JASA INVITED-PAPER

Classifying the emotional speech content of participants in group meetings using convolutional long short-term memory network[a)]

Mallory M. Morgan,[1,b)] Indrani Bhattacharya,[2,c)] Richard J. Radke,[2,d)] and Jonas Braasch[1,e)]
[1]*School of Architecture, Rensselaer Polytechnic Institute, Troy, New York 12180, USA*
[2]*Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA*

ABSTRACT:
Emotion is a central component of verbal communication between humans. Due to advances in machine learning and the development of affective computing, automatic emotion recognition is increasingly possible and sought after. To examine the connection between emotional speech and significant group dynamics perceptions, such as leadership and contribution, a new dataset (14 group meetings, 45 participants) is collected for analyzing collaborative group work based on the lunar survival task. To establish a training database, each participant's audio is manually annotated both categorically and along a three-dimensional scale with axes of activation, dominance, and valence and then converted to spectrograms. The performance of several neural network architectures for predicting speech emotion are compared for two tasks: categorical emotion classification and 3D emotion regression using multitask learning. Pretraining each neural network architecture on the well-known IEMOCAP (Interactive Emotional Dyadic Motion Capture) corpus improves the performance on this new group dynamics dataset. For both tasks, the two-dimensional convolutional long short-term memory network achieves the highest overall performance. By regressing the annotated emotions against post-task questionnaire variables for each participant, it is shown that the emotional speech content of a meeting can predict 71% of perceived group leaders and 86% of major contributors.
 https://doi.org/10.1121/10.0003433

(Received 28 July 2020; revised 5 January 2021; accepted 9 January 2021; published online 4 February 2021)

[Editor: Bozena Kostek] Pages: 885–894

I. INTRODUCTION

Collaborative group meetings are a ubiquitous aspect of modern working life, and a large body of research involves the automatic analysis of such meetings to understand the dynamics of productivity, leadership, and rapport. For example, emergent leaders have a significant impact on group efficacy by helping set concrete, achievable goals and encouraging group members to focus on specific tasks.[1,2] However, emergent leadership cannot be measured directly without relying on time-consuming manual annotation from outside observers or unreliable self-reporting statistics, therefore, emergent leadership and other group meeting metrics are instead frequently predicted from derived features such as facial behaviors (e.g., face, eye gaze direction) and auditory nonverbal cues (e.g., energy, pitch).[3–8]

This paper focuses on deep-learning-facilitated classification of speech emotion with the purpose of using the result as a group meeting metric with more direct meaning than the aforementioned low-level metrics that are commonly extracted in the group dynamics literature. Specifically, while metrics, such as eye gaze direction, may have some degree of correlation with the perceived group leader (e.g., the perceived group leader is more likely to look in a certain direction), it is not immediately clear that this behavior is *causing* them to be perceived as the group leader. Our hope is that with a more intuitive metric, such as the fraction of happy/sad/excited utterances for each participant, the causal link to important group meeting signifiers, in this case, the identity of the emergent group leader/contributor, may be clearer. However, attempting to use a metric such as speech emotion comes at a cost; it is difficult to extract without significant manual annotation efforts. For this reason, we work toward predicting this information automatically and reliably using deep learning techniques (see Fig. 1).

Two paradigms for characterizing "emotion space" include a discrete/categorical emotion model and a dimensional emotion model. Discrete emotion theory was first developed by Ekman and Oster in 1979 and is based on the premise that there are six culturally universal emotions: anger, disgust, fear, happiness, sorrow, and surprise.[9] However, this categorization approach can be ill-equipped to handle the nuanced emotional shifts common during many communication events. A dimensional emotional model, by contrast, typically characterizes emotions on a set of continuous (though often discretized for convenience) axes, typically activation, dominance, and valence.[10] Activation is defined

[a)]This paper is part of a special issue on Machine Learning in Acoustics.
[b)]Electronic mail: morgam11@rpi.edu, ORCID: 0000-0003-1690-6491.
[c)]ORCID: 0000-0003-1557-6961.
[d)]ORCID: 0000-0001-5064-7775.
[e)]ORCID: 0000-0001-8619-4819.

