Effect of Modality on Human and Machine Scoring of Presentation Videos

Haley Lepp, Chee Wee Leong, Katrina Roohr, Michelle Martin-Raugh & Vikram Ramanarayanan
Educational Testing Service R&D
<hlepp,cleong,kroohr,mmartin-raugh,vramanarayanan>@ets.org

ABSTRACT

We investigate the effect of observed data modality on human and machine scoring of informative presentations in the context of oral English communication training and assessment. Three sets of raters scored the content of three minute presentations by college students on the basis of either the video, the audio or the text transcript using a custom scoring rubric. We find significant differences between the scores assigned when raters view a transcript or listen to audio recordings in comparison to watching a video of the same presentation, and present an analysis of those differences. Using the human scores, we train machine learning models to score a given presentation using text, audio, and video features separately. We analyze the distribution of machine scores against the modality and label bias we observe in human scores, discuss its implications for machine scoring and recommend best practices for future work in this direction. Our results demonstrate the importance of checking and correcting for bias across different modalities in evaluations of multi-modal performances.

KEYWORDS

presentation scoring, human-computer interaction, multimodal assessment

ACM Reference Format:
Haley Lepp, Chee Wee Leong, Katrina Roohr, Michelle Martin-Raugh & Vikram Ramanarayanan. 2020. Effect of Modality on Human and Machine Scoring of Presentation Videos. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3382507.3418880

1 INTRODUCTION

Presentations are an inherently multimodal activity. Delivering an effective presentation requires the appropriate and synergistic use of delivery (fluency, pronunciation, rhythm, intonation, stress, etc.), language use and content (grammar, vocabulary, topic development, argumentation, discourse, etc.) and interactive gestures or kinesics to maintain engagement and rapport with the intended audience.

Multiple studies have also observed the complementarity of using features of various modalities in *automated* scoring of oral communication tasks [6, 10, 11, 14, 15, 17, 20]. Widely-used automated scoring algorithms are trained to consume these features and predict scores assigned by human expert raters.

However, the scores used in all these studies to train the algorithms are the result of human reviewers observing video data in its entirety in order to judge various aspects of oral communication proficiency, including those that can be judged purely based on the content of the presentation, rather than the kinesics or delivery thereof. None of these studies examined how raters were influenced by the modality of the data presented to them, and how the modality a human rater observes can affect the distribution of predicted scores.

A comprehensive understanding of the interplay between modality and proficiency judgement remains an important gap in our understanding of human and automated scoring of oral communication performance with practical implications – deciding whom to admit to a university, whom to hire in job interviews [4, 10], or whom to vote for as the best performer in a debate [3] – to name a few. Indeed, even as schools increasingly use online tools which assess and teach learners through audio, text, and video interactions, understanding the relationship between modality and assessment is of urgent import.

This study bridges this gap by examining how the modality of data – be it text only, audio only or the entire video – presented to a rater affects his/her score judgement. It further examines how modality-specific machine learning features and models perform comparatively, and discusses the implications for both human and machine scoring of oral communication tasks that are inherently multimodal in nature.

2 DATA

2.1 Task

The task required participants to record an informative video for a group of high school freshmen about what to consider when selecting and applying to college. The prompt encouraged participants to concentrate on providing the necessary information to increase the high school students' knowledge and understanding of the topic, rather than trying to persuade them to go to college. It further instructed participants to talk about at least two factors – one to consider early in high school and one later – and to support these points with personal experiences and/or other examples. The task therefore involved (i) reviewing the list of things to consider when selecting and applying to colleges, (ii) preparing their presentation (5 minutes), and (iii) the actual presentation itself for three minutes. It is worth noting that while we collect video/audio data of the participant during the entire task, we only analyze the three minute presentation.

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

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418880

2.2 Collection

Our data consists of 80 videos of oral presentations completed by college students from the United States who were recruited through Amazon Mechanical Turk. Participants interacted with HALEF[1], an open-source modular cloud-based dialog system that is compatible with multiple W3C and open industry standards [18], that delivers the presentation task via a call-in webpage provided to the participants. The HALEF dialog system logs speech data collected from participants, which is then transcribed and scored. Participants had five minutes to prepare their response to the oral presentation prompt, three minutes to deliver their presentation, and 4-7 minutes for background questions. We paid US$3.50 for completion of the 12-15 minute task.

