to scale up. For applications such as public speaking coaching, which require near real-time assessment results, the logistics can be prohibitive for a large deployment based on human scoring. In these cases, automated scoring becomes an attractive alternative. In this paper, we describe our research on the automated evaluation of public speaking performance utilizing multimodal sensing technologies, namely motion tracking, head-pose/gaze tracking, facial expression analysis, speech analysis, and lexical analysis.

The remainder of the paper is organized as follows. Section II provides an overview of prior research on evaluating public speaking skills. Section III describes the multimodal corpus we collected, including task design, data collection methods, and human scoring process. Section IV describes the multimodal features. Section V describes our experiments on model-building to predict presentation performance using the extracted multimodal features. Finally, in section VI, we summarize the findings of the paper and outline future research directions.

II. Previous Research

Public speaking has received attention from a number of different areas of research. For example, in affective computing, [24] developed a state-of-the-art speech-based emotion detection system and applied the system to assessing public-speaking samples. In order to investigate human multimodal behaviors under stressful conditions, [13] used the public-speaking scenario to create a corpus containing presentations in which stress was elicited. 19 participants' multimodal behaviors were recorded, including speech, video of the facial expressions, and body movements (by using the Kinect depth camera), balance (via a force plate), and other physiological measurements.

Automated solutions for public speaking evaluation depends critically on the multimodal sensing technology. Early attempts tended to rely on a single modality or require cumbersome setups. For example, [17] developed a presentation coaching system and used a marker-based object tracking method for tracking head orientations. However, marker-based and video-only methods for tracking body movements tend to be cumbersome and error-prone. With the introduction of consumer depth cameras such as Microsoft Kinect, tracking human movements

becomes increasingly accurate and convenient [34]. Depth cameras have been used for recording presenters' full body behaviors [1], [6], [21], [23]. For example, [1] created a public speaking skill training system with a combination of advanced multimodal sensing and virtual human technologies. In particular, MultiSense [28] was used to record and recognize a set of the multimodal behaviors of presenters. Meanwhile, a virtual audience would respond to the quality of the presentation in real time to provide feedback and training opportunities.

The new multimodal sensing hardware and software technologies afford the possibility of developing automated systems for evaluating public speaking performances. [6] proposed using multimodal cues to automatically assess public-speaking performance. In addition, a multimodal presentation corpus, known as the Oral Presentation Quality Corpus [10], containing presentations' audio, video, and motion traces, has been provided in the Multimodal Learning Analytics (MLA) 2014 Grand Challenge and Workshop [23]. This data set was collected in Ecuador as a part of a college level course. 1 to 6 students formed a group and developed a presentation (in either PowerPoint or PDF format) and later delivered their presentations individually. In total, 441 multimodal presentations (approximately 19 hours of multimodal data, i.e., audio, video, and depth data) were recorded using a video camera and a Kinect for Xbox device. The presentations were rated on several dimensions. A summary of the research projects from the three teams participating the MLA Challenge can be found in [23].

The emerging field of automated assessment of public speaking performance is a testament to the need for evaluation and training with respect to the skills, as well as the optimism that current multimodal sensing technologies will finally enable such applications. While we share this enthusiasm, we caution that the development of such assessment should follow best practices in assessment design [19]. Not all researchers have clearly identified the constructs they attempt to measure, or paid enough attention to the validity and reliability of human scores.

In this paper, we will report our research on developing an automated public-speaking assessment. As compared to previous research, we highlight the following: (a) applying the rigorous human rating practices widely used in the assessment development area in order to provide more reliable human rating results, (b) utilizing the rich information from multimodal channels used by presenters, and (c) following a data-driven approach to build an automatic scoring system to rate public speaking performance.

III. Corpus

A. Task and data recording

Two types of public speaking tasks were used in our data collection, namely *informative* and *impromptu* presentations. In an informative speech, presenters were given a pre-prepared slide deck and up to 10 minutes to prepare their presentation deliveries. In an impromptu speech, presenters were required to recommend something that was not actually favorable to them. No visual aids were provided for this type of speech. Each task type contained two speech sessions, and therefore each presenter was required to deliver 4 presentations.

