DBATES: dataset for Discerning Benefits of Audio, Textual, and facial Expression features in competitive debate Speeches

Taylan K. Sen*, Gazi Naven*, Luke Gerstner, Daryl Bagley, Raiyan Abdul Baten, Wasifur Rahman, Md Kamrul Hasan, Kurtis Haut, Abdullah Al Mamun, Samiha Samrose, Anne Solbu, R. Eric Barnes, Mark G. Frank, Ehsan Hoque, *Member, IEEE*

Abstract—In this work, we present a database of multimodal communication features extracted from debate speeches in the 2019 North American Universities Debate Championships (NAUDC). Feature sets were extracted from the visual (facial expression, gaze, and head pose), audio (PRAAT), and textual (word sentiment and linguistic category) modalities of raw video recordings of competitive collegiate debaters (N=716 6-minute recordings from 140 unique debaters). Each speech has an associated competition debate score (range: 67-96) from experienced judges as well as competitor demographic and per-round reflection surveys. We observe the fully multimodal model performs best in comparison to models trained on various compositions of individual modalities. We also find that the weights of some features (such as the expression of *joy* and the use of the word "we") change in direction between the aforementioned models. We use these results to highlight the value of a multimodal dataset for studying competitive, collegiate debate.

1 Introduction

In the first 1960 United States presidential debate between John Kennedy and Richard Nixon, initial analysis suggested that the radio audience predominantly found that Nixon won the debate, while the television audience found that Kennedy won [1], [2]. Could their use of facial expressions during the debate help explain this? Several studies have established that nonverbal communication, including facial expressions, pose, and speech audio characteristics, often account for a substantial portion of the meaning conveyed [3], [4], [5]. Yet, a fundamental question remains: *How are the different modalities of communication, including textual, auditory, and visual, interdependent in affecting a communication's effectiveness?* In this paper, we present a multimodal, expert-labeled, debate dataset for public release that includes several nonverbal communication data features that have often been omitted from prior studies. We then use this dataset to show how focusing on a single modality of interpersonal communication in isolation (as opposed to considering multiple modalities together), often leads one to develop opposite conclusions as to a feature's association with debate performance. For example, we show that when considering facial expressions alone, one is led to believe that smiling with both mouth and eyes (which is often, but not always, associated with joy) are negatively associated with debate score. However, when considered together with the context of textual word category, sentiment, and speech audio features, smiling with both mouth and eyes is shown to be positively associated with debate score. We present these findings along with several others through examination of the multimodal dataset DBATES: Dataset for discerning Benefits of Audio, Text, and facial Expression features in competitive debate Speeches.

The importance of studying debate goes beyond predicting the outcomes of presidential elections. Studies have shown that education and practice in debate improves one's ability to think critically. For instance, Green and Klug conducted an experimental study to find how debates promoted the quality of critical thinking [6]. They found that students who learned issue through debate had significantly larger increases in critical thinking test performance compared to students in the control group [6]. Another study demonstrated how debate encourages open thought [7]. More specifically, in their study, Kennedy, et al., showed that through a series of in-class debates, 31% to 58% of participants changed their views after participation [7]. This suggests that students were capable of learning new ideas through debate and recognized merit to different viewpoints to the extent that some ended up adopting alternative views. These studies provide evidence that debates play an integral role in learning new and different perspectives. Perhaps, it is not surprising that the majority of policy makers in the three branches of U.S. government, as well as many world leaders and architects of social change including Nelson Mandela and Dr. Martin Luther King, were debaters in school [8].

- *Taylan Sen, Gazi Naven, Luke Gerstner, Daryl Bagley, Raiyan Abdul Baten, Wasifur Rahman, Md Kamrul Hasan, Kurtis Haut, Abdullah Al Mamun, Samiha Samrose, and Ehsan Hoque are with the Department of Computer Science, University of Rochester, New York.*
- *Anne Solbu and Mark G. Frank are with the Department of Communications, University at Buffalo, New York.*
- *R. Eric Barnes is with the Department of Philosophy, Hobart and William Smith Colleges, New York.*

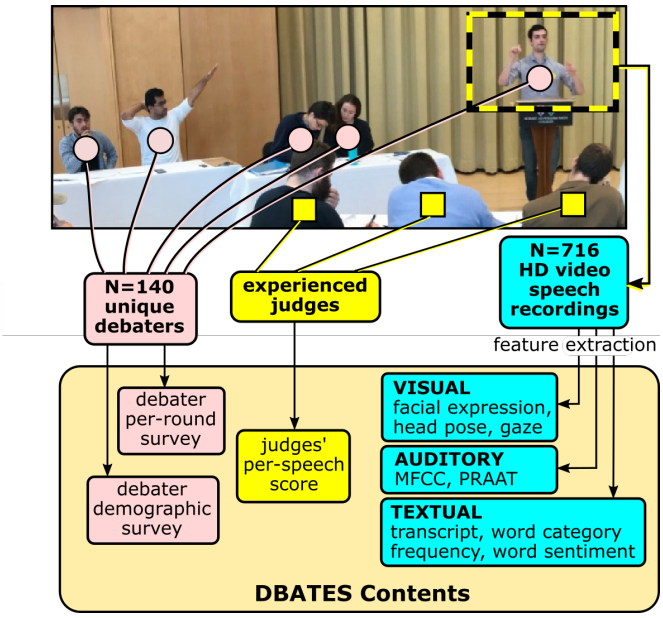


Fig. 1. **DBATES Data Collection and Dataset Contents.** Here we show a portion of a live debate scene at the North American Universities Debate Championships (NAUDC) at Hobart and William Smith Colleges from which the DBATES dataset is based. The yellow and black rectangular box represents the camera view of recordings from which communication features have been extracted. Extracted features include textual, auditory and visual modalities. Each speech also has an associated judge's score and debater survey.

While these studies make evident the importance of studying debate, the verbal and nonverbal factors which make a good debater are not as clear. As will be shown in the Related Work section, there is currently a lack of published debate datasets that allow for the study of facial expressions and gestures in a multimodal manner that could resolve these. In this paper we address these issues, with our major contributions comprising:

- providing DBATES - the Dataset for discerning Audio, Textual, and facial Expression features in competitive debate Speeches from the collegiate North American Universities Debate Championships (NAUDC) at urdebate.org,
- finding linear associations of multimodal debate features with expert-judged performance, with each feature considered independently with score (unimodal model) as well with all features simultaneously (multimodal model), and
- identifying that it is necessary to consider audio, facial expressions, and textual features simultaneously to avoid incorrect interpretations in the association of each feature with debate score.

2 Related Work

Several groups have made public debate-related datasets (see Table 1). The Green Persuasive Dataset ("GPD") (2007) is part of part of the larger HUMAINE dataset for the study of emotional data [9]. GPD contains audio-video recordings of eight discussions between two people lasting approximately 30 minutes each [9]. In the discussions, a single *persuader* with an attested genuine pro-green political beliefs tries to convince a *persuadee* to adopt a greener lifestyle [10]. In each of the eight discussions the persuader is the same person. Video recordings of the participants were made by cameras at a 45 degree side angle, which limits the ability of facial analysis tools [10]. The dataset includes ratings from the eight persuadees of how persuasive they found the persuader. While the GPD dataset includes audio and video, and the potential to extract text, is utility is limited by its size in having only a single persuader and only eight persuadees.

The Canal 9 Political Debates Dataset (2009) comprises 70 televised French language audio video recordings of real-world moderated Swiss political debates between two to four participants [11]. Participants were often, but not always politicians. Participants were presented with a yes/no type political question which was debated for on average 37 minutes. Audio and video was captured with 720x576 pixel DV recordings (PAL). The recording was live-edited from multiple cameras in which the view changed from full group views, individual participant views, and subgroup views. Thus, not all participants are visible at all times, and often the camera view is at an angle which does not support facial expression analysis [10]. The Canal 9 dataset includes manual annotations for segmentation of the video indicating which frames involve multiple participants vs. a single participant, and when a single participant, which participant it is. Similarly, the audio stream is annotated to indicate who is speaking. While the Canal 9 dataset contains audio, video, and the potential to extract text, it is limited in that there is not a consistent, individual video recording for each participant. Thus, any automated facial expression analysis will be limited to only portions of the debate. Additionally, the Canal 9 dataset does not contain any ratings of each speaker's performance.

The Internet Argument Corpus (IAC), released in 2012 contains 390,704 posts from 11,800 discussions, sourced from 4forums.com, an online debate forum [12]. This dataset contains a subset of 103,206 posts and offers a diverse set of labeled annotations across different metrics. While the IAC provides a relatively large sample set, it is not focused on debate nor does it involve modalities other than text.

IBM has been working on "Project Debater" over the past 8 years, which includes a number of textual and audio samples in the study of debate [13]. While Project Debater has a number of datasets with over 800 speeches, this project does not include any visual data. Mirkin, et. al created the public Recorded Debating Dataset in 2017, including 60 speeches by professional debaters on various topics comprising audio and text modalities. In 2018 Zhang et. al released ArgRewrite [14], an argumentative writing dataset that contains 180 essays with custom content and surface annotations.

It should also be noted that while not directly related to debate, there are some multimodal datasets have been made available to examine human interaction involving emotion. For example, the SEMAINE database provides 959 emotionally rich acted conversations approximately 5 minutes each [15], [16].

Unfortunately, within the above corpora of debate-related datasets no single dataset has 1.) audio, video, and text modalities, 2.) an experienced evaluator rating of

TABLE 1
Existing Datasets: The previous premiere datasets on debate in comparison to the DBATES dataset.

