Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues

Lei Chen
Educational Testing Service (ETS)
660 Rosedale Rd
Princeton, NJ, USA
LChen@ets.org

Gary Feng
ETS
660 Rosedale Rd
Princeton, NJ, USA
GFeng@ets.org

Jilliam Joe
ETS
660 Rosedale Rd
Princeton, NJ, USA
JJoe@ets.org

Chee Wee Leong
ETS
660 Rosedale Rd
Princeton, NJ, USA
CLeong@ets.org

Christopher Kitchen
ETS
660 Rosedale Rd
Princeton, NJ, USA
CKitchen@ets.org

Chong Min Lee
ETS
660 Rosedale Rd
Princeton, NJ, USA
CLee001@ets.org

ABSTRACT

Traditional assessments of public speaking skills rely on human scoring. We report an initial study on the development of an automated scoring model for public speaking performances using multimodal technologies. Task design, rubric development, and human rating were conducted according to standards in educational assessment. An initial corpus of 17 speakers with 4 speaking tasks was collected using audio, video, and 3D motion capturing devices. A scoring model based on basic features in the speech content, speech delivery, and hand, body, and head movements significantly predicts human rating, suggesting the feasibility of using multimodal technologies in the assessment of public speaking skills.

Categories and Subject Descriptors

H.5.1 [Information interfaces and presentation]: General—Multimedia Information Systems

Keywords

Multimodal Presentation Assessment, Public Speaking, Multimodal Corpus, Body Tracking, Educational Applications

1. INTRODUCTION

Oral communication is consistently rated as one of the most valued workforce skills in national surveys [8], and its importance is reflected in the Common Core State Standards for K12 education. Public speaking is the epitome - and evidently the most feared form [12] - of oral communication that is often overlooked in educational assessment.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI’14, November 12–16, 2014, Istanbul, Turkey.
Copyright 2014 ACM 978-1-4503-2885-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663204.2663265.

A successful public speaking performance is distributed across multiple modalities - e.g., the speech content, voice and intonation, facial expressions, head poses, hand gestures, and body postures. While current rubrics for evaluating public speaking performances attend to both verbal and non-verbal aspects, virtually all existing assessments require human rating [14]. In this paper, we describe research using multimodal sensing, namely motion tracking, head tracking, and speech processing, toward developing an automated assessment system for scoring presentation skills.

The remainder of the paper is organized as follows: Section 2 reviews the previous research on (a) analyzing presentation performance by using oral and multimodal cues and (b) multimodal corpus using Kinect tracking device. Section 3 describes the multimodal presentation corpus we created, including tasks, data recording methods, and the human scoring process. Section 4 describes experiments predicting presentation skills using multimodal features extracted from audio, video, and motion data. Finally, Section 5 summarizes the findings of the paper and discuss future research directions.

2. PREVIOUS RESEARCH

A number of publications described research prototypes of using audio and video cues to evaluate presentation skills. [13] conducted an audio/visual analysis on a political speech corpus and found some clues related to speaking performance rating. [7] developed a presentation coaching system by using Automatic Speech Recognition (ASR), prosody analysis, and image processing. A marker-based computer vision object tracking method was used for head tracking. However, marker-based and video-only methods for tracking body movements tend to be cumbersome and error-prone.

The emergence of 3D motion tracking devices such as the Microsoft Kinect has greatly facilitated multimodal research. [11] proposed building a public speaking evaluation system. Students participating in a scientific presentation course were recorded by using a Kinect device. [1] created a public speaking skill training system with a combination of advanced multimodal sensing and virtual human technologies. Meanwhile, a virtual audience would respond to the quality of the presentation in real time to provide feedback...
and training opportunities. In addition, affective computing has been utilized to analyze public speaking, especially focusing on speakers’ stress responses [4, 5].

Two observations can be drawn from the literature. With regard to the technology, there is a convergence on 3D sensors such as Kinect in combination with audio/video recording as the basis for evaluating presentation performances. With regard to applications, there is much interest in designing interactive systems for training presentation skills. We believe the multimodal technology can transform traditional human-scored public speaking assessments [14] and make them more reliable and cost-efficient.