3 HUMAN SCORING

Human raters scored each presentation using a content scoring rubric on a 0 to 4 scale, where 0 is "off topic," 1 is "deficient," 2 is "weak," 3 is "competent," and 4 is "proficient." For scores of 1 to 4, raters evaluated whether the speaker demonstrated effective organization, transitions, and time management, and whether they tailored the presentation to the targeted audience using informative language. Speakers were also evaluated on whether they supported the presentation with relevant information as indicated in the prompt and using personal/outside experiences. The scoring rubric, which is the same with each modality, was developed based on a synthesis of existing scoring rubrics in the oral communication literature and with experienced assessment developers [19]. The rubric only covered the content dimension of oral communication, and not the delivery dimension.

All presentations received six scores: two raters with high-levels of experience scoring spoken responses scored each of the three modalities. A third rater was brought in if presentations had score discrepancies of 2 or greater. All raters held at least a master's degree, and spoke English as their native language. Five of the six raters were female. For scoring, two raters were randomly assigned to one of three scoring modalities: 1) video, 2) audio only, or 3) transcript only. After being assigned, each rater completed an online training session where they were familiarized with the task, the custom-designed content scoring rubric and scoring modality. Raters were then given two opportunities to calibrate on practice videos, audio files, and transcript files prior to completing the scoring activity. Raters scored the videos in two separate, roughly equal batches. After completing the scoring, raters were asked to provide their confidence rating on a scale of 0 to 100. Across modalities, the average confidence rating was 77.5, 59.5, and 88 for video, audio only, and transcript only, respectively.

4 ANALYSIS OF MODALITY BIAS IN HUMAN SCORING

We use the rounded-down median of the two (or rounded median of three) scores for each presentation, and combine the "weak" and "deficient" score classes into a single class, given the smaller number of video samples that were assigned these scores. As shown

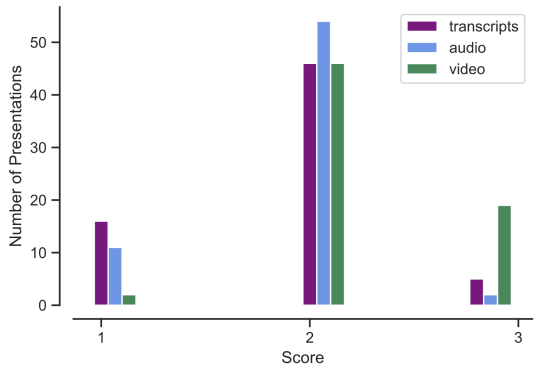

Figure 1: Distribution of scores by modality

by Figure 1, the scores are primarily centered in the middle, with very few presentations having median score of 3.0, or "proficient."

As can be seen in in Figure 2, the video modality has relatively more high scores, the audio is relatively centered, and the transcript modality has the most low scores. One reason for this could be that the formality of language expected by audiences for text is generally higher than what one might expect in speech. For example, one presentation which received a score of 1.0 for transcript but a 3.0 for video includes a significant number of disfluencies and reductions ("*Um* and then yo- from there you're *gonna wanna* start planning *um* around your future."). While these elements may look inappropriate in a text transcript, they may be perceived as less so in the audio or video modalities.

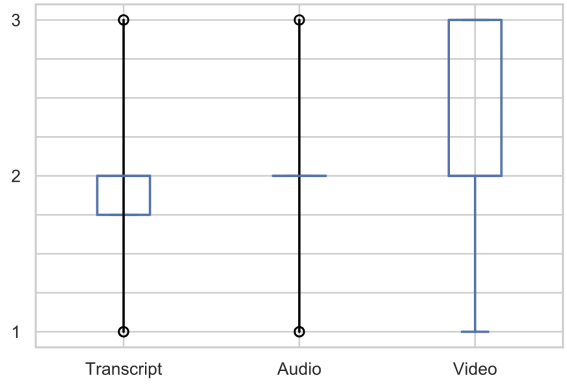

Figure 2: Distribution of scores within each modality

To determine whether there is a significant difference between the distribution of human scores by modality, we ran a Kruskal-Wallis test, or non-parametric 1-way ANOVA, which demonstrates a significant difference in distributions with a $p<0.00001$.These results are further bolstered by post-hoc Wilcoxon rank-sum tests, which indicate no significant differences between the medians of transcript and audio scores ($p=0.670$), but significant differences between those of video scores and the other two modalities ($p<0.01$). χ^2 tests of differences in categorical distributions also confirm similar patterns in modality-specific median differences at the $\alpha = 0.99$ significance level. From these tests, it is clear that there is a bias toward higher

[1] http://halef.org

scores when the video modality is scored, in comparison to other modalities.