Year	Dataset	Datapoints	Modalities
2007	Green Persuasive Dataset	8 discussions	Audio & Video Recorded at 45 degree angle
2009	Canal 9 Political Debates	70 debates	Audio & Video in French
2012	Internet Arg. Corpus	390k Debate Posts	Text
2012-	IBM-Rank-30k	30k Arg. Elements	Text
2012-	IBMPairs	9.1k Arg. Pairs	Text
2012-	IBM-Debater	800 Speeches	Audio, Text
2017	Recorded Debating Dataset	60 Speeches	Audio, Text
2018	ArgRewrite	180 Arg. Essays	Text
2020	**DBATES**	**716 speeches**	**Audio, Text, Visual**

each speaker, and 3.) more than 100 datapoints to enable advanced analysis. Additionally, many of the mentioned datasets lack a high-stakes, competitive atmosphere. The DBATES dataset vitally addresses these limitations by providing a multimodal, expert labelled, database of competitive debate speeches. DBATES fundamentally includes facial expressions and head pose data, which our analysis in the following sections show are fundamental in order to properly interpret textual features. The DBATES dataset thus enables exploring the subtleties of debate by observing the interdependencies that are exclusive to a rich, multimodal dataset.

3 METHODS

3.1 Raw Data & Collection

Data was gathered from the 2019 North American Universities Debate Championship (NAUDC) at Hobart and Williams Smith Colleges. This competition was held over three days and involved a total of 224 students, of which 140 participated in our study (i.e., the number of unique participants that we have at least one speech from).

3.1.1 Recruitment

Ethical approval was obtained from our university Institutional Review Board (IRB) prior to any participant recruitment. Individuals were recruited through email and electronic flyers which were sent to all tournament registrants. Consent of all participants was obtained prior to any recording or surveying. In addition to a global participation consent, participants data was not used unless they also provided additional consent on a per-round basis. Individuals were motivated to participate through being offered i) high quality video recordings of their debate speeches, and ii) $5/debate round for answering a short (<2min) survey.

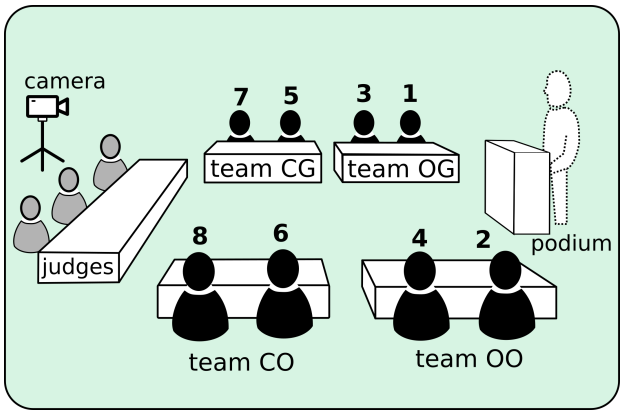


Fig. 2. **Debate Session Room Typical Layout.** Seated at the tables are the two-person teams: Opening Government (OG), Opening Opposition (OO), Closing Government (CG), and Closing Opposition (CO), as well as three to seven judges. The numbers indicate the sequence of the speakers in the debate session.

3.1.2 Debate Format

The tournament followed the British parliamentary debate format for varsity level [17].

Debate Teams. Competitors registered for the tournament as a *debate team* containing two *debater slots*. Although for most teams each member slot was occupied by a different person, in special circumstances, a single person occupied both slots (aka "maverick"). Each debater was designated as either “novice" or “varsity" depending on the debater’s experience. A debater is novice if they are either in their first year of debating at the university level or they have competed in 3 or fewer university level tournaments; otherwise a debater is considered “varsity." A team is a *novice team* only if both members are novice debaters; if one or more team members are “varsity" then the team is designated as a *varsity team*. The level of success at previous tournaments has no impact on novice or varsity status. During the tournament, each debate team took part in a number of individual *debate sessions* held over several *rounds*.

Debate Sessions. Each debate session group consisted of four teams. Each of the four teams were assigned to fulfill one of the four possible *Team Roles* consisting of: *Opening Government*, *Closing Government*, *Opening Opposition*, and *Closing Opposition*. The Government teams were tasked with supporting a given motion, while the Opposition teams were tasked with opposing the given motion. Each of these Teams had two debater slots with specific *titles* which determine the order of speaking. The speaker order and titles followed the following sequence:

While the Team Roles were assigned to each debate team, it was up to each debate team to decide which members would have which title associated with their team (in other words, within a given debate team, the members decided which member spoke in their first slot, and which would spoke in their second slot). Each debate session had its own room which roughly followed the physical layout shown in Fig. 2. The finals and semifinal rounds, however, were held in more auditorium style rooms. Each speech was limited to seven minutes with an additional 15 seconds of grace time. Speakers had the option of accepting questions from

the debaters on the other side of the house.

Debate Rounds. The tournament was organized into six *preliminary rounds* (in which all debate teams competed), followed by several varsity and novice *playoff rounds* for qualifying teams. During each round, a number of concurrent *debate sessions* took place, with each debate session in its own room. Each round had its own *motion* (i.e. issue statement) for teams to debate, and each round lasted approximately one hour. The novice playoff consisted of single elimination semifinals, and finals. The varsity playoff consisted of single elimination quarterfinals, semifinals, and finals. Each motion was secretly selected before the tournament started, by a group of experienced debaters/coaches who were selected as the Chief Adjudicators (CA) for the tournament. Table 2 in the supplementary data shows the motions for each round in the tournament.

At the beginning of each round, all debaters came together to a common auditorium in order to learn their *Debate Session* room assignment and which *Team Role* they were assigned (i.e Opening Government, Opening Opposition, Closing Government, and Closing Opposition). In the first open round (i.e. “Round 1"), varsity debate teams were randomly assigned to a varsity debate session and novice teams were randomly assigned to a novice debate session. In order to ensure each debate session had a full complement of eight teams, noncompeting debaters from the tournament host school were used to fill any remaining open team positions after the initial random assignment. After receiving their debate session room and team role assignment, all debaters hear the round’s motion together at the same time. Debaters are then given 15 minutes to prepare their arguments and report to their designated debate session room.

Although for most teams, each member slot was occupied by a different person, in special circumstances, a single person occupied both slots (aka "maverick"). Thus, each debate group normally consisted of eight competitors. A debater is considered “novice" if they are either in their first year of debating at the university level or they have competed in 3 or fewer university level tournaments. If they are not a novice, then they are considered “varsity." A team is a novice team only if both members are novice debaters. In other words, the level of success at previous tournaments has no impact on novice or varsity status. Only experience is taken into account when assigning these definitions.

There were 6 preliminary rounds in which all teams competed. Top scoring novice teams were entered into a novice playoff, and the top scoring varsity teams were entered into a varsity playoff. The novice playoff consisted of single elimination semifinals, and finals. The varsity playoff consisted of single elimination quarterfinals, semifinals, and finals.Each motion was secretly selected before the tournament started, by a group of experienced debaters/coaches who were selected as the Chief Adjudicators (CA) for the tournament. Table 2 in the supplementary data shows the motions for each round in the tournament. The CA team ensured that the CHAIR judge in every round was an experienced coach or debater. Many of the other judges serving on the panel with the CHAIR judge were also team coaches or experienced debate competition veterans. However, in some cases a few of the panelists were relatively new to the activity.

Judging and Scoring. The Chief Adjudicators ensured that a CHAIR judge in every round was an experienced coach or debater. Many of the other judges serving on the panel with the CHAIR judge were also team coaches or experienced debate competition veterans. However, in some cases a few of the panelists were relatively new to the activity. For the 2019 NAUDC tournament, each speech was provided a score between 50 and 100. The judges were asked to follow scoring guidelines which emphasize the relevance and number of arguments, strength of the reasoning, clarity, completeness, and vulnerability to rebuttal (see Table 4 in the supplemental material for complete guidelines). Each debate group of four teams had three to seven judges. While judges were to each provide their own score for each speech, only the average score among the judges was made public. It should be noted there is some disagreement in the debating community about how debaters should be judged. Until recently, the rules stipulated that matter (i.e., argumentative content) and manner (i.e., effective speaking style) should count equally. However, it has been more common practice for years that the judges have increasingly skewed toward deciding based on matter rather than manner . Still, almost everyone acknowledges that the two qualities are not completely independent. Inevitably good manner impacts what judges perceive as good matter. As a result, there are some judges today who believe that manner is currently undervalued in judging these competitive collegiate debates.

3.1.3 Video Recording

Sony-HDRCX405 HD Video Recording Handycam Camcorders with optical zoom were used to record speeches at 30 FPS in each of the 30 rooms concurrently where the debate rounds took place. These cameras were used to record both the audio and video of the speakers. The cameras were setup using a tripod for steady recording throughout the speech, and all camera operators were trained to make sure the recording captured the speaker from the top of their head to their waist (See fig 2 yellow and black rectangle and fig 3 example). Each speech was captured with a single recording.

3.1.4 Demographic Survey

We developed an online registration application for debaters who were interested in participating in our research study. Through this online portal we also collected information on the debaters age, gender identity, whether they are a native English speaker or not, and their college major.

3.1.5 Per Round Survey

After every round of debate, debaters would leave the room and allow the judges to reflect on the round and discuss scores to be given to each debater. During these 15 minutes we released a timed survey, where we asked debaters post round questions. The questions and the possible answers are shown in Table 3 in the supplementary data.

3.2 Extracted Features

From the raw debate video recordings several data features were extracted including manual text transcription, text sentence sentiment, word category, automated facial

expressions, head pose, and speech audio characteristics. Each of these extracted features are provided in the DBATES database.