Fig. 1. Multimodal data example (informative speech) : the left panel shows a video of a subject with the tracking of facial features activated; the right panel shows the corresponding motion-trace

A Microsoft Kinect for Windows Version 1 device was used to record 3D body motions. Brekel Pro Body Kinect tracking software (v1.30 64 bit version) was used to record 48 body joints, and the motion data was stored using the BioVision Hierarchical (BVH) format. A consumer-level digital camcorder was used for audio/video recording. Note that we only used the camcorder's built-in microphone with a goal to reduce the complexity of the equipment setup in order to allow the developed technology to be used in practice. The Kinect and camcorder devices were mounted together on a tripod and placed 6ft away from the front of the speaking zone, which was marked on the ground. For the informative speech task, a SMART Board projector system was used to show the PowerPoint slides.

The subjects were 17 volunteers recruited from ETS, with 10 male and 7 female participants[1]. After being familiarized with the recording equipment, participants were informed that they were expected to speak for 4 to 5 minutes for the informative speech task and 2 to 3 minutes for the impromptu speech task. Before each recording, the speaker was asked to clap, which served as a synchronization signal common to all the multimodal data streams. Due to data loss caused by equipment failures, in total, we obtained 56 presentations with complete multimodal recordings from a total of 14 speakers. Only one subject was a non-native English speaker with a noticeable accent. Figure 1 provides an example of the collected video and motion tracking results. More details on other data processing steps, e.g., video conversion, speech transcription and alignment, for creating this corpus can be found in [6].

B. Human rating

Human-generated scores are important for building a system to automatically assess public speaking presentations. The quality of human-generated scores is determined by at least

[1]7 of the participants were experienced public speakers from the Toastmasters Club. The rest varied widely in their experience in public speaking. The participants' time was paid by ETS so they did not receive additional compensation.

two aspects, namely the validity of the rubric and the reliability of the human scores. Ideally the scoring rubric should cover important constructs and can be applied reliably across raters.

A limited number of rubrics for measuring the core competencies of public speaking has been published, e.g., [20], [27]. [29] conducted an in-depth and systematic analysis on these rubrics and identified 9 core competencies (roughly corresponding to the list presented in [27]) and provided the Public Speaking Competence Rubric (PSCR). In addition, two human rating studies have been conducted using the PSCR and this rubric's psychometric properties have been investigated. Among the studies in the Computer Science (CS) area on evaluating public-speaking performance, however, not enough attention has been paid to the human scoring. For example, few used established scoring rubrics and reported any details on how the scores were obtained. In some cases a single scorer was used [10], [24], making it difficult to evaluate the psychometric properties of the human scores. Taken together, there is a pressing need to improve the quality of the scoring rubrics and the reliability of the human scores.

In the present research we adopted the PSCR [29], an established rating rubric reported in the literature, as the basis for our human scores. Specifically, we used the following 9 items, *introduction*, *Organization*, *Conclusion*, *Word choice*, *Vocal expression*, *Nonverbal behavior*, *Adapts to audience*, *Visual aids*, and *Persuasive*. Note that these 9 ratings are not independent; in fact a good performance is often high in all dimensions, and vice verse. For example, [29] conducted a factor analysis showing three strongly correlated factors, though the result is likely unstable given the small sample size. In addition to the 9 item-scores, we also asked raters to provide a holistic judgement of the speaking performance.

In addition to choosing an established scoring rubric, we also followed best practices in psychometrics in conducting the human rating. We recruited five raters with experience in rating language proficiency. All presentations were double-scored by two raters using a score-scale from 1 to 4. In the cases in which the two scores differed by more than 1.0 point, a third scorer was called in to adjudicate. The final "operational" score is the average of all scores assigned to a given dimension for a given presenter. A psychometric analysis of the human scores was published in [15] and is not reported here for space reasons[2].