We computed inter-rater reliability for each of the modalities using the 0 to 4 scale with the original two raters, as shown in Table 1 [2]. We observed the highest inter-rater reliability for the transcript modality, followed by video and audio. It is interesting that the video modality has a relatively low agreement even though coders of videos had relatively high confidence in their ratings. Coders may have been more confident because they had additional information around nonverbal communication that could help to ease their understanding, as has been found in other studies [2]. However, since the ratings were about content, not about delivery, the inclusion of the video may have introduced additional sources of bias that could have influenced rater decisions and ultimately impacted their level of agreement.

Table 1: Inter-rater agreement statistics.

Metric	Video	Audio	Transcript
ICC (two-way random, type=consistency)	0.502	0.481	0.694
Quadratic Weighted Kappa (QWκ)	0.28	0.27	0.65
Exact Agree	0.543	0.531	0.407
Adj Agree	0.309	0.383	0.407
Ex+Adj Agree	0.852	0.914	0.926
Spearman ρ	0.365	0.39	0.533

5 ANALYSIS OF MODALITY BIAS IN MACHINE SCORING

The previous section demonstrates a clear difference in the way human raters perceive and score the proficiency of content in informative presentations. This section extends these findings to analyze how this bias could potentially affect machine scoring and examines the efficacy of machine learners to predict scores of presentations by emulating the type of modal information a human scorer would observe when reading a transcript, listening to a recording, or watching a video. Note that the aim of these experiments is *not* to push the state of the art in presentation scoring, but to investigate the effect of modality bias. We therefore focus on feature sets and learners that have been shown to be effective in previous work on multimodal scoring tasks.

5.1 Feature Extraction

We compute time-aggregated feature sets based on the features a human would have when rating a specific modality. The number of features increases from text, to audio, to video, just as it would for a human rater. Specifically, the following features are extracted due to their ability to represent the given modality in a time-aggregated manner.

For the **text** modality (transcriptions), we extract document embedding features for **text** computed using the *doc2vec* method. Doc2Vec is an algorithm which creates dense vectors based on variable length sections of words within a document. In this way, the algorithm is able to represent a full document in a numerical vector and capture information beyond simply counts of the words in a document[12].

For the **audio** modality, we extract speech features using the OpenSMILE toolkit [9]. We use the minimalistic set of voice parameters collected in the eGeMAPS feature set, which has been shown to capture physiological changes in voice production and support predictions of emotion and mental state in speech[8]. This modality also includes the text features which capture the content of the presentation.

To capture the **video** modality, we extract 427-dimensional time-series features at a 10 FPS sampling rate using the OpenFace toolkit[1] based on (a) tracking head poses, (b) tracking gaze directions, and (c) facial landmarks in both 2D and 3D estimations, (d) parameters describing both rigid and non-rigid face shapes, and (d) estimation of occurrence and intensity of Action Units. Following [7], we average these 427-D frame vectors using a moving window size of 10 frames. Then we apply an unsupervised clustering method on these averaged vector sequences to find K clusters. We then represent the entire sequence by the corresponding discrete cluster identifiers (called "visual words"). After this process of converting frame-based vectors into 'text documents', we use the term frequency-inverse document frequency (TF-IDF) of all visual word tokens as feature inputs for each video. These visual features are appended to the speech and content features from the **audio** and **text** modalities to form the **video** feature set.

5.2 Machine Learning Models

We use SKLL[3], an open-source Python package that wraps around the `scikit-learn` package [16], to run 10-fold cross-validation experiments based on the three score classes described in Section 4. We train the models using the rounded median of all the modality scores (henceforth, *overall* median score) for a particular presentation as labels, under the assumption that such a label can represent a holistic score across modalities. We evaluate these models in two ways: first, with the *overall* labels described above, and second, based on the *modality-specific* labels (scores assigned based on the modality of features used to train the model). We use the same folds across all experiments. We select folds by randomized stratified sampling to ensure that the fold distributions are equally sampled from all label classes.

For all machine learning experiments with *time-aggregated* features, we experiment with three classification learners: linear support vector machines, random forests, and logistic regression classifiers, and use quadratic weighted kappa as objective functions for optimizing learner performance. We further tune and optimize the free parameters of each learner using a grid-search method.