3.2.1 Audio Features

We used Praat [18], an application for analyzing the phonetic properties of recorded speech, to extract audio features from the recordings in our dataset. Each video file was converted into a WAV file, then imported into Praat to extract 14 features including *Mean Pitch (F0)*, *Harmonics to Noise Ratio (HNR)*, and different kinds of *Shimmer (local, apq3, apq5, apq11)* and *Jitter (local, absolute, rap, ppq5)*. Praat calculates each of these features over an automatically adjusted time window. In general, PRAAT uses a 0.01 second window unless it computationally identifies that it can save time by using a faster to compute window without sacrificing accuracy[18] To understand how each feature relates to debate speech score, we calculate the average value of the feature for all windows over an entire speech for initial analysis. In addition, we add standard deviation of pitch as a feature. We surmised that pitch and standard deviation of pitch are useful in predicting how dynamic, as opposed to monotone, a given speech is which could affect debate score. The Shimmer and Jitter features have been found to be useful in identifying emotion and stress from speech [19]. Similarly, pitch, the standard deviation of pitch, shimmer, jitter, and HNR have been used to predict deception in corporate executive speech [20].

3.2.2 Text Features

As each speech was recorded as a separate file, each recording was transcribed using a pool of professional transcribers. Each transcription contains both the speech text of the debater who has the floor, as well as the text of the questions (“points of information") raised by opposing debaters that the speaker decided to take. The transcripts contain speaker labels, allowing for the removal of the points of information when analyzing the text spoken by the debater.

The positive and negative sentiment of the speech test was extracted using the sentence-level analysis tool provided by VADER [21]. In summary, VADER calculates positive, negative, and neutral sentiment on the sentence-level through the use of a grammatical and syntactical rule-based model. Thus, VADER is capable of incorporating word order sensitive effects, as well as punctuation as well as slang. Each of the postive, neutral, and negative measures are on a 0 to 1 scale. Instead of a concrete definition of positve, negative, and neutral sentiment, the features are a collective result of a large number of human raters’ understanding of positive, negative, and neutral emotion associated with sentences. Additionally, VADER provides a compound score which which ranges from -1 to 1. VADER data was analyzed on an aggregated debate speech basis, where the sentence-level features were averaged together by speech and each speech was considered a single datapoint.

The Linguistic Inquiry and Word Count (LIWC) [22] tool was used to analyze word semantic category usage within each speech. LIWC measures the frequency of word category usage in text, in which the word categories were designed, established, and validated through several rounds of evaluation by language experts [23]. The frequency of word usage in each of the categories has shown to be an effective way of capturing style and in many cases characterizing/classifying high level behaviors. For our analysis, we normalized the LIWC category counts by the total number of words spoken in each debate transcript. Table 2 provides several examples of the LIWC word categories.

TABLE 2
LIWC Word Category Examples

Category	Examples
Inclusion	with, add, addition, and, plus, we, both, each
Conjunction	for, and, if, but, or/nor, though, till, whether
Motion	walk, run, ride, bring, go, bounce
We	we, us, lets, our, ours, ourselves
You	you, your, yours, yourself, thou, thy
Future Tense	will, won’t, would, may, might
Present Tense	are, aren’t, become, can, need, take, use
Swear	*a selection of common swear words*
Numbers	one, two, thirty, thousand

3.2.3 Facial Expression Features

The OpenFace tool was used to extract 17 facial action coding system (FACS) *action unit* facial expression levels from the raw video of the debaters [24]. The facial action units are provided on a scale from 0 to 5, with 0 representing no expression of the given facial action unit and 5 being the maximum possible intensity of the expression of the given facial action unit. In addition to action units, OpenFace was used to extract eye gaze (x,y) and head pose (6 degree of freedom, i.e. Tx, Ty, Tz, Rx, Ry, Rz).

The Affdex tool was used to extract affective-based facial expression features from the raw debate speech videos [25]. Affdex was been trained on over one million facial expression videos that contain rich affective content, and provides expression levels for expressions commonly associated with *joy*, *fear*, *disgust*, *sadness*, *anger*, *surprise*, *contempt*, *valence*, and *engagement* [25]. These features were extracted at a rate of 30 frames per second over the course of the speeches. Each of these features are provided on a 0 to 100 scale, with 0 representing no expression of that emotion and 100 representing the maximum possible level of expression (with the exception of valence which is measured from -100 to 100).

3.3 Analysis Methods

Several statistical and regression-based analyses were conducted for each of the various feature sets.

3.3.1 Statistical Analysis

In order to identify the differences between high and low scoring debaters, we compared the lower (25th percentile) and upper (75th percentile) quartile scoring groups.

We chose the Mann-Whitney test, for hypothesis testing, because data distribution for most features were found to be nonparametric. [26] The hypothesis test were used to determine statistical significance in any differences between the high quartile’s median and the low quartile’s median for each feature. We also evaluated the Cohen’s d effect size and the means of the upper and lower quartiles. Furthermore, to

account for multiple test comparisons, we apply a Bonferroni correction to the significance test results [27]. Differing views exist on how to calculate the Bonferroni multiplier [28], [29]. For this paper, we take a conservative approach and treat all features resulting from a single feature group as part of the same hypothesis and thus use a multiplier equal to the number of features in each group.

Additionally for each feature we evaluate the Pearson correlation coefficient [30] with the judges' score as well as the correlation between each feature (except where otherwise noted). For the p-values associated with each correlation coefficient Bonferroni correction was applied in the same manner as above.

3.3.2 Regression Analysis

The feature to score correlations and the high/low score quartile feature median comparisons provide little insight as to whether there is any interaction between features. In other words, when considering a single feature in isolation of the other features, the other features may act as confounding variables which, in worst cases, may cause the correlations to provide values opposite from their unconfounded effect. In order to directly see how features are mutually interacting with debater score

linear regression models are trained on each feature modality separately, as well as on a combined model of with all modalities. Specifically, ridge regression is used (linear regression in which the feature weights are regularized with l2 regularization) and the features and scores are standardized before fit. Different combinations of features from different modalities are reported to show how features behave differently and provide insight on how the model's predictive ability changes. For feature extraction tools that extract features for every frame, such as OpenFace and Affdex, the average over all the frames for a particular speech is computed. Cross validation (10-fold) is used to used to estimate the test and training set errors in predicting the speech scores. The models are evaluated using a mean squared error on both the train and test sets. The MSE reported in the subsequent sections are averaged across all of the folds. For each linear regression model, we report the average model weights for each of the features as well. This allows us to understand the importance of each feature in predicting the score, as well as directly see how each of the features are associated with the score in the models.

4 Results

The DBATES dataset comprises extracted multimodal feature data for N=716 speeches, along with judges scores, demographic data for speakers, and per-round speaker survey data. This data will be made publicly available at urdebate.org upon publication.

The high/low debate score quartile analysis results are shown in (Table 3) and correlation analysis is shown in (Table 4). The Mean Squared Error (MSE) for each model is summarized in (Table 5) and the feature weights of each model is shown in (Fig. 3).

TABLE 3
Debate Score High/Low Quartile Comparisons

Features	High Quartile Median	Low Quartile Median	Bonf. scaled P-value	Effect Size (Cohen's *d*)
LIWC				
Inclusion	0.015	0.018	<0.0001	0.42
Conjunction	0.024	0.025	<0.0001	0.28
Motion	0.0060	0.0069	<0.0001	0.34
Future Tense	0.0028	0.0036	0.00023	0.48
We	0.0068	0.0079	0.00091	0.31
VADER				
positive	0.11	0.12	0.020	0.30
negative	0.059	0.049	<0.0001	-0.51
compound	0.097	0.13	0.016	0.34
Affdex				
Surprise	14.8	10.1	<0.0001	0.41
Engagement	38	31	<0.0001	0.46
OpenFace				
AU01	0.37	0.33	0.0005	0.29
Praat				
F0 - Mean	220	198	<0.0001	-0.57
F0 - SD	88	79	<0.0001	-0.44
HNR	4.4	5.0	0.0006	0.39
Jitter				
local	0.030	0.027	<0.0001	-0.73
rap	0.017	0.015	<0.0001	-0.66
ppq5	0.019	0.017	<0.0001	-0.76
Shimmer				
local	0.19	0.19	0.00010	-0.46
local, dB	1.7	1.7	<0.0001	-0.50
apq3	0.090	0.088	0.028	-0.31
apq5	0.13	0.12	0.0064	-0.36

4.1 Debate Score High/Low Quartile Comparisons

Table 3 lists the features which show a statistically significant difference between medians of the debate score high and low quartiles (using a 0.05 significance level). P-values shown are from the two-tailed Mann-Whitney test, and have been scaled by their associated Bonferroni multiplier for number of features in each modality group. The Effect Size column represents the difference between the quartile means represented in estimated number of standard deviations (effect sizes of 0.2 have been characterized as small, 0.5 as medium, and 0.8 as large.)[31]

4.1.1 Text Features (LIWC)

The LIWC word categories that show a significant difference between the top quartile debaters and the bottom quartile quartile debaters are *Inclusion*, *Conjunction*, *Motion*, *Future Tense*, and *We*. As shown, the median of each of these word categories was greater in thelow scoring group compared to the high scoring group of debaters. The *Inclusion* category includes words in both the *Conjunction* and *We* categories. Thus, it is not surprising that the *Conjunction* and *We* categories have smaller effect sizes than the *Inclusion* category.

4.1.2 Text Features (VADER)

As shown in Table 3, speeches given by top debaters contained more negative sentiment and less positive sentiment compared to low scoring debaters. That being said, both groups had a lower median negative sentiment than positive sentiment during their debates. Although we omitted showing neutral sentiment in the table due to the fact that there was no significant difference in the amount of

neutral sentiment, it should be noted that neutral sentiment was much more common than both positive and negative sentiment for both groups.