To further understand the internal structure of the human scores, we conducted a Principal Component Analysis (PCA) on all adjudicated sub-scores available in each task type. We were interested in whether there is a common principal component (PC) that can be used as an index of the public speaking performance, and if so, whether it correlates with the holistic human score. Indeed, we found that the first PC explains 68.8% of variance for the informative task and 61.1% of variance for the impromptu task. No other PCs had an eigenvalue larger than 1.0. Unlike in [29], a single factor model is the most parsimonious model for our data. However, [29] chose to use an oblique factor analysis model that generated three correlated factors, which is not inconsistent with our model. On our corpus, the single PC model seems to provide a valid measure of public speaking performance. The Pearson correlation r between the holistic scores with the single PC scores is 0.941 for the informative task and 0.93 for the impromptu task.

In summary, in this paper we used the PSCR as a comprehensive rubric for public speaking performance. Human scoring was based on best practices in psychometrics. Evidence suggests that item-scores in PSCR measures a single construct that is highly correlated with the holistic judgement of public speaking performance. Given this finding, we focus on using multimodal features to predict the holistic score in subsequent analyses.

[2]Low to medium reliability between human raters was observed. For example, intraclass correlation coefficients (ICCs) on several key items and holistic scores are: 0.60 (Vocal expression), 0.41 (Nonverbal behavior), 0.39 (holistic score)

IV. Multimodal Features

A. Kinect Features

Various methods have been proposed to compute expressive features related to body language in several research areas, e.g., affective computing, virtual agents, and multimodal dialogic systems. For example, [22] systematically summarized the methods used for analyzing expressive body language performances. They categorized the extractable features into three layers: (a) low-level, such as *hands velocity*, (b) medium-level, such as *hands symmetry*, and (c) high-level, such as gestures bearing communicational meanings. [14] presented a framework for finding a minimal representation set of affective gestures. They first processed head and hand motion data through an array of expressive feature extraction modules for measuring energy, spatial extent, smoothness, symmetry, and head posture. Then, they computed statistics such as the maximum, mean, and standard deviation (SD) from these measurements. Finally, the minimal set of representational features was obtained using PCA-based dimension reduction.

Following [14], [22], we extracted a number of visual features related to spatial, temporal and dynamic aspects from the Kinect motion data, a frame-by-frame 3D-coordinated (XYZ) recording of each body part as stored in the BVH file. Specifically, we focus on the following body parts: hip, spine, left/right forearms, and left/right hands, due to these body parts' dynamism during presentation.

- **Spatial**: for each motion data frame, (a) the distance between the hands and the body (**hands-spine**); (b) the distance between the arms and the body (**arms-spine**), and (c) the distance between the two hands (**hands**);
- **Temporal**: the first order derivatives of the above spatial measurements;
- **Power**: the second order derivatives of the above spatial measurements;

[22] also suggested several derived features to measure full-body motions, such as Kinetic Energy (KE), Posture, and Symmetry.

- **Energy**: We postulate the KE as a measure of the power/energy exhibited by the presenter. KE is computed

based on the velocities of upper-body segments and their corresponding percentage of mass, as described in the formula below:

$$KE = 0.5 \times \sum_i m_i v_i^2$$

where m_i and v_i refer to the normalized mass ratio and velocity of body part i respectively. In our work, we focus exclusively on body parts that are in the upper-body region – namely, hips, spine, shoulders, arms, forearms and hands – noted for their importance in a related study in Kinesiology [25], from which we also obtained the normalized mass of these body parts[3].

