Multimodal Assessment of Oral Presentations using HMMs

Everlyne Kimani, Prasanth Murali, Ameneh Shamekhi, Dhaval Parmar, Sumanth Munikoti and Timothy Bickmore
Khoury College of Computer Sciences, Northeastern University
Boston, Massachusetts, USA
{kimani15, ameneh, dhavalparmar, sumanthmunikoti, bickmore}@ccs.neu.edu,
murali.pr@northeastern.edu

ABSTRACT

Audience perceptions of public speakers' performance change over time. Some speakers start strong but quickly transition to mundane delivery, while others may have a few impactful and engaging portions of their talk preceded and followed by more pedestrian delivery. In this work, we model the time-varying qualities of a presentation as perceived by the audience and use these models both to provide diagnostic information to presenters and to improve the quality of automated performance assessments. In particular, we use HMMs to model various dimensions of perceived quality and how they change over time and use the sequence of quality states to improve feedback and predictions. We evaluate this approach on a corpus of 74 presentations given in a controlled environment. Multimodal features—spanning acoustic qualities, speech disfluencies, and nonverbal behavior—were derived both automatically and manually using crowdsourcing. Ground truth on audience perceptions was obtained using judge ratings on both overall presentations (aggregate) and portions of presentations segmented by topic. We distilled the overall presentation quality into states representing the presenter's gaze, audio, gesture, audience interaction, and proxemic behaviors. We demonstrate that an HMM of state-based representation of presentations improves the performance assessments.

CCS CONCEPTS

• Human-centered computing ~ Human computer interaction (HCI) ~ Empirical studies in HCI.

KEYWORDS

Multimodal; Hidden Markov Models; Presentations

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418888

ACM Reference format:

Everlyne Kimani, Prasanth Murali, Ameneh Shamekhi, Dhaval Parmar, Sumanth Munikoti and Timothy Bickmore. 2020. Multimodal Assessment of Oral Presentations using HMMs**.** In *Proceedings of 2020 ACM Conference International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands*, 5 pages. https://doi.org/10.1145/3382507.3418888

1 Introduction

Oral presentations to an audience continue to be a central feature of most industries today, and is where colleagues, clients, the media, and the public hear about the latest findings, become engaged and inspired, and where reputations and deals are made. To help presenters improve their performance, several researchers have developed systems that observe presentations using a variety of sensors and provide automated feedback on presentation quality, allowing presenters to avoid the cost, time, and potential stigma of hiring a public speaking coach [1-3]. However, most of this work is aimed at providing aggregate ratings of an overall presentation that, while useful, does not provide information about which part of the presentation was particularly good or bad from the audience's perspective.

We report on work towards the automated assessment of presentation quality as judged by an audience, provided for each segment of a presentation, in addition to an overall quality rating. Here we define 'presentation segment' as a contiguous time interval during a presentation which the speaker uses to satisfy a communicative goal (analogous to the notion of "discourse segment" in dialogue [4]), roughly corresponding to the notion of 'topic.' Providing quality assessments per segment may allow speakers to more rapidly identify the parts of their presentation that require work, rather than having to infer this from low-level features such as speaking rate or head nods [5].

Modeling presentations as sequences of quality states may also aid in automated quality assessment. There are common patterns of quality states in oral presentations: some presenters start strong and become fatigued (decreasing quality), some start anxious and become more comfortable once they adapt to their audience (increasing quality), some become most anxious at the end of their presentation when they fear negative evaluation [6], and many speakers maintain a constant quality throughout. We seek to model these common sequences using Hidden Markov Models (HMMs) and use them to both provide more meaningful feedback

to speakers and improve the accuracy of automated quality assessment.

In the rest of this paper, we briefly review related work, then describe our corpus of presentations, features, and judge ratings used in the modeling work. We then present the results of our HMM models and how these compare to more standard stateless quality assessments. We also describe an analysis of the relationship of segment quality ratings to aggregate ratings, before concluding.

