Improved Visual Focus of Attention Estimation and Prosodic Features for Analyzing Group Interactions

Lingyu Zhang*
zhangl34@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York

Mallory Morgan*
morgam11@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York

Indrani Bhattacharya
Rensselaer Polytechnic Institute
Troy, New York
bhatti@rpi.edu

Michael Foley
Northeastern University
Boston, Massachusetts
foley.mic@husky.neu.edu

Jonas Braasch
Rensselaer Polytechnic Institute
Troy, New York
braasj@rpi.edu

Christoph Riedl
Northeastern University
Boston, Massachusetts
c.riedl@neu.edu

Brooke Foucault Welles
Northeastern University
Boston, Massachusetts
b.welles@northeastern.edu

Richard J. Radke
Rensselaer Polytechnic Institute
Troy, New York
rjradke@ecse.rpi.edu

ABSTRACT

Collaborative group tasks require efficient and productive verbal and non-verbal interactions among the participants. Studying such interaction patterns could help groups perform more efficiently, but the detection and measurement of human behavior is challenging since it is inherently multimodal and changes on a millisecond time frame. In this paper, we present a method to study groups performing a collaborative decision-making task using non-verbal behavioral cues. First, we present a novel algorithm to estimate the visual focus of attention (VFOA) of participants using frontal cameras. The algorithm can be used in various group settings, and performs with a state-of-the-art accuracy of 90%. Secondly, we present prosodic features for non-verbal speech analysis. These features are commonly used in speech/music classification tasks, but are rarely used in human group interaction analysis. We validate our algorithms on a multimodal dataset of 14 group meetings with 45 participants, and show that a combination of VFOA-based visual metrics and prosodic-feature-based metrics can predict emergent group leaders with 64% accuracy and dominant contributors with 86% accuracy. We also report our findings on the correlations between the non-verbal behavioral metrics with gender, emotional intelligence, and the Big 5 personality traits.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Applied computing** → *Psychology*.

KEYWORDS

Multimodal sensing; smart rooms; visual focus of attention; prosodic acoustic features; group meeting analysis

ACM Reference Format:
Lingyu Zhang, Mallory Morgan, Indrani Bhattacharya, Michael Foley, Jonas Braasch, Christoph Riedl, Brooke Foucault Welles, and Richard J. Radke. 2019. Improved Visual Focus of Attention Estimation and Prosodic Features for Analyzing Group Interactions. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353761

1 INTRODUCTION

Productive group meetings can help teams achieve goals efficiently, and consequently researchers have long studied the factors that can affect such interactions [12, 31, 35]. For example, research in group meeting analysis has shown that a single dominant participant may prevent other valuable opinions from being presented, or that a lack of engagement from team members may cause the entire group to fail [31]. Emergent leaders can boost group efficacy by helping set concrete, achievable goals [24, 54] and encouraging group members to focus on specific tasks [26]. Further, a diversity

*Both authors contributed equally to this research.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353761

of personality traits such as openness and emotional stability can also affect the decision making during a meeting [35].

To study these social interactions, researchers have considered facial behaviors such as eye gaze direction and auditory non-verbal cues such as tone of voice for estimating levels of engagement, personality type, and leadership style, among other communication patterns [7, 9]. For example, results in [40] show a correlation between group perceptions and visual cues such as the fraction of convergent gaze, mutual gaze, and shared gaze. Group participation cues such as speaking length and speaking turns are also important factors in group performance, particularly as it relates to group composition.

In this paper, we propose a new visual focus of attention (VFOA) estimation algorithm that performs with state-of-the-art accuracy on both our new dataset (90%) and the AMI Corpus (64.5%). We also propose the use of nontraditional prosodic acoustic features for non-verbal speech analysis. Finally, we use the automated VFOA analysis to extract visual metrics, and combine these with the prosodic features to analyze groups performing a collaborative decision-making task. We used the widely-adopted Lunar Survival Task experiment to study the interactions between 45 participants in 14 groups. We focus our analysis on emotional intelligence, perceived leadership, perceived contribution, and the "Big Five" personality traits. Our results show that using a small number of metrics derived from the estimated VFOA and prosodic features, we can predict group leaders and major contributors with good accuracy. Our analysis also shows some interesting correlations between these extracted metrics and emotional intelligence and the Big Five personality traits.

2 RELATED WORK

Research on dynamic group meeting interactions heavily depends on time-consuming manual annotation of pre-recorded data. For example, video frames are often manually annotated to record variables such as the location, head pose, gaze direction, and speaking duration of each participant [48]. To alleviate this difficulty, developing robust methods to automatically derive group dynamic interaction metrics has become a popular topic of research, bolstered by recent novel signal processing and computer vision techniques. To this end, many multimodal corpora of group meetings have been created and released, including the AMI [17], ATR [16], ELEA [59], ICSI [39], ISL [15], UGI [10], and NTT corpora [51–53]. These efforts have included data collected using a variety of modalities – including camera arrays, microphones, Microsoft Kinects, and wearable sensors – in order to boost automatic analysis as much as possible.