- **Posture**: The posture of a presenter can be approximated using the concept of a Bounding Volume (BV) described in [22], whereby using the hip as the center, we construct an imaginary, pseudo-cuboid whose length, width and height are formed by the furthest distance spanning any two arbitrary body parts in the X, Y and Z dimensions respectively. The BV of a presenter at any frame can be computed simply by taking the volume of this imaginary cuboid. Intuitively, BV provides an approximation of the degree of body "openness" shown by the presenter.
- **Symmetry**: Following [22], we compute the symmetry index (SI) of a presenter's two-handed gestures. SIs measure the degree of symmetry exhibited by both hands along a specified spatial dimension. For example, we compute X dimension's SI as

$$SI_X = \frac{|| \, X_P - X_L \, | - | \, X_P - X_R \, ||}{| \, X_L - X_R \, |}$$

where X_L and X_R are the X-dimension coordinates of the left and right hands, and X_P is the X coordinate of the pivoting body part i.e. hips. A value of 0 for SI_X suggests perfect symmetry while a value of 1 suggests perfect dissymmetry. SI for the Y and Z dimensions can be computed accordingly. From the symmetry measurements of the individual axes, we also derived the following averages as additional measurements: $SI_{XY} = (SI_X + SI_Y)/2$ and $SI_{XYZ} = (SI_X + SI_Y + SI_Z)/3$.

Finally, for each frame-wise measurement described above (f), a set of statistics was computed to serve as features for the entire response, including:

- **mean**: mean value, which is $mean(f)$
- **mmrt**: ratio of the max value to the mean, which is computed as $max(f)/mean(f)$
- **mean-log**: mean of log-scaled value, which is computed as $mean(log(f))$
- **mmrt-log**: ratio of max log-scaled value to the mean of log-scaled value, which is computed as $max(log(f))/mean(log(f))$
- **SD**: standard deviation, which is computed as $SD(f)$
- **SD-log**: standard deviation of log-scaled values, which is computed as $SD(log(f))$

[3]When a body part listed in our BVH motion tracking result is missing from [25], we selected the normalized mass weight of the closest body part as its replacement e.g. pelvis weight is used for hips, abdomen weight is used for spine.

B. Head pose and eye gaze

A successful presentation entails speaker engagement with the audience, which translates to head postures and eye gazes that are necessarily directed towards the audience. Here, we extract a set of features that target these aspects of the presentation.

Head postures are approximated using the *rotation* attribute (i.e., *pitch*, *yaw*, and *roll*) of the head through Visage's SDK FaceTrack[4], a robust head and face tracking engine. The tracking is activated if and only if the detector has detected a face in the current frame. Additionally, gaze directions are approximated through the *gazeDirectionGlobal* attribute of the Visage tracker SDK, which tracks gaze directions taking into account both head pose and eye rotation. Note that, different from head rotation, gaze directions represent estimated "eyeball" directions regardless of head postures, and can potentially measure a speaker's level of engagement with the audience.

For each presentation's basic head pose measurements, (i.e., pitch, yaw, and roll) and gaze tracking measurements (i.e., X, Y, and Z) over the entire presentation, we computed their SD, kurtosis, and skewness. Additionally, a feature measuring extreme values' ratio (*ert*) is computed as follows: for each measurement, obtain the $10th$ percentile and $90th$ percentile from our entire data set to be the lower-bound and the upper-bound. For each contour, we then use the proportion of the values beyond these two bounds as a feature.

C. Facial expression

Facial expressions from presenters also contribute to an effective presentation. Therefore, we utilized an off-the-shelf emotion detection toolkit, Emotient's FACET SDK[5], to analyze facial expressions. FACET outputs the intensity (ranging from 0.0 to 1.0) and confidence values for seven primary emotions (i.e., *anger*, *contempt*, *disgust*, *joy*, *fear*, *sadness* and *surprise*). Due to several technical challenges arising in our data set, including (a) that presenters' faces are typically small given the large distance from the camera, (b) that there was extreme background brightness from using a SmartBoard during the informative task, and (c) that presenters frequently turned their heads during their presentations, we have low expectations that accurate and precise emotion tracking on these individual emotion categories can be obtained. To provide a coarse representation of variation of facial expressions, we computed the following metrics: total emotion (mean of positive emotion (joyness) and negative emotions, which is the mean of the other six emotions), and the ratio of positive to negative emotion ($p2n$). Then, for each response, these measurements' SD, kurtosis, and skewness were computed to serve as emotion features.