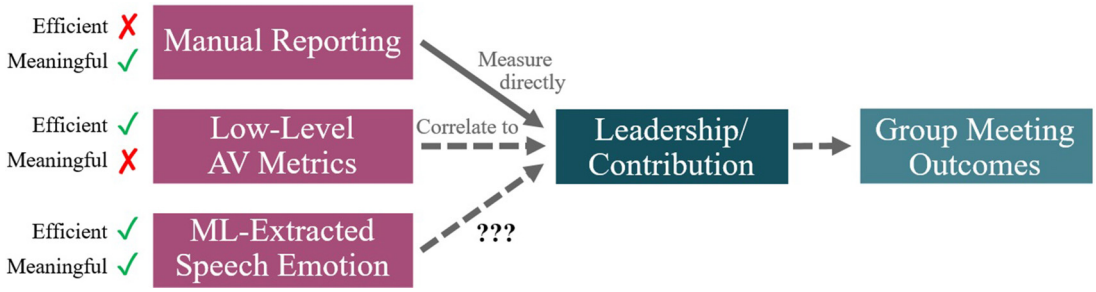

FIG. 1. (Color online) Diagrammatic representation of the application of speech emotion to the study of group dynamics. In this paper, the ability of different deep learning techniques to extract the emotional speech content from group meeting participants is explored. The degree of correlation between that emotional speech content and the perception of the group leader and contributor is calculated, motivated by the fact that such leaders have a demonstrable effect on the efficiency of group meetings (Refs. 1 and 2).

as the energy or arousal level of an emotion, ranging from apathy to excitement. Dominance refers to the level of emotional control and ranges from weak to strong. Valence describes the positivity of the emotion and ranges from unpleasant to pleasant.

Research in this area relies on annotated speech emotion databases that consist of some combination of acted/simulated speech segments, elicited/induced speech segments, and natural/spontaneous speech segments.[11] Acted speech databases are usually recorded by professional actors under ideal acoustic conditions and include the Emo-DB (Berlin Database of Emotional Speech)[12] and parts of the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database.[13] While convenient to assemble, such an approach can lead to exaggerated expressions lacking realistic nuance. Elicited speech databases offer greater authenticity because they are comprised of simulated emotional situations where actors are free to improvise their reactions. Parts of the IEMOCAP database contain elicited speech. Both elicited and acted databases often involve a very small number of participants speaking in isolation, which can limit the generalizability needed for robust machine learning classification.

To capture more realistic behavior, natural speech databases, such as the VAM (Vera-am-Mittag) dataset[14] and the SAFE (Situation Analysis in a Fictional and Emotional) corpus,[15] source speech segments from television and talk shows, films, radio programs, and call-center recordings. To avoid the legal and ethical considerations associated with assembling databases that are almost exclusively drawn from the entertainment sphere, the RECOLA (Remote Collaborative and Affective Interactions) database consists of the online interactions between 46 participants as they perform a collaborative task.[16] However, to the authors' knowledge, only the first 5 minutes of data from 23 speakers have been annotated and made public.

To construct these databases, time-consuming and cumbersome manual annotations are required, motivating the use of signal processing and machine learning techniques to predict the correct annotations automatically. Deep learning methods are increasingly preferred over traditional machine learning methods because manual feature selection, resulting in accurate network classification performance, is laborious and rarely straightforward.[17] By contrast, deep learning offers an end-to-end approach whereby high-level feature selection and network training occur in tandem through an automatic, iterative process. Trigeorgis *et al.* first proposed using such an approach for speech emotion classification using a one-dimensional (1D) convolutional neural network–long short-term memory (CNN-LSTM) network to perform regression tasks on the RECOLA database in 2016.[18]

Etienne *et al.* and Zhao *et al.* have also achieved four- and six-class classification success on the IEMOCAP database with two-dimensional (2D) CNN-LSTM networks classifying spectrogram images.[19] 2D convolutional neural networks (CNNs) have also been successfully used for such tasks[20,21] as have recurrent 1D LSTM networks (classifying audio waveforms).[22,23] Combining these two approaches, Yang and Hirschberg had success classifying emotion in the sustained emotionally colored machine-human interaction using nonverbal expression (SEMAINE) and RECOLA databases by creating a "fusion" network in which the audio waveforms were processed using a 1D CNN, the spectrogram representations were processed using a 2D CNN, and then both networks were joined with multiple LSTM layers.[24]