Our primary research question, therefore, is whether machine-generated multimodal features can predict human scores based on a well-designed rubric for public speaking performance. The aforementioned multimodal research [11, 1] has so far been based on non-standardized scoring rubrics and human scoring process, leaving it vulnerable to reliability threats. In addition, the content of the speech, which is central in traditional public speaking assessments, is seldom taken into account in these studies. We apply natural language processing techniques on speech transcripts as an initial attempt to calculate content-related features.

3. MULTIMODAL CORPUS

3.1 Tasks

Four public speaking tasks were developed. Tasks B and C were modeled after prepared informational speeches, in which the speaker was given a pre-prepared slide deck and up to 10 minutes to prepare for the presentation. Task B was a business presentation, where the speaker was to present a financial report. Task C is a simulated teaching task on the topic targeting middle school students. The other two tasks were persuasive and impromptu speeches. No visual aids were provided for Task D and Task E.

3.2 Multimodal Data Collection

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 stored the motions in the Biovision hierarchical data format (BVH). A JVC Everio GZ-HM35BUSD digital camcorder was used for audio/video recording. The camcorder was mounted together with Kinect on a tripod. The raw video resolution used UXP high definition setting and was saved in the MTS format. Both Kinect and camcorder were placed 6ft away from the front of the speaking zone that was marked on the ground. For Tasks B and C, a SMART Board projector system was used to show the PowerPoint slides.

Speakers were 17 volunteers recruited from an education assessment company, with 10 male participants and 7 female participants. 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 covered by the employer so they did not additional compensation. After being familiarized with the recording equipment, participants were informed that they were expected to speak for 4 to 5 minutes for Task B and C and 2 to 3 minutes for Task D and E. For Tasks B and C, which involved PowerPoint slides, they were given 10 minute to prepare for their presentation. They were not allowed to bring notes during the presentation. In Task D and E, the participants were given no preparation time. They would start speaking as soon as they were given the topic of the impromptu speech. Before each recording, the speaker was asked to clap, which served as a signal synchronizing the multimodal speech. Some data were lost due to equipment failures. In total, we obtained 56 presentations with complete multimodal recordings.

3.3 Data Processing

After getting raw recordings from our lab sessions, the following steps were taken to process the data for further analysis. The motion and video data streams were synchronized. The HD video files were converted to another format (with a higher compression rate) for playing back, e.g., in Xvid or H.263 codec. The audio channel from the video was extracted for several speech processing procedures, including (a) manual transcription, (b) using P2FA [16] forced alignment tool to obtain time stamps for all transcribed words (and phonemes), (c) using Praat [2] acoustics analysis software to extract pitch and intensity measurements. In addition, nonverbal features related to hand movements, body locomotion, and head orientation, were computed. More details will be provided in Section 4.1.

3.4 Human Rating

A key emphasis of our research is a valid and comprehensive assessment of the public speaking skills. To this end we designed the human rating process according to the standard practice in educational assessment. We began with a construct analysis of skills involved in public speaking and a review of existing assessments and rubrics. The final rubric was based on the Public Speaking Competence Rubric (PSCR) [14] due to its psychometric properties. According to their study using five trained faculty member raters on 45 speech presentations by college students, the intraclass correlation coefficients (ICCs; a measure of variance in scores associated with raters) across the 10 dimensions and holistic score ranged from .37 to .93, making it one the most reliable human-scored instruments for assessing public speaking performance.

Using the PCSR tailored to our tasks, human raters scored these presentation videos on 10 dimensions, such as vocal expression (VE), which measures the efficiency of using vocal expression and paralinguistic cues to engage the audience, and nonverbal behavior (NVB), which measures the effect of using nonverbal behaviors to reinforce verbal messages, as well as overall holistic scores.

Five raters were recruited from the same educational assessment company. Two expert raters had background in oral communication/public speaking instruction at the higher education level. The other three (non-expert raters) had extensive experience in scoring essays, but not in scoring public speaking performances. For reliability purposes, the presentations were double-scored. The scoring process was organized in two phases. Phase 1 involved working with the two expert raters in order to develop criterion scores for a randomly selected subset of presentation videos. The criterion videos and scores were used to train non-expert raters in Phase 2. During the training, non-expert raters review exemplar videos from the criterion set and discuss with expert raters in order to come to the same scores. The process continued until they could reliably rate the criterion set. In
Phase 2 expert and non-expert raters worked together to score the remainder of the presentations. Each presentation was randomly assigned to an expert and a non-expert rater, who scored independently. Regarding holistic scores, between human raters, ICC is 0.385 and Pearson correlation is 0.385. The following adjudication process was used to generate final scores. If two raters agree with each other, the score will be used as the final score. Otherwise, a third rater (expert) will be brought in to make the final judgment. In our experiments described in Section 4, we focused on the final holistic scores after the adjudication process.