2 Related Work

Prior research has approached assessing public speaking performance in multiple ways and shown the potential of using machine learning algorithms on basic features of oral presentations to predict audience perception and presentation quality. Goberman et al. looked at the relationship of acoustic characteristics of public speaking with self and audience ratings of anxiety, finding that acoustic characteristics are significantly related to ratings of anxiety, and that analysis of speech characteristics can be sensitive to anxiety differences [7]. Curtis et al. analyzed the relationship between visual and acoustic features of scientific talks and judges' ratings of presenters and their audience [8]. They used the analysis to build a classifier to predict audience engagement in presentations. Tanveer et al. analyzed a dataset of over 2200 TED talk transcripts, audio features, and viewer ratings to create a prediction model for human ratings [9]. Similarly, Sharma et al. analyzed visual cues related to facial and physical appearance, facial expressions, and pose variations within a database of over 1800 TED talk videos and their viewer ratings on YouTube to predict the popularity of the presentations [10].

The approach to automated multimodal presentation scoring by Ramanarayanan et al. is closely related to the current research [11]. Their work applied multimodal features of speech, head orientation and gaze, body movements, and facial emotions to automate the scoring of presentations. Batrinca et al. similarly automated the overall assessment of presentations given in front of a virtual audience with the Cicero platform using multimodal features [12]. Chen et al. formulated a scoring model based on content, delivery, hand, body, and head movements to predict human rating [13]. Wörtwein et al. modeled the assessment of audiovisual nonverbal behavior of the presenters and used it to drive the behavior of a virtual audience [14]. With ROC Speak, Fung et al. combined automated analysis with crowdsourced human analysis of smiles, body movement, and volume modulation within recorded presentations [15]. While these approaches looked at time-varying features, they were used to predict human judge ratings of entire presentations and not of individual segments. On the other hand, Chollet and Scherer studied segments of a presentation for predicting audience ratings of the talks [16]. However, the segments were not analyzed in a holistic manner compared to the full presentations, and similar to the previous studies, their models could not provide feedback to the presenters about which parts of their presentation were particularly good or problematic.

Some work has looked at providing time-based feedback to presenters but only at the individual feature level rather than estimated audience ratings of quality. The MACH system provides summarized and focused feedback on smiles, head movement, audio, and speech features [5]. RoboCOP uses a robotic presentation coach to provide feedback on content coverage, speech features, and eye gaze after each slide and at the end of the presentation [1]. The AutoManner system provides feedback to make presenters self-aware of their body language [2]. The Presentation Trainer system provides real-time visual and haptic feedback for speech features, gestures, and posture during rehearsals [17]. The Rhema [3] and Logue [18] systems provide real-time presentation feedback on features such as speech, energy, and body openness using the Google Glass platform. These systems provide detailed feature-level feedback to presenters without providing a computed estimate of the overall quality or audience perception and ratings, leaving the interpretation of presentation quality to the user.

3 Multimodal Presentation Corpus

For this work, we used a corpus of audio-video recordings of oral PowerPoint presentations conducted in a controlled lab setting.

3.1 Data Collection

3.1.1 Study Setup

The presentations were performed in a room equipped with video and audio recorders, a large screen to display the slides, a podium and a laptop which the presenters used to control their slides, and took place in the presence of one to three live audience members sitting in front of the presenter, trained to remain neutral.

3.1.2 Participants

The presentations were part of a public speaking anxiety intervention study [19]. Participants were asked to deliver two recorded presentations on prepared topics (nutrition and exercise). The prepared slide decks contained 6-7 slides and speaker notes, and participants were given 20 minutes to prepare and rehearse. The 41 participants (15 female and 26 male) were mostly university students with presentation experience, but also included five professional actors (2 female, 3 male) delivering the same presentations. The study was approved by the University IRB, and participants were compensated for their time.

3.1.3 Corpus Curation

Our full corpus comprised of 92 presentations, of which we selected only those 74 presentations lasting 4.5 minutes or longer and for which participants consented to have their presentation data used for secondary analyses. Presentations lasted 4.5 to 6 minutes, with the corpus totaling 7.1 hours in length.

3.1.4 Segmentation

To analyze and model the time-varying qualities of the presentations, we split all presentation videos into one-minute segments based on subtopics, resulting in 413 total segments.

3.2 Data Labeling of Presentation Quality

We were interested in estimating lay audience quality ratings of the presentations, using Amazon Mechanical Turk workers (Turkers) as judges. We limited recruitment to judges in the US, having approval ratings over 95%, and performed several quality checks on the results. Judges rated presentations and presentation segments using a custom 28-item, seven-point Likert scale instrument, with items spanning presenters' internal state (e.g., nervousness), gaze, audio, gesture, audience interaction, and proxemic behaviors. The instrument demonstrated strong reliability (Cronbach's α =0.96). We also assessed inter-rater reliability using the intraclass correlation coefficient (ICC) of the overall presentation and presentation segment ratings, retaining the averaged ratings of the 3 judges with the highest ICC values, yielding a mean ICC of 0.5.