[4]http://www.visagetechnologies.com/

[5]http://www.emotient.com

D. Lexical features

We also extracted lexical features measuring the presentations' subjectivity/sentiment and language usage. Sentiment and subjectivity play an important role in delivering powerful speeches, especially on the persuasive and impromptu speech task. Therefore, we use subjectivity information provided by the MPQA subjectivity lexicon [31] and sentiment information provided by a sentiment lexicon (*ASSESS*)[6] [2]. The features consist of (a) the presence and count of existing polar/neutral words from the MPQA lexicon and (b) the presence and count of existing polar words from the ASSESS lexicon.

Using appropriate combinations of words is a skill expected in fluent presentations. Therefore, we utilized the collocation analysis method widely used in NLP for extracting word usage features. In particular, the Point-wise Mutual Information (PMI) of all adjacent word pairs (bi-grams) as well as all adjacent word triples (tri-grams) in the Google 1T web corpus [5] have been computed. Word tuples' PMI values show how common a word tuple is – the higher the value of the PMI the more common the tuple is to be observed. 16 bins (8 for bi-grams and 8 for tri-grams) were computed from all of the PMI values from the Google IT web corpus. For each presentation's transcript, we used its bi/tri-grams' proportions into these bins as collocation features. Also, the maximum, minimum, and median PMI values from bi-grams and tri-grams were encoded as features.

Using vivid descriptions to make presentations more colorful is also helpful for delivering well-performed presentations. For this aspect, we also modeled certain syntactic categories, e.g., adjectives and adverbs, by using their presence and count to serve as lexical features. More details about these lexical features from presentations' text transcripts can be found in [30].

E. Speech features

Speaking skill comprises multiple dimensions, including fluency, pronunciation, prosody, language usage, and so on. In the past two decades, automatic speech scoring technology (ASST) has been developing to provide objective and comprehensive measurements on these dimensions. For example, Automatic Speech Recognition (ASR) is used for measuring pronunciation skills [32] and several commercial speech-scoring products have been built [4], [11], [33]. Although the ASST's main focus is on language learners, it is used to track native speakers' progresses as well, as shown in [9]. Given the rich measurement of speaking competency provided by the ASST, we believed that this technology is also useful for providing speech-related features for scoring public speaking. Therefore, following the feature extraction method described in [7], [33], we used speech and transcription to generate a series of features on the multiple dimensions of speaking skills, e.g., speaking rate, prosodic variation, pausing profile, pronunciation, and so on.

[6]Note that the MPQA lexicon provides a positive, negative or neutral polarity category to its entries, while the ASSESS lexicon provides a positive/negative/neutral probability distribution.

TABLE I
THE CORRELATIONS BETWEEN THE PRE-PROCESSED PCA FEATURES FROM EACH FEATURE CATEGORY AND HUMAN RATED HOLISTIC SCORES. NOTE THAT THE PC NUMBER FOR THE REPORTED CORRELATION $|R|$ WAS REPORTED IN PARENTHESES.

Category	Feat.#	PC #	$\|r\| > 0.280$ for significant corr.
head/gaze	24	6	0.310 (PC2)
emotion	6	2	0.380 (PC2)
Kinect motions	78	7	0.284 (PC4), 0.347 (PC6)
speech	74	6	0.310 (PC2)
lexical	56	4	0.310 (PC2)

F. Features' indicative capabilities

In order to evaluate the indicative capabilities of the multimodal features described in Section IV, we run the following analysis to cope with dealing with a large number of features. In particular, for each feature group, i.e., head poses and eye gazes (head/pose), facial expressions (emotion), motion tracking (Kinect motions), speech analysis (speech), and lexical analysis (lexical), using the feature values from the total of 56 presentations, we run a PCA to map the multiple features to several PCs that can explain 75% of cumulative variance. Then, we computed the Pearson correlation of these PC features with the human-rated holistic scores (adjudicated version). Table I reports the obtained results. Note that we only reported PC features with an absolute correlation $|r|$ higher than 0.28, which corresponds to a statistically significant correlation for the small sample size ($n = 56$). We find that each feature group provides useful information for evaluating human-rated holistic scores.