Visual focus of attention (VFOA) is one powerful non-verbal indicator used to quantify group meeting interaction and productivity. VFOA can be estimated from wearable sensors [14, 52], by using head pose as a surrogate for gaze estimation (extracted from overhead depth sensors) [11], or using frontal-facing cameras [2, 3, 9, 40, 47]. In recent research, a more accurate approach for VFOA estimation is to combine head movements with eye gaze direction data [50]. Other methods include dynamic Bayesian networks to construct switching state-space models based on continuous changes in participant location and head pose [3, 47], as well as Support Vector Machines (SVMs) to classify VFOA based on estimated head pose [9]. In our approach, we recorded individual closeup videos to estimate head pose orientation (yaw, pitch, roll) and eye gaze direction (azimuth and elevation). Frontal-facing closeup recordings and a neural network-based algorithm enable us to substantially improve the reported accuracy of visual focus of attention estimation.

Prosodic acoustic metrics serve as another significant source of non-verbal interaction cues. Common metrics such as time-domain energy and frequency-domain pitch variation have been successfully employed for emergent leadership and group performance analysis [8, 60], and are also automatically extracted in this study. However, we also introduce a new suite of largely spectral features that have been used in speech and music mood classification tasks [23, 25, 28, 33]. Our goal in introducing these frequency-based features is to capitalize on the demonstrated effect that pitch has on perceived leadership.

Sanchez-Cortez et al. [60] and Beyan et al. [7] extracted similar non-verbal visual and audio features in order to predict the emergent leader from their own group interaction datasets. Non-verbal visual metrics included VFOA, head and body position, and activity features, while non-verbal audio features included speaking activity and acoustic features such as energy and pitch variation. Sanchez-Cortez et al. found that leaders talked and interrupted more frequently, and that the energy and pitch of their voices varied more. Both groups found that no single modality achieved a higher leadership prediction score than did a combination of all modalities used [7]. Beyan et al. took a similar approach to distinguish non-leaders from leaders, and further, to predict the leadership style (autocratic or democratic) of a given emergent leader [8].

Emergent leadership research often includes the use of personality trait questionnaires such as NEO-FFI [42] or the General Leader Impression Scale (GLIS) [46] to assess the effectiveness of team performance at various cooperative tasks. The Big Five Inventory-10 (BFI-10), or simply "Big Five", is the personality trait-based system distributed as a post-task questionnaire in this study; it includes the traits agreeableness, conscientiousness, extroversion, neuroticism, and openness to experience [55]. Numerous group interaction studies observed that a certain amount of extroversion,

agreeableness, and conscientiousness are positively correlated with team success as well as individual perceived contribution [6, 22, 43]. A test of emotional intelligence (EI) is also incorporated in this study, since high EI is positively correlated with group performance, productivity, and focus [18, 20].

3 PARTICIPANT SENSING ENVIRONMENT

An 11′×28′ conference room, modified from the environment in [11], was used as our testbed to record audiovisual information during group meetings. Specifically, we used (1) four 960×720 RGB cameras for closeup recordings of each participant's face (20 fps), (2) two ceiling-mounted, downward-pointed Microsoft Kinect sensors for room layout capture (10 fps), (3) individual lapel microphones for each participant (48 kHz), and (4) a spherical 16-channel microphone hanging from the ceiling (48 kHz). We also used two reference video cameras at the two ends of the room for ground truth validation. In this paper, we only present our analysis using the frontal video cameras and the lapel microphones.

In our system, the four RGB cameras are rigidly mounted on a wooden bar for individual frontal recordings as shown in Figure 1. In contrast to the camera setups in [3, 27, 47], which used one camera on each side of the table, our camera array enables individual closeup recordings, improving the image quality and the ability to capture detailed facial actions for higher-accuracy VFOA estimation. In [17], the camera is not front-facing for each participant, causing major occlusions of participants' eyes and potentially leading to larger estimation errors. Our camera rig also has two extra cameras on each end of the bar to help in calibrating the multi-camera system.

In order to synchronize the different modalities, each meeting began with a hand-clap from a non-participant. The lapel microphones, the two Kinects, and the reference video camera recordings were synchronized together using the clap as the audio-visual cue.

4 THE LUNAR SURVIVAL TASK DATASET

The instrumented meeting room was used to record 52 individuals across 16 groups performing the Lunar Survival Task, a widely-used group discussion task that assesses how decision-making is impacted by collaboration [34]. Groups of 3–4 participants were asked to rank the utility of 15 supplies for surviving a mission on the moon. First, each participant completed the task individually, and then the group was required to reach consensus on the 15 items in, at most, 15 minutes. During this discussion, the participants were seated in specific chairs, though they were generally free to move.

Each participant was asked to complete two pre-task questionnaires before the Lunar Survival Task. The first pre-task questionnaire had 36 images of different sets of eyes, and the participants were asked to choose from four options the

Figure 1: Camera rig and example fields of view.

word that best represented the mental state of the pictured individual. This task is used to test the emotional intelligence of an individual [5], i.e., the capability to understand one's own emotions as well as the emotional state of others. The second pre-task questionnaire was a short version of the Big Five Inventory-10 (BFI-10) questionnaire [55].

After completing the group collaboration part of the Lunar Survival task, each participant was asked to complete a post-task questionnaire. In addition to questions relating to the age, gender, and ethnicity of the participants, a 5-point scale (not at all, a little, somewhat, a lot, a great deal) was used to assess the answers to the following questions:

- How well did you know each of your group members before today?
- To what extent did the following group members contribute to the discussion?
- To what extent did the following group members act as a group leader?
- To what extent did you develop rapport with the following group members?

The discussions were conducted in English, and self-reporting statistics indicate that 45% of participants were White, 35% Asian, 10% Hispanic/Latino and 10% Black. Based on self-reports, 39% of the participants were female and the ages of the participants ranged from 18 to 38 years, with an average age of 22 years and a median age of 20 years. We removed two meetings (one in which the participants moved out of

the camera fields of view, and one in which there was a microphone failure), resulting in a dataset of 45 individuals across 14 meetings for final analysis.