4.1.3 Audio Features

Vocal audio features including pitch (mean and st. dev.), jitter (local, rap, ppq9), and shimmer (local, local in dB scale, apq3, and apq5) vary between top debaters and bottom debaters with a p-value below 0.05. Two features, Jitter (absolute) and Shimmer (apq11), are omitted from the table because they were not significantly different. The largest differences (in terms of effect size, or in other words, estimated standard deviations) were in the jitter (local, rap, ppq5) with effect sizes ranging from 0.66 to 0.76. The next largest differences are for median pitch (F0) and median pitch st. dev., both of were significantly higher in the high score quartile.

4.1.4 Visual Features

Overall the top quartile debaters showed a statistically significant higher level of the *surprise* expression feature, with a median value of 14.8 compared to just 10.1 in bottom performing debaters (d = 0.41, p-value <0.0001). Similarly, top quartile debaters displayed a higher level of *engagement* with a median level of 38 compared to 31 for bottom quartile debaters (d = 0.46, p-value <0.0001). Among the other remaining emotions they were expressed at a much more similar level between top and bottom debater score quartiles, suggesting that surprise and engagement may be important expressions for debaters to show.

As shown in Table 3, the only OpenFace facial action unit which showed a statistically significant difference between the high and low debater score quartiles was AU01 (inner brow raiser). Top performing debaters displayed this facial feature with a median value of 0.37 while bottom performing debaters had a median value of 0.33 (d = 0.29, p-value 0.0005). The remaining facial action units that OpenFace extracts did not have a statistically significant difference between the top and bottom performing groups.

4.2 Debate Score to Feature Correlations

Table 4 shows the correlations between the various features and the debate scores. Only correlations which were statistically significant at a level of 0.05 are shown. As was done with the quantile analysis, each of the features was averaged separately over each entire speech to create a single multidimensional data point per speech (as scores are assigned to speeches).

4.2.1 Text Features (LIWC)

The LIWC categories that showed a significant correlation with score are *Present Tense*, *Swear*, *You*, and *Numbers*. In addition, normalized usage count of words not in the LIWC dictionary significantly correlates with score. While none of the categories that significantly differ between top and bottom debaters (*Inclusion, Conjunction, Motion, Future Tense, We*) are significantly correlated with score, there is some connection between LIWC word category correlations and top and bottom quartile differences. For example, *Present Tense* correlates with score, while *Future Tense*, which is disjoint from *Present Tense* is used more often by lower scoring debaters. Similarly, *You* correlates with score, while *We*, which similarly is disjoint from *You* is used more often by lower scoring debaters. Thus, even though the LIWC categories that were found to be significant in the correlation analysis and high/low quartile comparisons are not identical, they show similar findings through consideration of the word categories that are disjoint (i.e. *Present Tense* vs. *Future Tense*, and *You* vs. *We*).

TABLE 4
Feature Correlations with Debate Score

Features	Correlation	P-value
LIWC		
Present Tense	0.157	0.0001
Swear	0.132	0.0017
You	0.119	0.0050
Not in LIWC dictionary	0.117	0.0053
Numbers	0.106	0.0115
VADER		
positive	-0.112	0.0159
negative	0.199	<0.0001
compound	-0.129	0.00369
Affdex		
Surprise	0.212	<0.0001
Engagement	0.172	<0.0001
OpenFace - AU01	0.145	0.0029
Praat		
F0 - Mean	0.202	<0.0001
F0 - SD	0.156	0.000704
HNR	-0.156	0.000669
Jitter (local)	0.260	<0.0001
Jitter (rap)	0.238	<0.0001
Jitter (ppq5)	0.269	<0.0001
Shimmer (local)	0.170	0.000139
Shimmer (local, dB)	0.182	<0.0001
Shimmer (apq5)	0.137	0.00493

4.2.2 Text Features (VADER)

The VADER correlation results generally align with the results demonstrated by the quartile analysis. As before, the largest correlation of the group is negative sentiment and positive sentiment and a positive compound score are both negatively related to score. Neutral sentiment, which has virtually no correlation, is not statistically significant and is consequently omitted from the table.

4.2.3 Audio Features

As shown in Table 4, the vocal aspects of pitch, jitter, and shimmer correlate positively with score. The audio features contain the largest correlations our analysis found. Three audio features, Jitter (local, absolute), Shimmer (apq3), and Shimmer (apq11), are omitted from the table because the correlation is not statistically significant.

4.2.4 Visual Features (using Spearman correlation)

From the Affdex tool we find that the *surprise* and *engagement* expression levels have a positive relationship with a debater's score, with correlation values of 0.212 and 0.172 respectively (p-value <0.0001, <0.0001). From OpenFace we discover that AU01 (Inner Brow Raiser) has a positive correlation value of 0.145 with the score of debaters (p-value 0.0029).

4.3 Regression Analysis

In this section we present the results of the multivariate linear regression models trained on combinations of the different modalities. While the previous results show individual features' associations with debate score, this section will specifically show results of how the different modalities interact with each other in association with debate score. The results first exhibit the different mean squared errors (MSE) for models using features from different modalities and then the model weights are presented.

4.3.1 Mean Squared Error

TABLE 5
MSE for Linear Regression models

Regression Models (i.e Feature Sets)	**Train Loss**	**Test Loss**
Null deviance (i.e. error when using average score as output with no input features)	7.87	8.14
Unimodal Models	7.71	7.94
Affdex	7.83	7.91
Vader	7.43	7.78
OpenFace	6.82	7.07
Praat	6.57	7.05
LIWC		
Mulimodal Model (i.e. all features)	5.87	6.61

Table 5 above shows the resulting mean squared errors from multiple different linear regression models. The "No Features" model represents the Mean Squared Error resulting from trying to predict the debate score without using any features, i.e. the error that results when just using the training set mean as the predicted output (this is also known as null deviance). The the following set of linear regression models are using all the features from one modality to predict the debate score. The model with the lowest MSE among the models using features from a single modality is the model with the LIWC features, with a train MSE of 6.57 and a test MSE of 7.05. The last model in Table 5 represents the linear regression model in which all features from all modalities are used to predict debate score. This model achieves a train loss of 5.87 and test of 6.61.

4.3.2 Model Weights

Figure 3 shows the coefficient weights for each regression model presented in the previous section. For each feature set that we used we provide a graph that shows the coefficient weights from using that feature set in isolation. We also provide a graph for each feature set showing the weights from using all the features together in a multimodal model. In general we are not considering a comparison between the magnitude of the weights from the isolated models compared to the multimodal models, but rather a comparison of the direction of the coefficient weights.

Interestingly, there are many feature weights that experience a change in direction in the multimodal model. Specifically the weight for *joy* changes from a negative association with debate score in the isolated Affdex model, to a positive association in the multimodal model. Also, the usage of "we" and its derivatives were originally a positive weight in the isolated LIWC model, but after considering all modalities the usage of we negatively impacts a debaters score.

5 Discussion

5.1 Comparison of Unimodal and Multimodal Model Disparities

Statues and paintings of Lady Justice, the personification of fairness and morality in a judge, often depict her wearing a blindfold and holding a scale. The blindfold cutting off her visual modality, represents a strive for impartiality, there to free her from visual bias as she uses the scale to accurately weigh conflicting arguments. There are many modern day examples in which one or more sensory modalities is obscured in hopes of obtaining a better judgement. Wine judges are served wine in opaque black wine glasses for some competitions in order to prevent being affected by the color of the wine, music school and orchestra candidates are given "blind" auditions behind a curtain, and even the common adage encourages us to "don't judge a book by its cover". Beyond instances of bias, input features which are known to not be relevant to the predicted output should generally not be added to a linear regression model as they will be another source of noise. However, using a unimodal model when the output really is a function of multiple modes of input can lead one to make flawed conclusions about given input variables' effects on the output. In multiple instances, analysis of the DBATES dataset shows that it is necessary to consider both verbal and nonverbal contexts in order to properly judge a speech. Specifically, the feature weight disparities between the unimodal and multimodal regression models demonstrate how considering one modality in isolation of the others leads one to make opposite conclusions about somespeech features' associations with debate success (i.e. score). Other features show little or no difference between unimodal and multimodal models.

5.1.1 LIWC We word category usage.

One of the features showing a disparity is the usage frequency of *We* category words (The *We* category words include: we, we'd, we're, we'll, us, our, ours, ourselves, lets, let's). The unimodal regression analysis utilizing only LIWC word category frequencies finds a positive weight association between *We* usage frequency and debate score , with the 3rd largest magnitude weight out of all the LIWC word categories (see Fig. 3e). Using the LIWC features alone, one may be inclined to deduce that a speech with increased *We* usage is more likely to be associated with a higher debate score. However, in the multimodal model (since the *We* weight is *negative* as shown in Fig. 3e'), increased *We* word category usage is associated with lower debate score. This demonstrates that among the full set of features across all modalities, there are interdependencies in their association with debate score. Another way of looking at this phenomenon is to consider textitWe to be the only dependent variable, while all other features are confounding factors. Because the confounding factors are not distributed evenly among the different debate scores , unimodal analysis ends up producing incorrect results. With the benefit of a rich multimodal dataset, we are more likely to discern the true association between each feature and debate score.