4. EXPERIMENTS

4.1 Multimodal Features

We focused on three types of features, namely speech delivery, speech content, and non-verbal behaviors. For speech delivery, we included features widely used in automated speech scoring. In particular, following the feature extraction method described in [3], we used speech and transcription to generate a series of features on the multiple dimensions of speaking skills, e.g., speaking rate, prosodic variations, pausing profile, and pronunciation.

Features of the content of speech were extracted using a syntactic complexity analyzer tool [9] on speech transcripts. This is a crude first attempt because speech transcripts differ importantly from written texts. This tool counts the frequencies of 9 types of syntactic structures, e.g., verb phrases (VP), T-units (T), clauses (C), etc., and computes 14 syntactic complexity feature values, such as the mean length of clause (MLC), verb phrases per T-unit (VP/T), etc. A complete list of the features can be found in [9].

For non-verbal features, we focused on the amount of locomotion, hand movements, and head orientation. The locomotion was indexed by the hip movement from the Kinect data. A basic feature set was extracted based on the mean and standard deviation (SD) of the hip and hand movement speeds and the log-transformed values. The orientation of the head is an approximation of the speaker’s attention to the audience. Head poses were tracked by using GAVAM head tracker [10] from the video data. Unreliable data (with confidence less than 7) were marked as missing. The head features consist of the mean, SD, as well as mean/SD on the log scale, of the horizontal and vertical head movements.

Given the modest size of the corpus, we chose to proceed with a simple feature selection process, by taking the top features from each category based on the Pearson correlation with the human rating scores on the entire data set 1, as seen in Table 1 on the next page.

4.2 Predicting Human Holistic Scores

We run a standard machine learning experiment for using these multimodal features to predict human judged holistic scores. In particular, we run a leave-one-out cross-validation among all of subjects \( n = 17 \). In each fold, presentations from 16 subjects were used to train a regression model and then the trained model was applied to the presentations from the remaining subjects. Two regression approaches were utilized, including Random Forest (RF) and Support Vector Machine (SVM) using a polynomial kernel. We used the implementations provided by R caret package [6]. Hyperparameters of these machine learning models were automatically tuned by using a 5-fold cross-validation on the training set. The whole process was repeated for 17 times to obtain the machine predicted scores for all of the presentations.

Table 2: Regression models of using multimodal features to predict final holistic scores

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Random Forest (RF)</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexical</td>
<td>0.283</td>
<td>0.220</td>
</tr>
<tr>
<td>speech</td>
<td>0.312</td>
<td>0.372</td>
</tr>
<tr>
<td>visual</td>
<td>0.202</td>
<td>0.132</td>
</tr>
<tr>
<td>lexical + speech</td>
<td>0.383</td>
<td>0.377</td>
</tr>
<tr>
<td>multimodal</td>
<td>0.416</td>
<td>0.447</td>
</tr>
</tbody>
</table>

Table 2 reports on the correlation between the human rated holistic scores and machine predicted ones.

By using the lexical features, speech features, and visual features from videos and motions tracked by Kinect, respectively, we find that each modality provides information for predicting overall presentation performance. Among the three modalities, the speech channel provides the most information. After combining both lexical and speech features together, for the RF model, we saw a performance increase. However, such gain was not observed for the SVM model. On the basis of the verbal model (using both lexical and speech features), adding visual features makes the multimodal model achieve the highest performance. The correlation has been increased from 0.383 to 0.416 for the RF model and from 0.377 to 0.447 for the SVM model.

5. DISCUSSIONS

We reported an initial study toward building an automated multimodal scoring model for public speaking assessments. Based on our multimodal presentation corpus, we found that simple multimodal features of the speech content, speech delivery (fluency, pronunciation, and prosody), and nonverbal behaviors (head, body, and hand motions) together significantly predict human scores on the presentation performance. The result is promising as a first step toward an automated assessment of public speaking skills. The study extends prior research (e.g., [11, 1]) with better grounding in construct definitions, more rigorous human scoring process, and the incorporation of content features along with speech and non-verbal features. This interim report based on our initial corpus only explored a few basic features and modeling techniques. Limitations notwithstanding, our on-going research continues to focus on expanding the multimodal corpus, developing advanced features such as gesture recognition, and building more robust scoring models for the multimodal presentation assessment.