5 HEAD POSE AND VFOA ESTIMATION

The visual focus of attention of a participant is the dynamic visual target where he/she is looking, which can be estimated from the head pose orientation and eye gaze direction.

Feature Extraction

In our scenario, we define possible VFOA targets as either one of the other participants or "somewhere else". Based on our previous experiments with the same task [11], we found that 98% of the time, "somewhere else" corresponded to the paper in front of the participant. To automatically extract the dynamic VFOA target for every participant, we process every frame of each individual's recording using a model based on convolutional neural networks (CNNs). At each frame, we estimate three head pose angles (yaw, roll, pitch) and two eye gaze angles (azimuth, elevation).

Since the head pose angle and eye gaze direction corresponding to looking at another participant depends on one's seating position, we built a different neural network model for each seat, which can be applied to all participants seated in the same position in different meetings. Though heights and head shapes vary significantly, and there are wide differences in person-to-person behavior (e.g., one person may turn her head to look at another participant while another may mainly move only her eyes), our goal is to design a general model that encompasses the underlying common behavior patterns involved in looking at each of the specific visual attention targets.

To do this, we first applied a pre-trained deep neural network model called MT-CNN [67], which is a multi-task CNN that contains three stages of networks for progressive refinement in order to extract the face bounding box in each input image. We then feed this bounding box into a feature extraction framework based on the OpenFace model [4], which estimates the relative 3D positions of 68 facial landmarks and 56 eye landmarks. As shown in Figure 2, the line between the estimated eye pupil and the center of the eyeball is taken as the eye gaze direction. OpenFace then solves the Perspective-n-Point [37] problem using the 68 3D facial landmarks to calculate the head pose. The first stage of this process is shown in Figure 3.

Neural Network Model for VFOA Estimation

We then concatenate the extracted head pose and eye gaze angles to form a 5-dimensional feature vector. The high variation in individuals' behavior patterns made it impossible to linearly separate the VFOA classes in the feature space, and support-vector machines (SVMs) [21] had poor estimation accuracy for our task. Therefore, we built a multiple-layer perceptron (MLP) model [58] able to accurately reflect the non-linear classification boundaries inherent in the VFOA classification task. Figure 4 illustrates the architecture of the neural network model after the feature extraction step. While a typical end-to-end CNN model contains a set of convolutional layers followed by one to two fully connected layers and one step of supervision at the end of the network, our combination of a pre-trained deep CNN for intermediate feature mapping and a separate neural network for classification means that the model starts from a relatively optimal point. The pre-trained CNN serves as an intermediate supervision step and the latter classification neural network serves as the final supervision step.

Figure 2: Extracted facial and eye landmarks, with gaze directions shown as green lines.

The classification network contains N stages of fully connected layers, each of which calculates the weighted inputs, followed by a ReLU activation layer to increase the non-linearity of the network. For each fully connected layer $l \in \{1, 2, ..., N\}$, we determine the number of nodes n_l and the number of layers N by first determining the dimension shape across different layers and then the actual number of nodes in each layer. Inspired by work in [38, 45, 57], we constructed a dual-pyramid-like neural network, by first mapping the input feature vector into a higher-dimensional space and then performing a dimension contraction at each latter stage of the network to carry out the final classification. We set $n_{l+1} = n_l/2, l = 2, \ldots, N$. Since the output layer $n_N = 4$ should correspond to the number of VFOA classes, the number of hidden layers, N is then determined. By varying n_1 and examining the training and validation accuracy to check the over- or underfitting of the current model, we arrived at an optimal value of $n_1 = 1024, N = 7$.