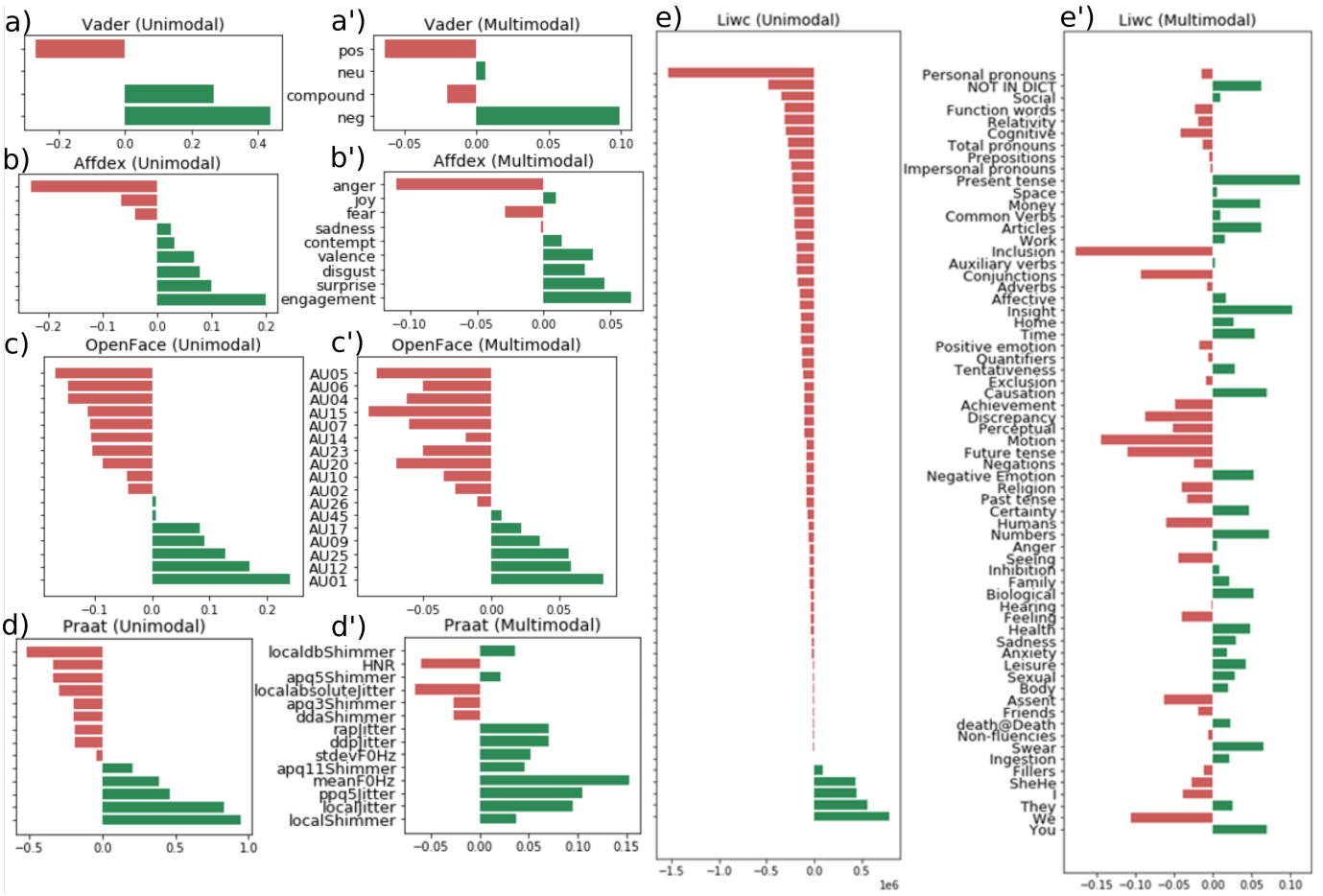


Fig. 3. **Comparison of Linear Regression model weights for unimodal models and multimodal models**. A separate unimodal model was generated using a) all VADER features only, b) all Affdex features only, c) all OpenFace features only, d) all Praat features only, and e) all LIWC features only. Immediately to the right of each unimodal model are their corresponding weights from the multimodal model (a',b',c',d',e'). Disparities in weights indicate features which are not independent of other features for score prediction.

While the models' predicted associations between debate score and *We* usage does not indicate causality, it does suggest that speakers may be able to improve their debate score by using less *We* language. This is perhaps surprising, since the characteristic of putting the group before oneself is often both culturally and socially extolled. For example, in the United States, the mantra *There is no I in TEAM* is commonly heard in school, sports, and workplace environments.

More generally, prior literature is divided on the utility of using *We*, suggesting a strong dependence upon domain and context.

For example, Simmons et. al suggest that "we" can signal a sense of group identity and have shown that in married couples, the more they used "we", the better their marriages [32].

Alternatively, Gonzales et. al found the use of first-person plural pronouns such as "we" was found to be negatively related to group cohesiveness in groups consisting of 4-6 members [33]. It may be the case that the individuals in Gonzales et. al's study were using "we" language to try to create, or give the appearance of, cohesion in an otherwise noncohesive group.

Tausczik, et al. showed that "we" and other pronouns people use often provide useful insights to where their focus is [34]. This has been shown in the realm of advertising, where positive political ads were shown to use more personal pronouns such as "we" and "I" than negative ads, which focused more on the opposition [35].

Individual debate speeches from the DBATES dataset were examined in attempt to identify characteristics of the potential harm or utility of using "we". In a debate round, "we" can be used either to refer to a debater and their partner, or can be used to refer to some larger community of people. The former usage is very common in debate. In some cases, it is used merely as a signpost to clearly demarcate what is being asserted by one's own team, in contrast to what others are asserting (e.g., "They tell you that it is worth trading some liberty for security, but *we* say that this is never a good bargain."). One can imagine a team using this signposting function of "we" too little. This could render it unclear what claims are being asserted by that team and what claims they are refuting. This could lead to a confusing speech that was unpersuasive, in which case the low frequency of "we' usage results in a lower score.

In other cases, "we" functions as a filler word that ultimately waters down what is being asserted (e.g., *We* think that the government has an obligation to protect the most vulnerable in society" as compared to just "The government

has an obligation to protect the most vulnerable in society"). For the same reason English teachers have been telling students to remove the "I think" (or worse "I feel") from their essays, it is wise for debaters to remove these uses of "we" from their speeches. For some debaters, "we" is largely just an empty filler word, and for others it also waters down the force of their assertions on top of wasting valuable time. In this case, higher frequency of "we" usage results in a lower score.

In contrast to both these cases, one can imagine other uses of "we" that are based on trying to include the audience in a broader community of like-minded thinkers. For example, "Some people may naively embrace neo-liberal capitalism as a pure meritocracy, but *we* all know that this is a terribly simplistic, cruel and ultimately racist perspective." This example sentence attempts to bring the judges into the wise in-crowd being defined by the speaker, potentially causing a higher score to be given for that particular debater using "we" in that specific case.

Clearly there are plenty of useful and harmful ways that "we" usage comes up in debate. We hypothesize that a mere frequency count of a word is insufficient to explain its effectiveness or lack thereof, and that that the context regarding whether *We* usage is good or bad can in some instances be explained by the facial expressions and audio features contemporaneous with "we" word usage.

5.1.2 Affdex Joy Expression

The level of *joy* expression's linear association with debate score, like *We* usage frequency, shows significant difference between the facial expression only unimodal model and the multimodal model. Specifically, in the unimodal model, higher levels of *joy* facial expression are associated with low overall debater score as shown by the negative value of the Joy weight in Fig. 3b. However, when we look to the multimodal model (Fig. 3b') for a more detailed understanding of how all features interact to predict debate score, we find that *joy* expression usage has a positive association with debate score.

An example of how the *We* word category usage frequency and the *joy* expression levels may simultaneously affect debate score is shown in Table 6 where two debate speeches are compared. In the first row, participant id433 from the study used "we" with 0.019 frequency (compared to the all debater mean of 0.0076) and showed an average expression of 0.004 for *joy* (compared to the all debater mean of 1.233), and achieved a score of 70 (bottom quartile of speakers). For this individual if only the LIWC feature unimodal model is considered, their above average use of "we" would have led to a higher than average performance prediction. Additionally, if only the Affdex feature unimodal model is considered, again, a higher than average score is predicted due to lower than average levels of *joy*. However, this debater received a lower than average score (70. compared to the all debater average of 76). Unlike the LIWC and Affdex unimodal models, the multimodal model predicts that participant id 433's *We* and *Joy* variable values both have a negative affect on debate score.

The second debater (participant id681) shown in Table 6 used "we" with 0.0034 frequency, showed an average expression of 8.6 for *joy*, and achieved a score of 82 (top quartile of speakers). This speaker would have been expected to perform poorly if only the LIWC or Affdex unimodal models were considered due to using "we" less often and having a very high expression of *joy* throughout their speech. However, in the multimodal model the weights for "we" frequency and *joy* undergo a change in direction, and a more accurate prediction for this debater's score is achieved.

These observations demonstrate the importance of utilizing all modalities to gain a better understanding of performance.

TABLE 6
We and *Joy* in two example debate speeches

Debater	***We* Frequency** example text	**Mean *Joy***	**Debate Score**
id433	**0.019** "**We** are telling you on your own side that is justified for **us** to take action because of this and because **we** want to represent the religious minorities and **we** want to be beneficial to them and **we** want to tell them no"	**0.0040**	**70.**
id681	**0.0034** "**We** are not talking about whether or not parents tell their kids, you are special because you are for the cute blue eyes and because you are my daughter."	**8.6**	**82**
mean all debaters	**0.0076**	**1.2**	**76**

In regards to our dataset from the NAUDC debate tournament, we argue that a rich multimodal model is essential to properly determine the appropriate frequency of "We" word category and *Joy* expressionusage in the varying contexts in which it can occur during a competitive debate climate. While the *We* and *Joy* variables showed the largest divergence between the unimodal and multimodal models, as shown in Fig. 3 there are many other variables with different weight valences. Thus, more generally than the above analysis of *We* and *Joy*, we are confident our work demonstrates the superiority of a modally comprehensive analysis in its ability to more accurately model the existing inter-dependencies contributing to the scoring of competitive debates and provides insights into some of the important variables regarding such a task.