6. REFERENCES


---

1If we have a large-sized data set, such feature selection will be conducted on the training set only.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Category</th>
<th>Description</th>
<th>( r_{HS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W )</td>
<td>transcript</td>
<td># words</td>
<td>0.417</td>
</tr>
<tr>
<td>( S )</td>
<td>transcript</td>
<td># sentences</td>
<td>0.382</td>
</tr>
<tr>
<td>( VP )</td>
<td>transcript</td>
<td># verb phrases</td>
<td>0.457</td>
</tr>
<tr>
<td>( C )</td>
<td>transcript</td>
<td># clauses</td>
<td>0.439</td>
</tr>
<tr>
<td>( T )</td>
<td>transcript</td>
<td># T-units</td>
<td>0.454</td>
</tr>
<tr>
<td>( DC )</td>
<td>transcript</td>
<td># dependent clauses</td>
<td>0.298</td>
</tr>
<tr>
<td>( CT )</td>
<td>transcript</td>
<td># complex T-units</td>
<td>0.298</td>
</tr>
<tr>
<td>( CP )</td>
<td>transcript</td>
<td># coordinate phrases</td>
<td>0.305</td>
</tr>
<tr>
<td>( CN )</td>
<td>transcript</td>
<td># complex nominals</td>
<td>0.339</td>
</tr>
<tr>
<td>( DC )</td>
<td>transcript</td>
<td># dependent clauses per clause</td>
<td>0.298</td>
</tr>
<tr>
<td>( MLC )</td>
<td>transcript</td>
<td>mean length of clause</td>
<td>-0.112</td>
</tr>
<tr>
<td>( DC.C )</td>
<td>transcript</td>
<td>dependent clauses per clause</td>
<td>-0.173</td>
</tr>
<tr>
<td>( T.S )</td>
<td>transcript</td>
<td>T-units per sentence</td>
<td>0.302</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Category</th>
<th>Description</th>
<th>( r_{HS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( SpeakRate )</td>
<td>fluency</td>
<td>speaking rate in words per second</td>
<td>0.463</td>
</tr>
<tr>
<td>( UttLenVar )</td>
<td>fluency</td>
<td>Mean deviation of the lengths (words) in chunks</td>
<td>0.447</td>
</tr>
<tr>
<td>( Pausing )</td>
<td>fluency</td>
<td>Mean of pauses’ lengths within clauses</td>
<td>-0.466</td>
</tr>
<tr>
<td>( Pronunciation )</td>
<td>pronunciation</td>
<td>Goodness of pronunciation [15, 3]</td>
<td>0.436</td>
</tr>
<tr>
<td>( Prosody )</td>
<td>prosody</td>
<td>Mean deviation of the prosodic units</td>
<td>0.326</td>
</tr>
<tr>
<td>( HipMeLgSd )</td>
<td>body</td>
<td>Hip movement (in log scale) SD</td>
<td>-0.328</td>
</tr>
<tr>
<td>( LHMeLgSd )</td>
<td>body</td>
<td>left hand movement (in log scale) SD</td>
<td>-0.393</td>
</tr>
<tr>
<td>( RHMeLgSd )</td>
<td>body</td>
<td>right hand movement (in log scale) SD</td>
<td>-0.275</td>
</tr>
<tr>
<td>( FaceUpDownLgMean )</td>
<td>head</td>
<td>Face Y-axis angle (in log scale) mean</td>
<td>-0.261</td>
</tr>
<tr>
<td>( FaceUpDownSd )</td>
<td>head</td>
<td>Face Y-axis angle SD</td>
<td>-0.254</td>
</tr>
<tr>
<td>( FaceLeftRightMean )</td>
<td>head</td>
<td>Face X-axis angle mean</td>
<td>-0.176</td>
</tr>
<tr>
<td>( FaceLeftRightSd )</td>
<td>head</td>
<td>Face X-axis angle SD</td>
<td>-0.158</td>
</tr>
</tbody>
</table>

Table 1: A list of multimodal features used in our experiment; Pearson correlations (\( r_s \)) between the features and human rated holistic scores were reported in the last column.