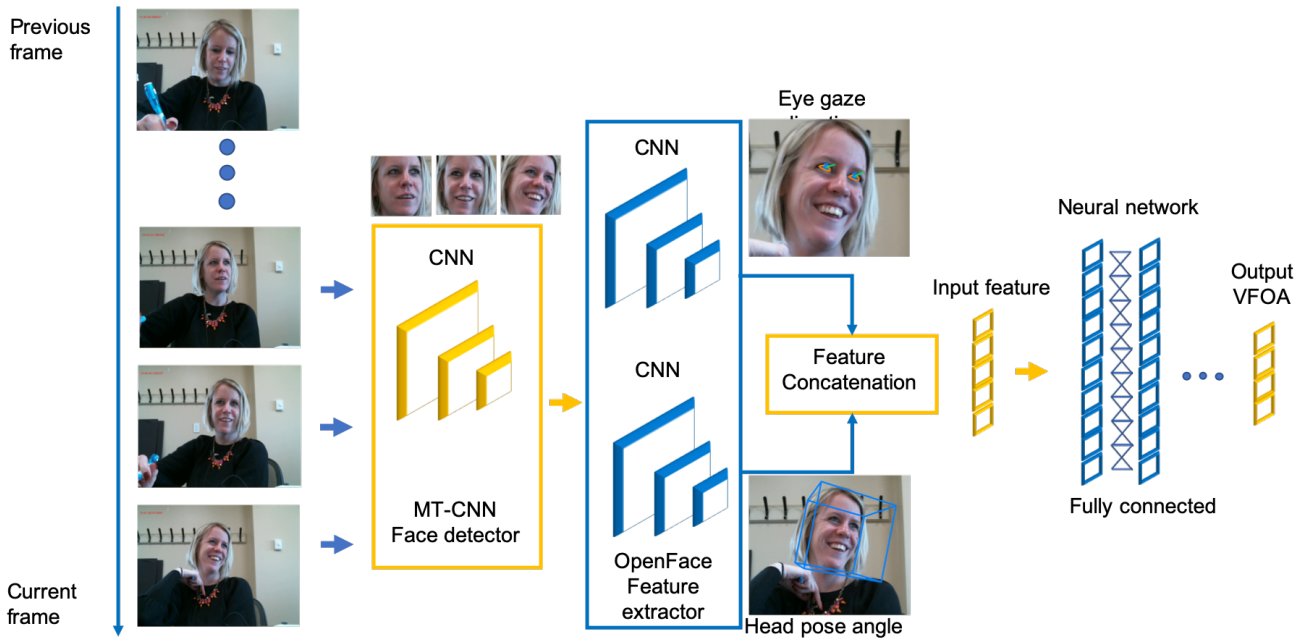

Figure 3: Overall VFOA estimation framework.

The camera angles and different illumination conditions result in different data distributions at each of the four seating positions. Therefore, during training, we tune each of the networks separately with different hyper-parameters. With varying numbers of epochs, batch sizes, initial learning rates and types of optimizer, each model was able to achieve its optimal prediction accuracy. The optimal hyper-parameter settings for each seat are shown in Table 1.

Table 1: Hyper-parameter settings for different seats.

Seat	Epochs	Batch size	Lr	Optimizer
A	45	200	0.00008	Adam
B	15	200	0.0001	RMSProp
C	150	400	0.00004	Adam
D	170	400	0.00004	Adam

Estimation Results

The VFOA labels in our study include four possible classes: {looking at the person straight ahead, looking at the person in the diagonal direction, looking at the person on the left/right, looking somewhere else}. To train and evaluate our algorithms, we manually annotated the VFOA for 52960 frames (about 7 to 10 minutes from each of 4 different meetings), roughly equally distributed across the 4 different seating positions. The details of the label distribution of the annotated data are shown in Table 2. We split the annotated data for each seating position into 2 parts: 85% for training and 15% for testing.

Table 2: VFOA label distribution in annotated data.

Seat	Ahead	Diagonal	Left/right	Other	Total
A	2824	3525	1732	6319	14400
B	2136	2495	1123	4606	10360
C	2423	2767	1274	4336	10800
D	2128	5125	2302	7845	17400

According to Table 2, for each seating position, the most common VFOA target is "somewhere else" (perhaps since the task involves referring frequently to a piece of paper on the table), and it was uncommon for a participant to look directly at the participant on the same side of the table. We computed the VFOA prediction accuracy for each seat as well as the F score [19] to evaluate the effectiveness of our model, where

$$F = 2 * \frac{precision * recall}{precision + recall}$$

Table 3 shows the results. Our method achieves an average accuracy of 90%, and a higher F score than the comparable CNN model trained on similar data recently reported in [50].

To further verify the effectiveness and generality of our model, we also evaluated it on the widely-used AMI corpus [17], which contains 14 meetings with VFOA annotation. In

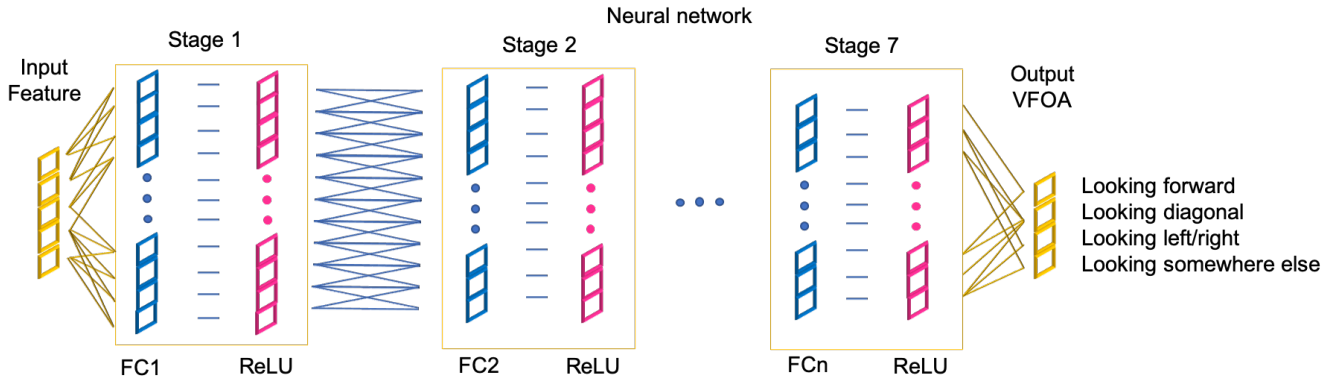

Figure 4: Details of the neural network architecture.

Table 3: VFOA prediction performance on self-collected dataset.

	Seat A		Seat B		Seat C		Seat D		Average	
	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F
CNNs [50]	-	-	-	-	-	-	-	-	-	0.799
Ours	90%	0.86	92%	0.90	88%	0.87	88%	0.85	90%	0.87

this scenario, there are 4 participants discussing a topic and the visual focus of attention targets could be one of the other participants, the table, the whiteboard on a side of the room, or somewhere else. The closeup camera positions and the relative angles between participants and cameras are totally different than our self-collected data. Nonetheless, Table 4 shows that in this different scenario, our VFOA estimation algorithm trained on the AMI Corpus outperforms the state-of-the-art results reported for DBN [3] and unsupervised incremental learning [27].