5.2 Features showing high agreement across different analyses

While the above analysis demonstrates features with divergent interpretations between the unimodal and multimodal models, as shown in Fig. 3, most features show similarity between the unimodal and multimodal models. Some features showed particularly high agreement in their association with debate score among unimodal, multimodal, correlation, and high/low quartile analyses.

5.2.1 Textual Modality - LIWC Conjunction, Future & Present Tense

The frequency of *Conjunction* words (e.g. or, and, but, also, although, unless, however, etc.) was lower in the higher performing debaters in the quartile analysis ($p<0.0001$) as well as both the unimodal and multimodal models (with the multimodal regression model finding *Conjunction* to have the 5th most negative weight out of all 108 features). Perhaps better debaters use more concise and concrete language, and less run on sentences. This may make their arguments easier to understand.

Similarly, the lower use of the *Future tense* word category in stronger debaters may indicate increased use of real-world data rather than hypothetical projections ($p<0.0003$ in quartile analysis, and the multimodal regression model finds *Future tense* to have the 3th most negative weight out of 108 features). In a similar vein, the strongest correlation with debate score was found with the *Present tense* word category, which is disjoint from the Future tense word category. The strongest statistical significance, as well as the second strongest effect size, is seen in the frequency in the use of words in the *Inclusion* category. This is likely due to the fact that the *Inclusion* category includes words in both the *Conjunction* and *We* categories.

5.2.2 Textual Modality - VADER negative sentiment

The quartile and regression results (both multimodal and unimodal) also show that higher scoring debaters tend towards using more negative sentiment and less positive sentiment compared to lower scoring debaters ($p<0.0001$, $p<0.02$). This suggests that choosing to use more words with a negative connotation and less words with a positive connotation leads to arguments appearing more persuasive, possibly by sounding more draconian. Perhaps in the context of debate, by using a negative sentiment, a debater can frame their argument in order to rally support to the negative element that needs to be changed. Similar results have been demonstrated in other domains. For example, evoking negative emotion has been shown to be more effective during fundraising activities [36].

5.2.3 Visual Modality - Affdex and OpenFace

Through quartile, correlation, and regression analysis we also found that both surprise and engagement expressions were significantly different between the top and bottom performing debaters (all $p<0.0001$). The difference in *engagement* rather obviously suggests that being more engaged with the judges (the judges were located directly in front of the camera) helps a debater. Furthermore, emotion expressions are important to consider when modeling debate speaker score because evidence has been shown that emotion plays a large role in how speeches are perceived. Gonzalez et al. showed when a leader speaks using a great amount of emotion it may be transferred to the audience and can increase the level of support for that leader [37]. A similar pattern may be happening in these debate speeches. When debaters are displaying more engagement and surprise, the judges take notice of this and it may influence higher scoring for that debater.

Overall AU01 (inner brow raiser) was the only OpenFace feature statistically different between the top and bottom performing debaters, with top performing debaters expressing higher average levels of AU01 ($p<0.0005$). This along with the regression weights in Fig. 3 show that having more expression delivered through the eyebrows and not having a static upper face region can positively impact the debater's score. Anecdotal accounts of AU01 action have been seen as an expression of conviction, and if this is how it appears to the judges, then that might explain why it would be more prevalent in the high scorers. It is possible, given that Affdex analysis indicated showing more *surprise* is associated with the top-tier debaters, that AU1+2 would also be significant (as AU1+2 is included in *surprise*). However, our initial analysis on action unit level intensities was limited to individual action units only, as we relied on Affdex to understand more complex combinations of facial expressions.

While expression levels of *Joy* were not significantly different between top and bottom performing debaters, our multimodal model found a small positive association between *Joy* levels and debate score. Our findings from OpenFace in that there was no statistically significant difference between AU06 (cheek raiser) or AU12 (lip corner puller) between the top and bottom quartile debaters. It is surprising that there is no a stronger association between smiling and debate score, as smiling has been shown to be an important expression to show while speaking [38]. Perhaps given the combative nature of debate, smiling frequently is ill advised. It is possible that expressions of joy were associated with a decreased perception of seriousness, and thus not as likely to have a strong positive effect on debate score as [37] and [38] would predict. This may especially be true due to the serious nature of the topics debated, including religion, climate change, military interference in Syria, and childcare (see Supplementary data Table V). Overall, as shown by the significance of expressing *Engagement* and *Surprise* expressions, the importance of not remaining in a neutral face is clear.

5.2.4 Audio Modality - Praat

We found that several audio features (see tables 3 and 4) were statistically significant between high scorers and low scorers, and that many of those features were correlated with score with statistical significance (at 95% confidence). This suggests that higher scoring debaters speak differently than lower scoring debaters based on a unimodal analysis of the audio modality.

Given that judges are expected to score debaters based on the quality of their arguments, rather than the quality of their presentation, one might expect to see little to no correlation between score and other features , except where such features correlate to the quality of the argument. Consequently, the relatively high effect of audio features on score is surprising. This suggests a connection between how a speaker expresses their arguments and the quality of the arguments they suggest through a mechanism such as confidence or pacing. Alternatively, this could be evidence of another confounding factor that happens to impact both audio and score, such as gender, posture, health, or stature. Many participants declined to specify their gender and in order to preserve the most data for the multimodal analysis, self-reported gender was not included. Of the participants

that reported male and female gender, there was no statistically significant difference in mean score (F=76.41,M=76.43, std.dev.=2.80, N=580).

5.3 Nonverbal Immediacy

The concept of presenters or teachers demonstrating nonverbal immediacy [39] has been shown to correlate significantly with learning outcomes across cultures [40] as well as student motivation [41]. Specifically, nonverbal immediacy consists of using body and facial gestures, smiles, more body lean toward the class, and non-monotonic speech [42]. It appears in the multimodal analyses that *engagement* may capture some of that concept, but we see other variables capturing other elements, such as the variability of voice tone (standard deviation of fundamental frequency in Fig. 3), and maybe as well the use of AU01, which is often paired with AU02 to generate eyebrow raises – which are in fact facial gestures. Thus, this data strongly suggests the more immediate debater is scored higher.

5.4 Regression Model Accuracies

The all-feature (multimodal) regression model had a mean squared error of 6.61, compared to 8.14 for the featureless model, showing that our model is able to explain 18% of the debate score variation. Given that the features are averaged over the entire duration of the speech, it is perhaps surprising that the features were able to provide any improvement over the featureless model in predicting debate score. In order to better predict debate score, it is likely necessary to understand the deeper meaning of each of the arguments that comprise a given speech. Additionally, as table IV shows that the error was still substantial for the training loss, it is clear that a model with more complexity than a linear model is necessary to learn the training data. In future work, we are hopeful that more sophisticated models will be developed to explain this dataset more completely. However, linear models are useful in that their interpretability allows the general trend to be understood in the association of the features with the dependent variable; debate scores. From Table 5 it is apparent that as features from different modalities were added, the mean squared error loss of the linear regression decreased. This is not surprising given our rationale for advocating multimodal modeling because with each additional modality comes more information. The additional information enables the model to be more certain of its predictions and this observation provides concrete evidence of the need for analyzing debates using multimodal data. Interestingly, unimodal feature sets such as Praat and LIWC turned out to be relatively more accurate than their unimodal visual counterparts (this is likely because predicting debate score based on AU intensity-level patterns has its limitations given that judges are instructed to focus on the words of the argument rather than the speaking style). However, the train and test error for the unimodal Praat and LIWC models are still higher than that of the multimodal model with all features. Thus, there is evidence suggesting that there exists information present in the visual modality that is missing from the others.

5.5 Limitations

5.5.1 Application to the first Kennedy-Nixon Debate

In the introduction, we mentioned that in the first presidential debate from Kennedy and Nixon in 1961, the radio audience found Nixon to have won, while the television audience found Kennedy to win [1], [2]. This interpretation, however, is controversial, with alternative analysis suggesting that there were no substantial disparity between television and radio audiences [43]. As an anecdotal example of applying the DBATES multimodal model to a real world example, as well as to help demonstrate model limitations, we compare how the model predicts the Kennedy and Nixon debate speeches with and without visual data. Shown in Fig. 4 are the predicted effects that the automatically extracted emotion expression levels have on debate score. The model predicts that both Kennedy's and Nixon's *debate score* would be significantly reduced by each candidates high *anger* (and low *valence*) expression levels, with Kennedy's reduction greater (4.048 score reduction for Kennedy and 2.415 reduction for Nixon). The model also predicts that Kennedy has higher expression levels of *engagement* compared to Nixon, which would give Kennedy 0.027 score advantage over Nixon. The effects of the other visual features are in comparison minimal. In summary, the model would suggest that Nixon would benefit more from the visual modality conveyed in television. However, it is important to note that the Affdex tool predicted average *anger* expression levels of 55.9 and 33.5 for JFK and Nixon respectively. The training data involving tournament competitors displayed no where near this high level of anger, with the largest average anger level detected being 11.2. Perhaps Affdex is over-estimating the perceived anger levels due to the prominent brow furrows displayed by both Kennedy and Nixon (see Fig. 4. Because our model is multimodal and dependent upon a large number of features, its estimation of debate score becomes more diversified and less susceptible to adverse performance due to a single feature being off. The large disparity in anger levels of JFK and Nixon brings light to a limitation of the DBATES dataset in that real world data may exhibit feature input feature values outside of the range of values found in the dataset. While a collegiate population is generally diverse in many aspects, the fact that the level of *anger* expression detected in the JFK - Nixon debate speeches fell outside the DBATES range raises caution in its application to other real world scenarios.