We found that, for each topic (exercise or nutrition), all actor presentations were given the highest overall presentation quality ratings. We also found a significant moderate correlation between the averaged ratings of all segments in a presentation and the overall presentation rating, r=0.477 p<.001. These lend additional validity to the ratings.

3.3 Multimodal Feature Extraction

We used a combination of automated and manual feature extraction methods on the overall presentations and presentation segments to support our modeling work.

3.3.1 Prosodic Features

Using the COVAREP toolbox [20], we extracted speech features that characterize glottal flow: NAQ, QOQ, H1H2, PSP, peakSlope, Rd, Rd conf, HMPDM 1-24, HMPDD 1-12 as well as MFCC 1-12 and MCEP 0-24 features. Glottal flow is known to vary significantly with changes in phonation type [19]. We computed mean, std, min, and max, of these features. Using the Praat toolbox [21], we extracted the fundamental frequency-F0, formants 1-5, and voice intensity and computed their mean, std, min, max, and range. We also extracted pause ratio, harmonicity, shimmer, jitter, and degree of voice breaks from each audio file. In total, we extracted and computed 188 prosodic features.

3.3.2 Disfluency Features

Speech disfluencies may be important markers of poor presentation quality [20]. One study found that the rate of disfluencies (repetitions, repairs, and filled pauses) negatively correlated with the charisma rating of public speakers [21]. Starting with verbatim transcripts of the presentations, five coders independently annotated the following features in the transcripts. 1) filler words (e.g., um, uh, like), 2) inaudible (words or phrases that were unable to be heard), 3) repetitions and corrections, 4) incomplete sentences, 5) false starts (beginning an utterance and subsequently aborting it before completion).

3.3.3 Non-Verbal Features

Presenters' nonverbal behavior can impact the perceived quality of their presentations [22]. We manually extracted nonverbal features indicating gaze behaviors, facial expressions, and hand gestures, coding the count, total duration, and average duration for each, resulting in 30 features in total. Turkers were provided with a user interface to mark the start and end times of each behavior as they watched a presentation video. Each video was annotated by three Turkers, with their ratings averaged. The annotated nonverbal features are: 1) gaze at audience, 2) gaze at notes, 3) gaze away, 4) gaze at slides (display), 5) smile, 6) head-nod, 7) cross arms and hands at back or hands in pocket, 8) moving on stage 9) posture shift and 10) pointing (deictic gesture).

4. Presentation Quality Assessment Framework

Here we describe the development of our Hidden Markov Models for automatically assessing the quality of oral presentation using the multimodal features we extracted.

4.1 Feature Pre-Processing

We concatenated all features described above to obtain one combined feature vector of 204 features. We normalized all features to have zero mean and unit variance, to remove any possible bias related to the range of values associated with a feature. To train and validate our HMM models, we used 413 video segments from the presentations described in 3.1. Given the large ratio of features to sample size, we reduced the size of our feature set by computing mutual information scores between each feature and the overall presentation quality rating and selected 20% of the highest scoring features.

4.2. Presentation Quality Assessment Models

We trained an HMM in addition to a stateless classification model to predict presentation segment quality scores. To derive the states for the HMM and prediction class labels, we created a frequency table of the ground truth presentation segment quality ratings and partitioned it at the 33rd percentile and 66th percentile to divide the distribution into three categories: Low, Medium, and High quality.

4.2.1 Baseline Stateless Model - Random Forest Classifier

For the baseline stateless model, we evaluated several classification models, including Support Vector Machine classifier (SVC), Random Forest (RF) Classifier, and Gradient Boosting Classifier. For model training and evaluation, we first divided our dataset into training and test sets with a ratio of 7:3. The Random Forest Classifier generally performed the best with our datasets, and thus we used it to train our baseline stateless model. Using Grid Search with 5-fold cross-validation, we tuned five hyperparameters of our random forest classifiers, including the number of tree estimators, maximum depth of individual tree estimators, the minimum number of samples required to split an internal node in a tree, and the minimum number of samples required to be a leaf node in a tree. We evaluated the performance of our classification model on our held-out test set, using four metrics: accuracy, F1, precision, and recall.