Table 4: VFOA prediction accuracy on the AMI corpus.

Model	Seat A	Seat B	Seat C	Seat D	Avg
DBN [3]	63%	56%	46%	55%	55.0%
Increm [27]	-	-	-	-	52.8%
Ours	64%	68%	58%	68%	64.5%

Visual Metrics Extracted from VFOA

We estimated the VFOA for each participant in each frame using our algorithm, and then extracted derived visual metrics for subsequent group meeting analysis. Since the attention received by a participant and the attention given by a participant during the entire meeting were observed to be good indicators of group perception [40], we extracted similar metrics, as detailed in Table 5.

6 PROSODIC ACOUSTIC FEATURES

Since group communication occurs across multiple modalities, in addition to these derived visual metrics, we extracted

Table 5: Visual metrics based on VFOA estimation.

Visual feature	Metric
Attention received by a participant	ATR
Attention given by a participant	ATG
Attention Quotient (Ratio of ATR and ATG)	ATQ
Attention Center (fraction of time a participant is looked at by all other participants)	ATC
Attention Center (2 people)	ATC2
Attention Center (2 or more people)	ATC2+
Fraction of mutual gaze	FMG

prosodic acoustic features directly from each participant's lapel microphone. We consider the commonly-used energy, fundamental frequency (or pitch), and zero-crossing rate, as well as several non-traditional spectral metrics often used in speech/music classification applications, described in further detail below. All of our acoustic features are prosodic, i.e., they convey linguistic meaning regarding emotional state, intonation, or rhythm, but are not attached to single phonetic speech segments. Energy metrics have been used numerous times in group dynamics analysis [8, 49, 60]; for example, emergent leaders are usually louder, particularly if they are male. However, leaders who deliver feedback with a "softer" voice are found to be more supportive [56]. During data collection, some signal leveling was necessary to prevent clipping or to boost quieter signals. Therefore, we compute energy variance rather than absolute energy metrics.

The zero-crossing rate is another metric calculated in the time domain and can be defined as the number of times that

a sign change occurs between two consecutive samples in the waveform. This reflects the dominant frequency of the signal, and has been used for speech/music discrimination and voice activity detection in noisy conditions [29, 61].

Fundamental frequency is closely related to the perceptual measure known as pitch, the *perceived* fundamental frequency of a sound. Since speech and music signals are not perfectly periodic, estimation techniques using autocorrelation, spectral, or cepstral-based methods are common [1]. The PEFAC (Pitch Estimation Filter with Amplitude Compression) algorithm was selected for this study because it demonstrates accurate performance under noisy conditions that may have a low signal-to-noise ratio [30, 63] and has been successfully applied in other studies analyzing group interactions [8, 49, 68]. Previous research has found that emergent leaders typically have lower pitch [28, 33].

Since music and speech are both structured sounds generated by humans for a specific purpose, they have traits that distinguish them from unstructured environmental sounds. Specifically, they both share a harmonic structure in which discrete acoustic units (e.g., phonemes or musical notes) are arranged into deliberate sequences [1]. Thus, we included a suite of features in our analysis that characterize the shape of the spectrum of a signal, which are typically used for tasks such as song, genre, or mood classification [65, 66]. These features include Mel frequency cepstrum coefficients (MFCCs), chroma, Tonnetz, and spectral centroid, rolloff, contrast, flatness, and bandwidth.

MFCCs are commonly used in speech recognition and audio content classification due to their efficient representation of speech data on a frequency scale that mirrors the human auditory system [1, 13, 44]. For this study, we extracted 40 coefficients. Chroma features project the spectrum of a signal into 12 bins corresponding to the 12 semitones on a musical octave. Outputting frequency relationships rather than frequency absolutes may reveal more information about the degree of pitch similarity not apparent in other metrics [36]. The Tonnetz feature reveals the amount of close harmonic relationships present in a signal, using Euclidian distance as a measure for harmonic change.

The spectral centroid is the predominant frequency of the signal, or the center of gravity of the spectral energy for a given frame of audio. Similarly, the spectral rolloff point is the frequency below which 85–95% of the spectral energy is concentrated (a 50% rolloff typically yields a frequency band close to the spectral centroid) [62, 64]. Both features have been successfully used for music and environmental sound classification and recognition [1]. Spectral contrast represents the relative spectral distribution of a signal, rather than the more traditional average spectral envelope, and has been known to perform better at music-type classification tasks than MFCCs [41].

Spectral flatness is a measure of how uniformly the frequency power spectrum is distributed, and in addition to the applications above, it is often used to discriminate between voiced and unvoiced speech [32, 60]. This metric was selected because emergent leaders typically modulate their voices to a greater degree [23], and increased pitch variation has been found to be positively correlated with professional success [25]. Spectral bandwidth is a metric that is useful for distinguishing between tone-like sounds or noise-like sounds because it indicates the degree to which the frequency band energies are concentrated around the spectral centroid.

Table 6 lists all of the prosodic features used in this study. Pitch metrics were computed using the VOICEBOX toolbox in MATLAB, and all other features were computed in Python. Specifically, the spectral metrics were calculated using the Librosa library, with 512 samples between successive frames. Each frame was windowed with the Hann function, with a window length of 2048. The metric values were averaged so that there was only one scalar measurement per audio file for each feature.

Table 6: Extracted prosodic non-verbal acoustic features.