V. EXPERIMENTS

We applied the multimodal features described in Section IV to the task of building an automatic scoring system for evaluating public speaking performance per task type. For the very limited data size ($n = 28$ for each task), dimension-reduction on features is typically helpful for improving prediction accuracy. Therefore, we follow the approach suggested in [14], which is to apply PCA to provide a more concise representation of the features. This process is explained in Figure 2. Two PCAs have been conducted on (a) the speech and lexical features and (b) the visual features representing head pose, eye gaze, facial expressions, and body language, for representing presentation competence on speech and visual modalities respectively. Note that the default 0.95 cumulative variance stopping criterion was used and for each feature type, and was sufficient to cover more than 100 individual features within 15 PCs.

For each task type, i.e., informative speech vs. impromptu speech, we employ a standard machine learning framework using the multimodal features obtained from the above PCA pre-processing steps to predict human-judged holistic scores. In particular, we run a leave-one-presenter-out cross-validation among all subjects ($n = 14$). The conducted experiments are divided into three feature groups, namely (a) visual features (visual), (b) speech and lexical features (speech+lexical), and (c) the combination (multimodal). For each fold, during

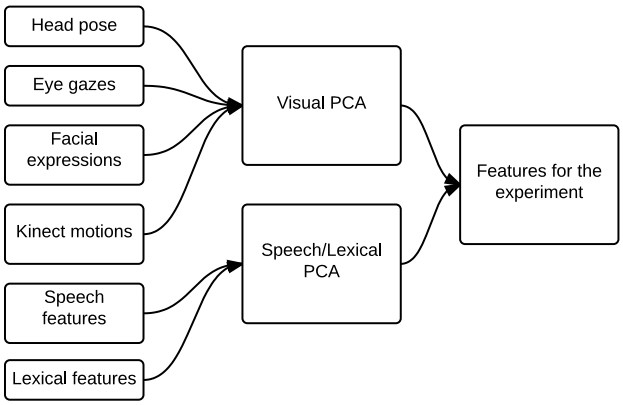

Fig. 2. A diagram showing the use of two PCAs to reduce both visual and speech features to a small set

TABLE II
USING MULTIMODAL FEATURES TO PREDICT FINAL HOLISTIC SCORES ON THE INFORMATIVE AND IMPROMPTU TASK TYPES

Feature set	SVM (poly)	*glmnet*	RF
Informative, $r_{H1} = 0.796$ $r_{H2} = 0.880$			
visual	0.178	0.204	0.343
speech+lexical	**0.457**	**0.526**	0.411
multimodal	0.396	0.411	**0.527**
Impromptu, $r_{H1} = 0.843$ $r_{H2} = 0.766$			
visual	0.504	0.465	0.226
speech+lexical	0.270	0.160	0.427
multimodal	**0.634**	**0.566**	**0.589**

training, we computed the Pearson correlation of each feature group to human holistic scores. Consequently, we only utilized the PC features with a correlation higher than 0.25. The three regression approaches widely employed in practice were utilized, with their implementations in the R **Caret** package [16]: (a) Support Vector Machine (SVM) using a polynomial kernel (svmPoly) as described in [8], (b) *glmnet* [12], a generalized linear model fitted via penalized maximum likelihood, and (c) random forest (RF) corresponding to the method "rf" in **Caret**. Hyper parameters of these machine learning models were automatically tuned by using a 10-fold cross-validation on the training set. For example, svmPoly [8] needs to specify three hyper parameters, i.e., polynomial degree, scale, and cost whereas RF only needs to specify the selected features for each tree node. The whole process was repeated for 14 times to obtain the machine predicted scores for all presentations. Table II reports on the correlation between the human-rated holistic scores and the machine-predicted scores. Note that the correlations between each individual rater's holistic scores (H1 or H2) and adjudicated holistic scores were provided to be the (upper) base line for each task.