4.2.2 Hidden Markov Model

We model oral presentations as a sequence of quality states that a presenter is in, based on each segment of the presentation. We

assume that the presenter's next state not only depends on the features corresponding to their presentation but also the audience's perception of the presentation at the previous time step. This represents a classic Markov chain of a sequence of presenter's presentation quality rating states.

We divided our dataset of 74 presentations into training and testing sets in the ratio of 7:3. For each presentation, the dataset consisted of videos corresponding to the 1-minute topic segments. Using features and presentation quality ratings (Low, Medium, High) corresponding to each presentation segment, we then defined an HMM as follows:

- A set of three hidden states: Low, Medium, High
- A transition probability matrix: representing the probability of a presenter going from one hidden state to the other
- A sequence of observations corresponding to each state: at every time step we considered disfluency, audio, and nonverbal features as the observables
- An initial probability distribution over the states
- Emission probabilities: that estimate the likelihood of a feature being observed at a particular state

We estimated the HMM emission and transition probabilities using the Expectation-Maximization algorithm.

5. Results

5.1 HMM-based Presentation Quality Prediction

We evaluated the ability of the HMM to predict presentation segment quality, to determine whether this approach yielded greater accuracy than a standard stateless estimator. Table 1 shows the results of the HMM compared to our best Random Forest classifier. For the three-class presentation-quality prediction, the HMM achieved an accuracy of 0.627 (F1 = 0.627), while the RF model achieved an accuracy of 0.504 (F1 = 0.500).

Table 1: Performance of presentation quality assessment models trained on presentation segments

Model	Accuracy	F1-Score	Precision	Recall
Baseline (RF)	0.504	0.500	0.500	0.504
HMM	0.627	0.627	0.628	0.630

5.2 Relationship of presentation segment ratings to overall presentation ratings

We also analyzed the relationship of the quality ratings of the segments to those of the overall presentations. To understand this temporal relationship, we calculated the correlation between the ratings for each segment and ratings for the entire presentation (Figure 1). Ratings of the second segment and the last segment of the presentation were significantly correlated with the overall presentation quality rating. Figure 1 also shows that a similar correlation pattern holds for ratings of other traits of the presenter's behavior, such as gesture, voice, gaze, audience interaction, internal state (e.g., anxiety and competence), and proxemics.

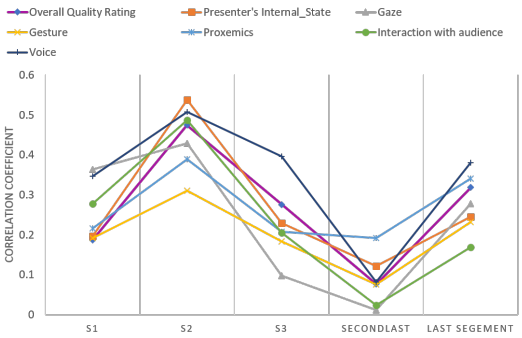

Figure 1: Correlation between ratings of presentation segments and overall presentation

7. Conclusion

We found that modeling the quality states for a presentation leads to increased quality prediction accuracy over stateless predictions, providing validation for this general approach. We also found that the final segments of a presentation have the highest correlation with overall presentation quality ratings, further validating the segment-based analysis. Our findings corroborate previous work [16] that found a significant correlation between ratings of randomly selected segments of a presentation and overall presentation ratings.

This work is an important step towards providing automated feedback to presenters, not only on the quality of their overall presentation but on the quality of each segment of their presentation. This can allow them to more readily identify the parts of their talk that need attention. This work can also pave the way towards identifying common temporal trajectories of presentation quality that could be used to provide more tailored diagnostics and corrective action, e.g., by identifying individuals who become anxious towards the end of their talk, or who need time to become comfortable with their audience.

Significant work remains to develop this approach into a fully automated presentation feedback system. First, our manual feature extraction steps must be automated. Several other researchers have automated gaze, gesture, and proxemic feature identification using Kinect data analytics [12, 13, 16], leading the way for this development. Second, our topic segmentation was performed manually, greatly aided by having pre-defined presentation media for presenters. Automated approaches to topic segmentation must be developed, taking into account segment boundaries that are useful and meaningful for presenters. Slide and slide section specifications in slideware (such as PowerPoint) may provide useful starting points for this. Finally, interactive visualizations that provide segment ratings to speakers, with the ability to drill down into the features that support the ratings along with layperson recommendations, need to be developed.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant Number IIS-1514490.