Prosodic acoustic feature	Abbreviation
Energy variance	varE
Zero-crossing rate	ZCR
Pitch minimum, maximum, mean, variation	minP, maxP, meanP, varP
Mel frequency cepstral coefficient	MFCC
Chromagram	Chr
Tonnetz	Ton
Spectral centroid	cenS
Spectral rolloff	rolS
Spectral contrast	conS
Spectral flatness	fltS
Spectral bandwidth	bndS

7 CORRELATION AND REGRESSION ANALYSIS

With these multimodal metrics calculated, encompassing both VFOA and the prosodic information contained in natural speech, we investigated how they correlated with the post and pre-task target variables of perceived leadership, contribution, and emotional intelligence.

We found that emotional intelligence (EI) has positive correlations with ATR ($\rho = 0.25, p = 0.08$), ATG ($\rho = 0.26, p = 0.07$), and FMG ($\rho = 0.33, p = 0.02$), suggesting that participants with higher EI tend to look more at others, are looked at more by others, and also share more mutual gaze. Interestingly, we find that EI has a negative correlation with ATQ ($\rho = -0.39, p = 0.006$), suggesting that individuals with higher EI look at others more than they are looked at, although individually both ATR and ATG positively correlate

with EI. We also note that EI has a positive correlation with perceived contribution ($\rho = 0.33, p = 0.01$), although we do not find any significant correlations between EI and perceived leadership. Participants with higher EI are also seen to be more open ($\rho = 0.34, p = 0.01$). One significant finding was that visual metrics alone could explain 30% ($F = 2.88$, $p = 0.02$) of the variance in EI.

We also found that of the Big Five personality traits, conscientiousness, in particular, has significant positive correlations with several acoustic metrics, MFCC, Chr, cenS, fltS, bndS, meanP, ZCR, and varE (all $0.29 \leq \rho \leq 0.32$, $p < 0.05$), and negative correlations with some visual metrics, ATR, ATC2, ATC2+, and FMG (all $-0.37 \leq \rho \leq -0.32$, all $p < 0.05$). Agreeableness has positive correlations with MFCC ($\rho = 0.29, p = 0.04$), while openness has positive correlations with Chr, maxP, varE ($\rho = 0.32, 0.26, 0.26$, all $p < 0.05$).

Through multiple linear regression, one significant finding was how well both metrics could account for gender diversity, suggesting that men and women have very different patterns of looking and speaking. Specifically, the full suite of audiovisual metrics could explain 74% ($F = 2.4, p = 0.03$) of the gender diversity, with the acoustic metrics accounting for 51% ($F = 1.7, p = 0.08$) and the visual metrics accounting for 27% ($= 2.6, p = 0.03$) of the variance in gender.

Based on the post-task questionnaire rankings, we average the perceived leadership and contribution scores each participant received from the others in the group, resulting in a quantized ground truth score. We then consider the participant who received the maximum leadership/contribution scores in a particular group as the perceived leader/major contributor of that group. Each group may have more than one perceived leader and major contributor.

We used multiple linear regression to regress individual leadership and contribution scores for each participant with the automatically extracted audiovisual metrics. Since the ground truth scores are quantized, we quantized the regressed scores to the nearest bin. Using this approach, we could predict emergent group leaders with 64% accuracy, and major contributors with 86% accuracy. This result is promising and shows that substantial information about perceptions of emergent leadership and contribution can be explained by using combined VFOA and prosodic features. Taken together, these results indicate that automatically computed audiovisual features, which incorporate no analysis of informational content, may closely align with perceptions of emergent leadership and contribution in group meetings.

8 CONCLUSIONS AND FUTURE WORK

In this work, we demonstrated several aspects of group discussion analysis based on features automatically extracted from frontal videos and non-verbal acoustic data. One possible direction to improve the analysis is to increase the accuracy of the VFOA estimation with additional sensors. In particular, we intend to leverage the collected overhead range measurements recorded by the ceiling-mounted Kinect sensors to dynamically locate the precise position of each participant's head, providing information about the relative position between participants, which should improve the accuracy of VFOA estimation. We could also expand our VFOA targets to broaden the "somewhere else" class, e.g., participants presenting at a projected screen.

The overhead Kinect sensors, although not used in this work, provide color and distance maps of the room. The participants can be tracked using these distance maps, and we are developing automated algorithms that estimate the seated body posture (leaning forward vs. leaning backward) and arm pose (arms on table, crossed arms, arms touching face, arms in conversational gestures) of each participant. These postural cues can be combined with VFOA-based cues for further analysis of group discussions.

In this work, we only used the non-verbal speech signals, without considering the spoken content. We are currently working on transcribing the speech to text using automatic software. The spoken content would provide richer information about the nature of the discussion and also help explain human perceptions in groups better when combined with the non-verbal cues. We also believe that a combination of the addressee and speaker information, together with the context of the meeting, can improve the existing VFOA estimation algorithm.

Though the 16-channel, custom-built spherical microphone in the conference room was not used in this analysis, we intend to combine this microphone with the overhead Kinects to both allow participants to move about the room and remove the need for individual lapel microphones. The spherical microphone can also precisely localize speaking participants using beamforming and source segregation techniques, leading to improvements in the signal-to-noise ratio.

Finally, we would like to create metrics that use both visual and acoustic information (e.g., time spent looking at another group member while speaking loudly, or time spent moving about the room while speaking about a given topic). Taking into account both body posture and the pitch and energy of a vocal utterance, for example, might unlock richer insights into the emotional content of the meeting.