When focusing on the SVM and *glmnet* models, we found that these two models show different result patterns between the two tasks. In particular, for the informative task, the models using visual features are worse than the models using speech+lexical features. Adding the visual features on top of the speech+lexical features could not further improve prediction performance. In contrast, for the impromptu task, the models using visual features are more accurate than the models using the speech+lexical features. More importantly, jointly using these two types of features generates the most accurate predictions. The fact that visual features play more of a role in accurately predicting presentation performance for the impromptu task seems consistent with our prediction based on its task design.

However, the results from the random forest (RF) models (shown in the fourth column) provide a different picture. Between the two tasks, the RF models show a consistent pattern: (a) the model using speech+lexical features is somehow better than the model using visual features and (b) the model using multimodal features achieves the best performance[7]. The contrasting patterns from using different machine learning models may be caused by the limited sample size ($n = 14$ for each task) in our experiment. This also calls our attention to the need to increase the corpus's size for more deterministic experiments. Another possible reason could be the RF's special nature of containing an internal ensemble scheme in its model-learning process. When comparing the human scoring (upper) baselines with the best machine prediction results, we observe some encouraging results. For the impromptu task, the highest correlation (0.634 from the SVM model using multimodal cues) is promisingly close to the values from human rated scores.

VI. DISCUSSION

Public speaking is an important type of oral communication, playing many roles for academic and job-related success. However, the evaluation of public speaking is still occasional and heavily relies on human rating, a costly process in which it is hard to meet all learners' needs. Therefore, with the rapid progress of multimodal sensing technology, it is important to build an automated scoring system to evaluate public speaking's inherently multimodal performance. This research starts from the construction of a presentation corpus containing multimodal recordings, i.e., speeches, transcripts, videos, and motion traces collected using a Kinect depth camera, and human-rated performance evaluation scores based on the standards in the assessment development field. The multimodal nature of the corpus enables us to evaluate presentation performance via both verbal and nonverbal dimensions. To our knowledge, this corpus is the first one to have human rating results using a well-established rubric and rigorous human rating quality control.

The remainder of the paper clearly illustrates a complete end-to-end data-driven method to build automatic scoring systems. Based on the human knowledge used for rating public-speaking and various feature extraction methods in related areas, e.g., NLP, spoken language assessment, and nonverbal communication analysis, we extracted a large set of multimodal features. The scoring models using either speech/lexical

[7]However, a significance test (using R **psych** package's paired.r function) on such small data set ($n = 14$) did not suggest that such correlation increases are significant.

or visual features alone show promising correlation values between machine-predicted scores and human rated scores, but more importantly, jointly using both visual and speech/lexical features shows improvement over the uni-modality features alone on the four experiments – each experiment corresponds to a combination of feature usage and a machine learning model, among the six experiments being tried.

The limited size of our corpus hinders us in making strong conclusions about the multimodal features' quality and in building more accurate machine prediction models. In addition, the features utilized are mostly from low/medium levels rather than high-level information, e.g., gestures with communicative purposes, and from simple statistics. Therefore, in our future work, we plan to expand our corpus size, to extract and utilize high-level nonverbal features with clear semantics and social meanings, and to use temporal structure-related information.

ACKNOWLEDGMENT

The authors would like to thank the following: (a) ETS Research Allocation Program for providing the funding for conducting the study, (b) Jilliam Joe and Christopher Kitchen for their important help in creating the corpus, (c) Saad Khan for his suggestion of using the Visage SDK, (d) David Suendermann-Oeft, Vikram Ramanarayanan, Xinhao Wang, and Keelan Evanini for providing internal reviews, (e) the ACII anonymous reviewers for their valuable comments, and (f) James Bruno for proofreading the manuscript.