REFERENCES

[1] Trinh, H., Asadi, R., Edge, D. and Bickmore, T. RoboCOP: A Robotic Coach for Oral Presentations. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1, 2 (2017), Article 27.

[2] Tanveer, M. I., Zhao, R., Chen, K., Tiet, Z. and Hoque, M. E. AutoManner: An Automated Interface for Making Public Speakers Aware of Their Mannerisms. In *Proceedings of the Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA, 2016). Association for Computing Machinery.

[3] Tanveer, M. I., Lin, E. and Hoque, M. Rhema: A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking. In *Proceedings of the Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA, 2015). Association for Computing Machinery.

[4] Grosz, B. and Sidner, C. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12, 3 (1986), 175-204.

[5] Hoque, M., Courgeon, M., Martin, J.-C., Mutlu, B. and Picard, R. W. MACH: my automated conversation coach. In *Proceedings of the Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (Zurich, Switzerland, 2013). Association for Computing Machinery.

[6] Witt, P., Brown, K., Roberts, J. and Behnke, R. Somatic Anxiety Patterns Before, During, and After Giving a Public Speech. *Southern Communication Journal*, 71, 1 (2006).

[7] Goberman, A. M., Hughes, S. and Haydock, T. Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech Communication*, 53, 6 (2011/07/01/ 2011), 867-876.

[8] Curtis, K., Jones, G. J. F. and Campbell, N. Effects of Good Speaking Techniques on Audience Engagement. In *Proceedings of the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA, 2015). Association for Computing Machinery.

[9] Tanveer, M. I., Hasan, M. K., Gildea, D. and Hoque, M. E. A Causality-Guided Prediction of the TED Talk Ratings from the Speech-Transcripts using Neural Networks. *arXiv preprint arXiv:1905.08392* (2019).

[10] Sharma, R., Guha, T. and Sharma, G. *Multichannel Attention Network for Analyzing Visual Behavior in Public Speaking*. City, 2018.

[11] Ramanarayanan, V., Leong, C. W., Chen, L., Feng, G. and Suendermann-Oeft, D. Evaluating Speech, Face, Emotion and Body Movement Time-series Features for Automated Multimodal Presentation Scoring. In *Proceedings of the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA, 2015). Association for Computing Machinery.

[12] Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P. and Scherer, S. *Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training*. Springer Berlin Heidelberg, City, 2013.

[13] Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C. and Lee, C. M. Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues. In *Proceedings of the Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey, 2014). Association for Computing Machinery.

[14] Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R. and Scherer, S. Multimodal Public Speaking Performance Assessment. In *Proceedings of the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA, 2015). Association for Computing Machinery.

[15] Fung, M., Jin, Y., Zhao, R. and Hoque, M. ROC speak: semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Proceedings of the Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan, 2015). Association for Computing Machinery.

[16] Chollet, M. and Scherer, S. *Assessing Public Speaking Ability from Thin Slices of Behavior*. City, 2017.

[17] Schneider, J., Börner, D., Rosmalen, P. v. and Specht, M. Presentation Trainer, your Public Speaking Multimodal Coach. In *Proceedings of the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA, 2015). Association for Computing Machinery.

[18] Damian, I., Tan, C. S., Baur, T., Schöning, J., Luyten, K. and André, E. Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques. In *Proceedings of the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea, 2015). Association for Computing Machinery.

[19] Kimani, E., Bickmore, T., Trinh, H. and Pedrelli, P. *You'll be Great: Virtual Agent-based Cognitive Restructuring to Reduce Public Speaking Anxiety*. IEEE, City, 2019.

[20] Strangert, E. and Gustafson, J. *What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations*. City, 2008.

[21] Hirschberg, J. B., Biadsy, F., Rosenberg, A., Carlson, R. and Strangert, E. A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech (2008).

[22] Slutsky, J. *The Toastmasters International Guide to Successful Speaking: Overcoming Your Fears, Winning Over Your Audience, Building Your Business & Careeer*. Dearborn Trade Pub, 1997.