5.5.2 Expert Judge Scoring

Although the DBATES dataset possesses the unique qualities of being the first debate dataset of its kind to include the visual modality and the dataset is expertly labeled, there is one potential limitation of these two qualities that should be mentioned. The debater score that is given by these judges may not accurately incorporate the visual modality into its assessment in a way that a lay person would. This is due to the observation that the scoring rubric the judges are given stresses matter (i.e., argumentative content) over manner (i.e., effective speaking style). It is reasonable to assume that due to the scoring guidelines, some judges would have the incentive to not look at the speaker throughout their debate and instead focus on listening to the argumentative

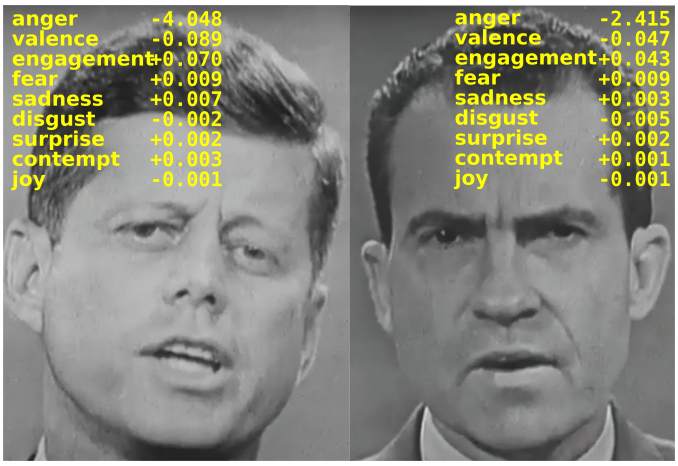


Fig. 4. **Predicted effect on debate score by averaged Affdex visual features on the Kennedy - Nixon debate speeches.**

content while taking notes. It would directly follow for that particular judge giving that specific score for that instance of speaker, that the visual modality is not reflected in their overall score. In other words, if a judge isn't looking at the speaker, then the speaker's facial expressions or other body language presumably cannot affect the judge's evaluation of that speaker. However, we argue that most judges do in fact look at the given speaker and that it is nearly impossible to completely ignore the influence of manner on debate quality even when explicitly trying to do so.

Another limitation associated with the judges' score is that due to the nature of the data, inter-annotator agreement is not available. Such data would help identify the potential judge-related subjectivity in the scores. However, the potential for subjectivity in the scores is perhaps limited since the judges were instructed to follow a scoring rubric (provided in the supplemental data), and since judges conducted a discussion immediately after each round which could help dispel an individual judge's initial unobjective predilections.

5.5.3 Feature Analysis

A limitation of the LIWC tool specifically is the low dictionary size of less than 6000 words/word roots, which represents only 6.2% of the debate transcript corpus overall. Automated facial expression analysis software has and continues to experience rapid advance in its ability to accurately detect facial landmarks [24]. However, there has also been recent criticism on the universality of facial expressions associated with truly experienced emotion. We would like to stress that the Affdex facial expression analysis is directly related to the expressed emotion, and our findings are not dependent upon an equivalence between actual and expressed emotion. Some of the auditory features (such as mean F0 pitch) may benefit from the use of gender labels. Our analysis does not make use of self reported gender since several participants did not report their gender, but we foresee future alternative analyses of the DBATES dataset investigating the use of demographic variables.

5.5.4 Data Collection Procedure

While pursuing to collect as much data as possible, inevitably, we encountered limitations during the data collection process. To protect the privacy of the speakers, we only recorded individuals that opted into being recorded. This means that for some debates we may not have all eight participants recorded. Another precaution we took to protect the privacy of the participants was agreeing to not release the raw audio or videos of the individuals. Instead, we have opted to release commonly analyzed features (such as Affdex and OpenFace AUs) and run any other analysis on the raw data by request. Indeed, not having the option to view the raw video footage could prevent researchers from drawing conclusions intuitively based on watching the videos.

In striving to record as many debaters as possible, we were required to enlist a large research staff to record debates because they happen simultaneously. This leads to potential inconsistencies in the way the videos were recorded, as well as in how easy it is for the audio to be transcribed. Consequently, the transcribers may have made errors or marked a portion "inaudible".

Finally, this paper focuses on the dataset and does only basic analysis. The machine learning models we used to understand the data are therefore simpler than one could use to try to best predict debater score (or another quality). However, by sacrificing performance scores, we acquire far more interpretability from employing said models.

5.6 Future Work

There are many exciting directions for this research moving forward. We have self-reported surveys on a debater's level of belief in the position they were arguing for. Using that information, we could work to develop a model that helps detect when a person is voicing a position that they do not actually hold. Such information could be valuable at a grand scale, such as for evaluating the speeches of foreign government officials (e.g., understanding whether a public speech is meant to inform or to propagandize). Such information can be applied to understand small scale phenomenon as well, such as detecting specific subtleties of human communication (e.g. sarcasm and insincere flattery). Machines that can better recognize the subtle nuances of human communication will be able to better serve users in a variety of contexts.

Another interesting idea is to develop a system where an individual can practice giving a speech/debating a topic. We could apply the insights learned from deploying that system to help people mediate their political disputes over social media platforms. Imagine if there were a software add-on that would assist in providing objective feedback such as identifying toxic exchanges, checking factual claims or emphasizing forgotten points. Given the current opinion wars being waged on social media, any tool that could enable a more productive digital discourse would be highly sought after on both sides of the house.

This research could be useful in quantifying a 'gut' impression. Often, people make 'gut' judgments – and while some have terrific track records making such judgments, others have terrible track records. One effort on 'gut' judges are the 'wizards' of deception detection who seem to routinely get 80% or better detecting deception [44]. They often cannot articulate why they believe someone is lying,

but yet they have shown consistent accuracy. Multimodal analysis methods may help detect those elements that make someone an expert. For example, interviews with horse race handicappers show these individuals use a six factor function when rating horses, and this cognitive skill is uncorrelated with IQ [45]. This decomposition was based on interviews and then correlated with publicly available information on each horse (closing speed, weight, weather, etc). The analysis methods discussed in this paper present a potentially big step in further decomposing how such individuals process their worlds.

Another area that would be interesting to pursue would be to have the debater speeches evaluated by third party non-experts (e.g., Amazon Mechanical Turkers). Barnes et al showed that there is a clear difference between experienced and lay judges in the realm of evaluating debate [46]. It would be interesting to compare the ratings from the non-experts with the experts and identify what features are contributing to the discrepancy of the ratings. This would have important ramifications for presidential elections specifically because most of the voters would be classified as non-experts for evaluating presidential speeches/debates.

In addition to the data planned for the initial release, we have more data on the self-reported surveys further increasing the richness of the dataset. In addition to the measures of how strongly a debater felt in support of the topic they were debating, we captured whether the debater was a novice or varsity and even had the participant rank the teams from first to last from their perspective. We also have data on how many times various teams were referenced by other teams in the forms of rebuttals and points of information. Although the video data needs to be sifted through manually and annotated, we are excited about the potential of including reference counts for predicting debater score. Looking into all these factors will be a very exciting frontier of future debate research.

6 Conclusion

Debate is a useful skill for explaining and exploring ideas for individuals and societies. Nevertheless, up to this point there has not existed a debate dataset that includes visual data. To fill that gap, this paper presents a novel, multimodal debate dataset containing over 700 unique speeches. From this dataset we are able to analyze facial expressions, affect, phonetics, text sentiment, and LIWC categories to see how those features relate to participant surveys and official judge scores.

Furthermore, this paper demonstrates how having multiple modalities allows for a greater understanding of debate scores and how to foster more accurate predictions. Naturally, adding more information leads to higher accuracy and lower error rates. In addition to that, by analyzing the various modalities both individually and simultaneously, we showed that some features, such as use of the word “we" and facial expressions of joy, change in meaning when considered in light of all modalities. This suggests that features can be misinterpreted when modalities are missing and better understood when taken in the context of modally diverse feature sets. Consequently, we believe that this dataset is a valuable resource for the continued study of debate and multimodal machine learning models and therefore encourage other researchers to use the dataset for future work.

Acknowledgment

This research was supported in part by grant W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO), and the National Science Foundation NRT-DESE #1449828. The authors would like to thank the HWS debate team for their help. The first two authors should be considered co-first authors.

References

[1] D. Gunderman, “The story of the first tv presidential debate between nixon and kennedy — ‘my god, they’ve embalmed him before he even died’,” *New York Daily News*, Sep 2016. [Online]. Available: https://www.nydailynews.com/news/politics/story-televised-debate-nixon-jfk-article-1.2803277

[2] B. Garsten, “The sorry state of american debate,” *The Wall Street Journal*, Sep 2016. [Online]. Available: https://www.wsj.com/articles/the-sorry-state-of-american-debate-1474558271

[3] J. S. Philpott, “The relative contribution to meaning of verbal and nonverbal channels of communication: A meta-analysis,” *Unpublished master’s thesis, University of Nebraska, Lincoln*, 1983.

[4] S. G. Henry, A. Fuhrel-Forbis, M. A. Rogers, and S. Eggly, “Association between nonverbal communication during clinical interactions and outcomes: a systematic review and meta-analysis,” *Patient education and counseling*, vol. 86, no. 3, pp. 297–315, 2012.

[5] A.-S. Colţa *et al.*, “The importance of non-verbal communication in business,” *Anale. Seria Ştiinţe Economice. Timişoara*, vol. 16, no. 16, pp. 776–781, 2010.