9 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIP-1631674 and by a Northeastern University Tier 1 Seed Grant. Thanks to Jon Mathews, Zhen Xu, Zengtian Deng, Arunas Tuzikas, and Gyanendra Sharma for assistance in room instrumentation.

REFERENCES

[1] F. Alías, J.C. Socoró, and X. Sevillano. 2016. A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Appl. Sci.* 6, 143 (2016).

[2] S.O. Ba and J. Odobez. 2009. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 16–33.

[3] S.O. Ba and J. Odobez. 2011. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011), 101–116.

[4] T. Baltrusaitis, A. Zadeh, Y. Lim, and L. Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[5] S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. 2001. The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines* 42, 2 (2001), 241–251.

[6] B. Barry and G.L. Stewart. 1997. Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology* 82, 1 (1997), 62–78.

[7] C. Beyan, F. Capozzi, C. Becchio, and V. Murino. 2017. Multi-task learning of social psychology assessments and nonverbal features for automatic leadership identification. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 451–455.

[8] C. Beyan, F. Capozzi, C. Becchio, and V. Murino. 2018. Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features. *IEEE Transactions on Multimedia* 20, 2 (2018), 441–456.

[9] C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 317–324.

[10] I. Bhattacharya, M. Foley, C. Ku, N. Zhang, T. Zhang, C. Mine, M. Li, H. Ji, C. Riedl, B. Foucault Welles, and R.J. Radke. 2019. The Unobtrusive Group Interaction (UGI) Corpus. In *Proceedings of the 10th ACM Multimedia Systems Conference (MMSys '19)*.

[11] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. Foucault Welles, and R.J. Radke. 2018. A Multimodal-Sensor-Enabled Room for Unobtrusive Group Meeting Analysis. In *Proceedings of the 2018 International Conference on Multimodal Interaction*. ACM, 347–355.

[12] J.H. Bradley and F.J. Hebert. 1997. The effect of personality type on team performance. *Journal of Management Development* 16, 5 (1997), 337–353.

[13] J.S. Bridle and M.D. Brown. 1974. *An Experimental Automatic Word-Recognition System*. JSU Report 1003. Joint Speech Research Unit, Ruislip, England.

[14] A. Bulling and H. Gellersen. 2010. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12.

[15] S. Burger, V. MacLaren, and H. Yu. 2002. The ISL meeting corpus: The impact of meeting type on speech style. In *INTERSPEECH*. Denver, CO.

[16] N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, and D. Douxchamps. 2006. A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow. In *Proc. Int. Conf. Lang. Resources Evaluation*. Genoa, Italy.

[17] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 28–39.

[18] J.W. Chang, T. Sy, and J.N. Change. 2012. Team Emotional Intelligence and Performance: Interactive Dynamics between Leaders and Members. *Small Group Research* 43, 1 (2012).

[19] N. Chinchor. 1992. MUC-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*. Association for Computational Linguistics, 22–29.

[20] D. Chrusciel. 2006. Considerations of emotional intelligence (EI) in dealing with change decision management. *Management Decision* 44, 5 (2006), 644–657.

[21] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

[22] P.L. Curşeu, R. Ilies, D. Virgă, L. Marticuţoiu, and F.A. Sava. 2018. Personality characteristics that are valued in teams: Not always "more is better"? *International Journal of Psychology* (2018).

[23] A. Darioly and M.S. Mast. 2014. The role of nonverbal behavior for leadership: An integrative review. In *Leader Interpersonal and Influence Skills: The Soft Skills of Leadership*, R.E. Riggio and S. Tan (Eds.). Taylor and Francis, 73–100.

[24] G. De Souza and H.J. Klein. 1995. Emergent leadership in the group goal-setting process. *Small Group Research* 26, 4 (1995), 475–496.

[25] B.M. DePaulo and H.S. Friedman. 1998. Nonverbal communication. In *Handbook of Social Psychology* (4 ed.), D. Gilbert, S. Fisker, and G. Lindzey (Eds.). McGraw Hill, Boston, MA, 3–40.

[26] V. Druskat and A.T. Pescosolido. 2006. The impact of emergent leader's emotionally competent behavior on team trust, communication, engagement, and effectiveness. *Research on Emotion in Organizations* 2 (2006), 25–55.

[27] S. Duffner and C. Garcia. 2016. Visual focus of attention estimation with unsupervised incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 12 (2016), 2264–2272.

[28] M. Frese, S. Beimel, and S. Schoenborn. 2003. Action training for charismatic leadership: Two evaluations of studies of a commercial training module on inspirational communication of a vision. *Personnel Psychology* 56, 3 (2003), 671–698.

[29] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan. 2010. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In *11th Annual Conference of the International Speech (InterSpeech)*. Makuhari, Japan, 3118–3121.

[30] S. Gonzalez and M. Brookes. 2014. PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 2 (Feb. 2014), 518–530.

[31] C. Gorse, I. McKinney, A. Shepherd, and P. Whitehead. 2006. Meetings: Factors that affect group interaction and performance. *Proceedings of the Association of Researchers in Construction Management* (2006), 4–6.

[32] J. Gray and J. Markel. 1974. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Trans. Acoust., Speech, Signal Process.* 22, 3 (1974), 207–217.

[33] J.A. Hall, E.J. Coats, and L.S. LeBeau. 2005. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin* 131, 6 (2005), 898–924.