5.3 Dealing with Label Bias

Recall from Figure 1 that the score distributions for each modality are not uniform, and we observe that machine learning models trained directly on our data *as is* predict the majority class for nearly all samples in the test set of each cross-validation fold.

To accommodate for the limited number of samples for the lowest and highest scored presentations, we expand our dataset by adding synthetic samples of these categories using Gaussian noise

[2]Note that we brought in a third rater to resolve any discrepancies larger than 2.

[3]`https://github.com/EducationalTestingService/skll`

Table 2: Automated score prediction results on our augmented dataset with modality-specific feature sets. We trained on the overall median scores across all modalities for all experiments reported, and report tests calculated with overall and modality-score labels.

System		Overall Score			Modality Score	
Modality	**Features**	**Learner**	**Acc.**	**QWκ**	**Acc.**	**QWκ**
Text	doc2vec,	Linear SVC	0.535	0.426	**0.462**	0.113
		Logistic Regression	0.489	0.394	0.275	0.104
		Random Forest	**0.561**	**0.539**	0.262	**0.130**
Audio	doc2vec,	Linear SVC	**0.584**	0.394	**0.663**	**0.202**
	eGeMAPS	Logistic Regression	0.545	0.306	0.600	0.033
		Random Forest	0.528	**0.509**	0.237	0.088
Video	doc2vec	Linear SVC	0.403	0.247	0.550	0.164
	eGeMaPS,	Logistic Regression	0.502	0.428	**0.525**	0.172
	OpenFace	Random Forest	**0.582**	**0.568**	0.350	**0.256**

augmentation. For each fold, we randomly sample with replacement from the minority classes within that fold[4], until each class has equal number of instances for every class, and perturb each sample with mean-zero Gaussian noise with a standard deviation corresponding to the sample standard deviation of that class. In other words:

$$synthetic\ sample = current\ sample + \sigma\mathcal{N}(\mathbf{0}, \mathbf{S})$$

where S is the standard deviation of the class being sampled from and σ is a multiplication factor (for the purposes of our experiments, we set $\sigma = 0.01$. Using this method, we generate ten folds of approximately 18 instances each, with 6 instances for each of the three classes.

5.4 Observations and Results

Table 2 shows the results of our automated content scoring experiments. In the Modality Score column, we report the results of training classifiers on the Overall Score and testing on the score related to the specific modality.

We observe that performance of the best learner in the case of the video feature set is generally higher than that of the text feature set, both in terms of quadratic weighted kappa (QWκ; recall that we optimized on this) and accuracy. This trend is also consistent both when we test on the overall score or the modality-specific score label. However, things do not pattern as clearly in the case of audio features. To investigate this further, we plotted the distribution of scores by modality as predicted by the Linear SVC classifier in Figure 3. This suggests that the distribution of scores across modality as predicted by the automated scoring models demonstrates a divergence from the distributions of modality-specific human scores. In particular, both the skew toward high scores in the video modality and the skew toward low scores in the text modality is not as acute in machine scores relative to the human score distributions.

In other words, we observe clear differences between how features extracted from different modalities predict the overall score versus scores assigned specifically to their counterpart modalities.

[4]Performing this step for every fold ensures that no generated instance ends up in a different fold than the original instance it is based upon (to ensure we aren't training and testing on the "same" data).

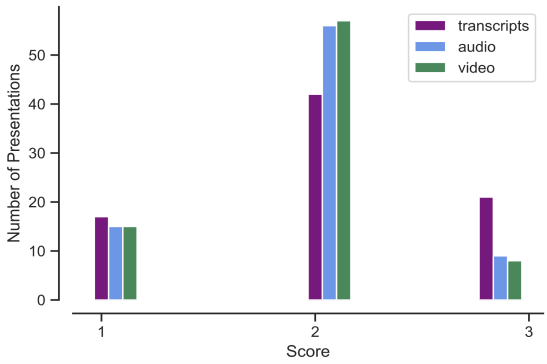

Figure 3: Distribution of scores by modality predicted by LinearSVC learners.

6 DISCUSSION: IMPLICATIONS OF BIAS FOR MULTIMODAL SCORING

Our work exposes multiple sources of bias that need further exploration and consideration for automated multimodal data analysis.