[6] C. S. Green and H. G. Klug, “Teaching critical thinking and writing through debates: An experimental evaluation,” *Teaching Sociology*, vol. 18, no. 4, pp. 462–471, 1990. [Online]. Available: http://www.jstor.org/stable/1317631

[7] R. R. Kennedy, “The power of in-class debates,” *Active Learning in Higher Education*, vol. 10, no. 3, pp. 225–236, 2009. [Online]. Available: https://doi.org/10.1177/1469787409343186

[8] D. Burek and C. Losos, “Debate: where speaking and listening come first,” *Voices from the Middle*, vol. 22, no. 1, p. 49, 2014.

[9] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, “The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data,” in *International conference on affective computing and intelligent interaction*. Springer, 2007, pp. 488–500.

[10] R. Cowie, A. Dielman, M. Mehu, M. Pantic, I. Poggi, and M. Valstar, “D8. 1: Report on available data and annotations, identification of experimental procedures,” *European Commission in the 7th Framework Programme*, 2008.

[11] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, “Canal9: A database of political debates for analysis of social interactions,” in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–4.

[12] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King, “A corpus for research on deliberation and debate.” in *LREC*, vol. 12. Istanbul, 2012, pp. 812–817.

[13] M. Orbach, Y. Bilu, A. Gera, Y. Kantor, L. Dankin, T. Lavee, L. Kotlerman, S. Mirkin, M. Jacovi, R. Aharonov *et al.*, “A dataset of general-purpose rebuttal,” *arXiv preprint arXiv:1909.00393*, 2019.

[14] F. Zhang, H. B. Hashemi, R. Hwa, and D. Litman, “A corpus of annotated revisions for studying argumentative writing,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1568–1578.

[15] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, “The semaine corpus of emotionally coloured character interactions,” in *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 2010, pp. 1079–1084.

[16] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

[17] J. Eckstein and M. Bartanen, "British parliamentary debate and the twenty-first-century student," *Communication Studies*, vol. 66, no. 4, pp. 458–473, 2015. [Online]. Available: https://doi.org/10.1080/10510974.2015.1056916

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.11)," 2020. [Online]. Available: http://www.praat.org

[19] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and emotion classification using jitter and shimmer features," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–1081.

[20] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decision Support Systems*, vol. 74, pp. 78–87, 2015.

[21] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[22] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Liwc2007: Linguistic inquiry and word count," *Austin, Texas: liwc. net*, 2007.

[23] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[24] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[25] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 2016, pp. 3723–3726.

[26] M. Neuhäuser, *Wilcoxon–Mann–Whitney Test*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1656–1658. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_615

[27] W. Rice, "The sequential bonferroni test," *Evolution*, vol. 43, pp. 223–225, 1989.

[28] R. A. Armstrong, "When to use the bonferroni correction," *Ophthalmic and Physiological Optics*, vol. 34, no. 5, pp. 502–508, 2014.

[29] R. J. Cabin and R. J. Mitchell, "To bonferroni or not to bonferroni: when and how are the questions," *Bulletin of the Ecological Society of America*, vol. 81, no. 3, pp. 246–248, 2000.

[30] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[31] J. Cohen, "Statistical power analysis," *NY: Academic*, 1988.

[32] R. A. Simmons, P. C. Gordon, and D. L. Chambless, "Pronouns in marital interaction: What do "you" and "i" say about marital health?" *Psychological science*, vol. 16, no. 12, pp. 932–936, 2005.

[33] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker, "Language style matching as a predictor of social dynamics in small groups," *Communication Research*, vol. 37, no. 1, pp. 3–19, 2010.

[34] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[35] M. A. Gunsch, S. Brownlow, S. E. Haynes, and Z. Mabe, "Differential forms linguistic content of various of political advertising," *Journal of Broadcasting & Electronic Media*, vol. 44, no. 1, pp. 27–42, 2000. [Online]. Available: https://doi.org/10.1207/s15506878jobem4401_3

[36] R. J. Fisher, M. Vandenbosch, and K. D. Antia, "An Empathy-Helping Perspective on Consumers' Responses to Fund-Raising Appeals," *Journal of Consumer Research*, vol. 35, no. 3, pp. 519–531, 02 2008. [Online]. Available: https://doi.org/10.1086/586909

[37] S. González-Bailón, R. E. Banchs, and A. Kaltenbrunner, "Emotions, public opinion, and us presidential approval rates: A 5-year analysis of online political discussions," *Human Communication Research*, vol. 38, no. 2, pp. 121–143, 2012.

[38] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception & psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[39] A. Mehrabian *et al.*, *Silent messages*. Wadsworth Belmont, CA, 1971, vol. 8, no. 152.

[40] L. L. Pogue and K. AhYun, "The effect of teacher nonverbal immediacy and credibility on student motivation and affective learning," *Communication Education*, vol. 55, no. 3, pp. 331–344, 2006.

[41] D. M. Christophel and J. Gorham, "A test-retest analysis of student motivation, teacher immediacy, and perceived sources of motivation and demotivation in college classes," *Communication education*, vol. 44, no. 4, pp. 292–306, 1995.

[42] D. Matsumoto, H. C. Hwang, and M. G. Frank, "Emotional language and political aggression," *Journal of Language and Social Psychology*, vol. 32, no. 4, pp. 452–468, 2013.

[43] W. J. Campbell, *Getting it wrong: Debunking the greatest myths in American journalism*. Univ of California Press, 2016.

[44] M. O'Sullivan and P. Ekman, "12 the wizards of deception detection," *The detection of deception in forensic contexts*, p. 269, 2004.

[45] S. J. Ceci and J. K. Liker, "A day at the races: A study of iq, expertise, and cognitive complexity." *Journal of Experimental Psychology: General*, vol. 115, no. 3, p. 255, 1986.

[46] R. E. Barnes and M. McPartlon, "Comparing experienced and lay judges," *Monash Debating Review*, vol. 13, 2015.

Taylan Sen Taylan Sen received a PhD in Computer Science from University of Rochester and JD from University at Buffalo. He is an assistant professor of computer and information sciences at Niagara University where his research focuses on computational modeling of behavior. Sen spent seven years industry experience as a software engineer and five years experience as an intellectual property attorney.

Gazi Naven Gazi Naven received a BS in Data Science and Economics and an MS in Computer Science from University of Rochester. He is currently a machine learning engineer at AI Zwei.

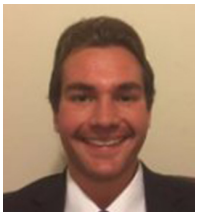
Luke Gerstner Luke Gerstner received a BS in Data Science from University of Rochester. He is currently a data scientist at Rosen.

Daryl Bagley Daryl Bagley received a BS in Mathematics and Computer Science from Harding University and an MS in Computer Science from University of Rochester. He is currently a software developer at Faithlife.

Raiyan Abdul Baten Raiyan Abdul Baten received an MS in Electrical and Computer Engineering from the University of Rochester. Raiyan is currently working towards a PhD in Electrical and Computer Engineering at the University of Rochester.

Wasifur Rahman Wasifur Rahman received an MS in Computer Science from University of Rochester. Wasifur is currently working towards a PhD in Computer Science from University of Rochester.

Md Kamrul Hasan Md Kamrul Hasan received an MS in Computer Science from University of Rochester. He is currently pursuing a PhD in Computer Science from University of Rochester.



Kurtis Haut Kurtis Haut received his B.A in Computer Science and Business from the University of Rochester in 2018. He is currently pursuing a joint PhD in Computer Science and Brain/Cognitive Science. His research interests include Computational Neuroscience, Human Cognition/Behavior, Artificial Intelligence, and Human Computer Interaction.

Abdullah Al Mamun Abdullah Al Mamun received a BA in Computer Science from University of Rochester in 2018. Afterwards he worked at the ROC HCI lab as a Research Associate to participate in more HCI oriented research. Currently he works as a Software Engineer at Kongsberg Digital.

Samiha Samrose Samiha Samrose received her B.Sc. in Computer Science and Engineering from Bangladesh University of Engineering and Technology in 2014, and M.Sc in Computer Science from University of Rochester in 2018. She worked as a Lecturer at United International University (UIU) and University of Information, Technology, and Sciences (UITS) in Bangladesh. Currently she is a PhD candidate in Computer Science at University of Rochester. Her research explores analyzing emotion and behavior patterns from group discussions and developing intelligent automated interventions.

Anne Solbu Anne Solbu has PhD from University at Buffalo. Anne Solbu studies human nonverbal behavior, such as expression of emotion, with an emphasis on contextual assessment. She researches deception and collaboration in social interaction.

R. Eric Barnes RR. Eric Barnes received a PhD in Philosophy from the University of North Carolina. He is a professor of Philosophy at Hobart and William Smith Colleges. His scholarly interests include moral and political theory, applied ethics (particularly bioethics), paradoxes, and competitive debate.

Mark G. Frank Mark Frank received his PhD in Social Psychology from Cornell University. He specializes in nonverbal communication, with a focus on understanding the complexities of facial expressions and deception in meaningful real world settings. He is currently Department Chair and Professor in the Department of Communication at University at Buffalo.

Ehsan Hoque Ehsan Hoque received his Ph.D. degree from the Massachusetts Institute of Technology in 2013. He is an associate professor of computer science with the University of Rochester where he co-leads the ROC HCI Group. Hoque's research aims to use techniques from artificial intelligence to amplify human ability. His research has been recognized with the MIT TR35 Award, NSF CAREER Award, ECASE-Army Award, among others. He is a member of the ACM, IEEE and AAAI.