[34] J. Hall and W.H. Watson. 1970. The effects of a normative intervention on group decision-making performance. *Human Relations* 23, 4 (1970), 299–317.

[35] M. Harris Bond and I. Wing-Chun Ng. 2004. The depth of a group's personality resources: Impacts on group process and group performance. *Asian Journal of Social Psychology* 7, 3 (2004), 285–300.

[36] C. Harte, M. Sandler, and M. Gasser. 2006. Detecting Harmonic Change in Musical Audio. In *1st ACM Workshop on Audio and Music Computing Multimedia*. ACM, Santa Barbara, CA, 21–26.

[37] J.A Hesch and S.I Roumeliotis. 2011. A direct least-squares (DLS) method for PnP. In *2011 International Conference on Computer Vision*. IEEE, 383–390.

[38] V. Iglovikov and A. Shvets. 2018. Ternausnet: U-net with VGG11 encoder pre-trained on Imagenet for image segmentation. *arXiv preprint arXiv:1801.05746* (2018).

[39] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, et al. 2003. The ICSI meeting corpus. In *Int. Conf. Acoust., Speech, and Signal Process.*

[40] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. 2012. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ACM, 433–440.

[41] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai. 2002. Music type classification by spectral contrast feature. In *International Conference on Multimedia and Expo*. 113–116.

[42] O.P. John and S. Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook Personality: Theory and Research* (2 ed.), L.A. Pervin and O.P. John (Eds.). McGraw Hill, Boston, MA, 102–138.

[43] S.L. Kichuk and W.H. Wiesner. 1997. The big five personality factors and team performance: implications for selecting successful product design teams. *Journal of Engineering and Technology Management* 14, 3-4 (1997), 195–221.

[44] S. Liang and X. Fan. 2014. Audio Content Classification Method Research Based on Two-step Strategy. *Int. J. Adv. Comput. Sci. Appl.* 5 (2014), 57–62.

[45] J. Long, E. Shelhamer, and T. Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[46] R.G Lord, R.J Foti, and C.L De Vader. 1984. A test of leadership categorization theory: Internal structure, information processing, and leadership. *Organizational Behavior and Human Performance* 34, 3 (1984), 343–378.

[47] B. Massé, S. Ba, and R. Horaud. 2018. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 11 (2018), 2711–2724.

[48] S. Mathur, M.S. Poole, F. Pena-Mora, M. Hasegawa-Johnson, and N. Contractor. 2012. Detecting interaction links in a collaborating group using manually annotated data. *Social Networks* 34, 4 (2012), 515–526.

[49] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. 2005. Automatic analysis of multimodal group actions in meeting. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 3 (March 2005), 305–317.

[50] K. Otsuka, K. Kasuga, and M. Köhler. 2018. Estimating Visual Focus of Attention in Multiparty Meetings using Deep Convolutional Neural Networks. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 191–199.

[51] K. Otsuka, H. Sawada, and J. Yamato. 2007. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: Who responds to whom, when, and how? From gaze, head gestures, and utterances. In *Proc. Int. Conf. Multimodal Interfaces*. ACM, Aichi, Japan.

[52] K. Otsuka, Y. Takemae, and J. Yamato. 2005. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th International Conference on Multimodal Interfaces*. ACM, 191–198.

[53] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. 2006. Conversation scene analysis with dynamic Bayesian Network based on visual head tracking. In *Proc. Int. Conf. Multimedia and Expo*. IEEE, Toronto, ON, Canada.

[54] A.T. Pescosolido. 2001. Informal leaders and the development of group efficacy. *Small Group Research* 32, 1 (2001), 74–93.

[55] B. Rammstedt and O.P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212.

[56] M. Remland. 1981. Developing leadership skills in nonverbal communication: A situational perspective. *Journal of Business Communication* 18, 3 (1981), 17–29.

[57] O. Ronneberger, P. Fischer, and T. Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 234–241.

[58] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. 1985. *Learning internal representations by error propagation*. Technical Report. California Univ San Diego La Jolla Inst for Cognitive Science.

[59] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez. 2011. An audio visual corpus for emergent leader analysis. In *Workshop Multimodal Corpora Mach. Learning: Taking Stock and Road Mapping the Future*. Alicante, Spain.

[60] D. Sanchez-Cortes, O. Aran, and M. Schmid Mast D. Gatica-Perez. 2012. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[61] J. Saunders. 1996. Real-time discrimination of broadcast speech/music. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 993–996.

[62] E. Scheirer and M. Slaney. 1997. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1331–1334.

[63] L. Sukhostat and Y. Imamverdiyev. 2015. A Comparative Analysis of Pitch Detection Methods Under the Influence of Different Noise Conditions. *Journal of Voice* 29, 4 (July 2015), 410–417.

[64] C. Thoman. 2009. *Model-Based Classification of Speech Audio*. Master's thesis. Florida Atlantic University, Florida, USA.

[65] A.L.C. Wang. 2003. An industrial-strength audio search algorithm. In *Proceedings of the 4th International Society for Music Information Retrieval Conference*. Baltimore, MD, 7–13.

[66] F. Wang, X. Wang, B. Shao, T. Li, and M. Ogihara. 2009. Tag Integrated Multi-Label Music Style Classification with Hypergraph. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*. Kobe, Japan, 363–368.

[67] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[68] T. Zhang and J.C.C. Kuo. 1999. Heuristic approach for generic audio data segmentation and annotation. In *Proceedings of the 7th ACM International Conference on Multimedia*. 67–76.