Chief among these is the issue of **modality bias** in both human and machine scoring. We noticed significant differences between content proficiency scores that human raters assigned to a presentation when they just viewed the video versus when they either listened to the audio or read the transcript. This in turn raises the following question that is crucial for automated machine scoring: *what is the "true/gold" content score to be used for training machine scoring algorithms*? Is it the score assigned just by looking at the text transcript or is it, as we have considered in this paper, the median score of all modalities? If the former, do other modalities bias the rater from the "true" score?

In our case, transcript scoring yielded higher inter-rater agreement and rater confidence. It's possible that the audio/video modalities introduced other factors that influenced the scores, rather than focusing strictly on content. Factors such as the vocal delivery (e.g., tone/pitch, pacing, etc.) could have impacted the way the content was spoken and therefore influenced the rater during the evaluation. Future research should explore whether other reviewer characteristics, such as race and gender, could affect scoring bias.

The second broad issue is that of how to deal with **label bias**, i.e., the problem of having significantly fewer examples of data in some score classes than others. This paper has presented one simple method of dealing with this bias, by artificially augmenting the dataset by perturbing samples with Gaussian noise. We could also investigate the efficacy of alternate methods, such as SMOTE oversampling [5], in future work.

There are two other important sources of bias that we did not focus on in this exposition, but which are nonetheless important to consider – gender and race [13]. Further evaluation of the relationship between the gender and race of the presenters and reviewers in conjunction with the the modality of the presentation is critical to ensuring fairness in scoring. Future studies should explore this interplay and how any biases reflected in human scoring can enter into automated scoring.

The continued study of human bias is crucial to factor in not only for machine scoring, but also for policy and educational research design going forward.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[2] Jorge Beltrán Zuniga. 2019. The Effects of Visual Input on Scoring a Speaking Achievement Test. 16 (09 2019), 1–23. https://doi.org/10.7916/salt.v16i2.1260

[3] Maarten Brilman and Stefan Scherer. 2015. A multimodal predictive model of successful debaters or how i learned to sway votes. In *Proceedings of the 23rd ACM international conference on Multimedia*. 149–158.

[4] J. R. Burnett and S. J. Motowidlo. 1998. Relations between different sources of information in the structured selection interview. *Personnel Psychology* 51, 4 (1998), 963–983.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. 2014. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 200–203.

[7] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 504–509.

[8] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. 2015. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7 (01 2015), 1–1. https://doi.org/10.1109/TAFFC.2015.2457417

[9] Florian Eyben and Björn Schuller. 2015. openSMILE:) The Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records* 6, 4 (2015), 4–13.

[10] Jelena Gorbova, Iiris Lusi, Andre Litvin, and Gholamreza Anbarjafari. 2017. Automated screening of job candidate based on multimodal video processing. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 29–35.

[11] Shan-Wen Hsiao, Hung-Ching Sun, Ming-Chuan Hsieh, Ming-Hsueh Tsai, Hsin-Chih Lin, and Chi-Chun Lee. 2015. A multimodal approach for automatic assessment of school principals' oral presentation during pre-service training program. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[12] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 1188–1196. http://proceedings.mlr.press/v32/le14.html

[13] Chee Wee Leong, Katrina Roohr, Vikram Ramanarayanan, Michelle P Martin-Raugh, Harrison Kell, Rutuja Ubale, Yao Qian, Zydrune Mladineo, and Laura McCulla. 2019. Are Humans Biased in Assessment of Video Interviews?. In *Adjunct of the 2019 International Conference on Multimodal Interaction*. 1–5.

[14] Mohana K Nambiar and Cecilia Goon. 1993. Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal* 24, 1 (1993), 15–31.

[15] Xavier Ochoa, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. 2018. The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th international conference on learning analytics and knowledge*. 360–364.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[17] Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. 2015. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 23–30.

[18] Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Alexei V Ivanov, Zhou Yu, Yao Qian, and Keelan Evanini. 2017. Assembling the jigsaw: How multiple open standards are synergistically combined in the HALEF multimodal dialog system. In *Multimodal Interaction with W3C Standards*. Springer, 295–310.

[19] K. C. Roohr, L. Mao, V. Belur, and O. L. Liu. 2015. Oral communication in higher education: Existing research and future directions. In *Designing next-generation assessments of student learning outcomes in higher education. Symposium conducted at the National Council on Measurement in Education Annual Meeting, Chicago, IL*.

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