REFERENCES

[1] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer. Cicero-towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents*, pages 116–128, 2013.

[2] B. Beigman Klebanov, J. Burstein, and N. Madnani. Sentiment profiles of multi-word expressions in test-taker essays: The case of noun-noun compounds. In *ACM Transactions on Speech and Language Processing*, volume 10(3), 2013.

[3] R. Bennett. Moving the field forward: Some thoughts on validity and automated scoring. *Automated scoring of complex tasks in computer-based testing*, pages 403–412, 2006.

[4] J. Bernstein, A. V. Moere, and J. Cheng. Validating automated speaking tests. *Language Testing*, 27(3):355, 2010.

[5] T. Brants and A. Franz. Web 1T 5-gram Version 1. In *Linguistic Data Consortium, Philadelphia*. 2006.

[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 16th International Conference on Multimodal Interaction*, pages 200–203. ACM, 2014.

[7] L. Chen, K. Zechner, and X. Xi. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *NAACL-HLT*, 2009.

[8] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. Misc functions of the Department of Statistics (e1071), TU Wien. *R package*, pages 1–5, 2008.

[9] R. Downey, D. Rubin, J. Cheng, and J. Bernstein. Performance of automated scoring for children's oral reading. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 46–55. Association for Computational Linguistics, 2011.

[10] ESPOL. Description of the oral presentation quality corpus. http://www.sigmla.org/datasets/, 2014.

[11] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda. EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401, 2010.

[12] J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

[13] T. Giraud, M. Soury, J. Hua, A. Delaborde, M. Tahon, G. Antonio, V. Eyharabide, E. Filaire, C. L. Scanff, L. Devillers, and others. Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 417–422. IEEE, 2013.

[14] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. Toward a minimal representation of affective gestures. *Affective Computing, IEEE Transactions on*, 2(2):106–118, 2011.

[15] J. Joe, C. Kitchen, L. Chen, and G. Feng. A prototype public speaking skills assessment: An evaluation of human scoring quality. *ETS Research Report*, in press.

[16] M. Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[17] 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 9th international conference on Multimodal interfaces*, pages 358–365. ACM, 2007.

[18] P. C. Kyllonen. Measurement of 21st century skills within the common core state standards. In *Invitational Research Symposium on Technology Enhanced Assessments. May*, pages 7–8, 2012.

[19] R. J. Mislevy and G. D. Haertel. Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4):6–20, 2006.

[20] S. P. Morreale, M. R. Moore, D. Surges-Tatum, and L. Webster. *"The competent speaker" speech evaluation form (2nd ed.)*. National Communication Association, 2007.

[21] A.-T. Nguyen, W. Chen, and M. Rauterberg. Online feedback system for public speakers. In *IEEE Symp. e-Learning, e-Management and e-Services*. Citeseer, 2012.

[22] R. Niewiadomski, M. Mancini, and S. Piana. Human and virtual agent expressive gesture quality analysis and synthesis. *Coverbal Synchrony in Human-Machine Interaction*, pages 269–292.

[23] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz. MLA'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 531–532. ACM, 2014.

[24] T. Pfister and P. Robinson. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *Affective Computing, IEEE Transactions on*, 2(2):66–78, 2011.

[25] S. Plagenhoef, F. G. Evans, and T. Abdelnour. Anatomical data for analyzing human motion. *Sport*, 54:169–178, 1983.

[26] A. Porter, J. McMaken, J. Hwang, and R. Yang. Common core standards the new us intended curriculum. *Educational Researcher*, 40(3):103–116, 2011.

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

[28] S. Scherer, G. Stratou, and L.-P. Morency. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 135–140. ACM, 2013.

[29] 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.

[30] S. Somasundaran, C. M. Lee, M. Chodorow, and X. Wang. Automated scoring of picture-based story narration. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48, 2015.

[31] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[32] S. M. Witt. *Use of Speech Recognition in Computer-assisted Language Learning*. PhD thesis, University of Cambridge, 1999.

[33] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895, 2009.

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