In this paper, a version of each of these four related neural network approaches—(i) 2D CNN, (ii) 1D LSTM, (iii) 2D CNN-LSTM, and (iv) 1D/2D fusion networks—were implemented to compare their performances on a new group dynamics dataset. For each network, both the categorical and three-dimensional (3D) emotional speech content of participant utterances were predicted. For the latter task, all three affective dimensions (activation, dominance, and valence) were jointly regressed using a multitask learning framework.[18,24–27] For both tasks, transfer learning was conducted on the IEMOCAP corpus to increase the generalizability of the model, significantly increasing the final performance and lending further evidence to the utility of cross-corpus training between exaggerated and natural speech emotion corpora.[28,29]

To create the group dynamics dataset, the widely adopted lunar survival task was conducted to study the naturally unfolding dynamics of small, collaborative group meetings. Pre- and post-task questionnaires were administered to assess the perceived group leadership and contributions of the participants among other attributes. After deep learning techniques, because each participant was linked to

https://doi.org/10.1121/10.0003433

this quantitative survey data, the degree to which the emotional speech content of a meeting is correlated to and predictive of emergent group leadership and contribution was explored. Ultimately, the proportion of emotional utterances attributed to each participant was able to correctly predict 71% of perceived group leaders and 86% of major contributors, drawing a promising link between speech emotion and group meeting dynamics.

II. THE LUNAR SURVIVAL TASK DATASET

The dataset introduced in this study was collected in a university conference room ($11' \times 28'$), equipped with lapel microphones (48 kHz), frontal-facing cameras, Microsoft Kinect sensors (Redmond, WA), and a ceiling-suspended spherical 16-channel microphone array (although in this paper, only the lapel microphone data are discussed). In the meeting room, groups of three or four participants were asked to complete the lunar survival task both individually and in a group. This task is widely used in group discussion research for assessing how collaboration impacts decision-making and consists of participants ranking the utility of 15 supplies for surviving a mission on the moon.[30] After coming up with an individual ranking, participants were then given 15 min to reach a consensus on the items. Although participants were generally free to move, they were seated in specific chairs. Altogether, the instrumented meeting room was used to record 45 individuals across 14 group meetings.

A. Pre-task questionnaires

Before the task began, each participant completed two pre-task questionnaires. The first questionnaire, called "Reading the Mind in the Eyes," had participants choose which of four emotions best represented the mental state of a pictured individual. The test consists of 36 images of different sets of eyes and is commonly used to test emotional intelligence (EI) or the ability of an individual to understand their own emotions as well as the emotions of others.[31] High EI has previously been positively correlated with group productivity, focus, and overall performance.[32,33]

The other pre-task questionnaire was a short version of the Big Five Inventory-10 (BFI-10), often used to assess team performance in emergent leadership research. The questionnaire is designed to rank participants on the traits agreeableness, conscientiousness, extroversion, neuroticism, and openness to experience.[34] Numerous group interaction studies observed that a certain amount of extroversion, agreeableness, and conscientiousness are positively correlated with team success as is individual perceived contribution.[35–37]

B. Post-task questionnaires

After completing the full lunar survival task, each participant was asked to complete post-task questionnaires. In addition to questions relating to the age, gender, and ethnicity of the participants, a five-point scale was used to gather the answers to the following four questions. On the five-point scale, 1 = not at all, 2 = a little, 3 = somewhat, 4 = a lot, and 5 = a great deal. The questions are as follows:

(1) How well did you know each of your group members before today?
(2) To what extent did the following group members contribute to the discussion?[38]
(3) To what extent did the following group members act as a group leader?
(4) To what extent did you develop rapport with the following group members?

In particular, the second and third questions are used to derive the perceived emergent leadership and contribution metrics for each participant and are used to conduct the analysis documented in Sec. V.

While much existing work assesses emergent leadership using manual annotation conducted by the researchers themselves[3,4] or personality trait-based questionnaires,[34,39,40] our work here concerns only *perceptions* of leadership. While comparatively easy to quantify, this perception is subjective; it will likely vary from individual to individual and is, thus, left open to interpretation. Crucially, however, the perception of leadership, independent of any outside assessment, particularly the convergence of that perception, has documented bearings on team performance.[41] Therefore, our leadership and contribution metrics are based solely on these participant responses and are not dependent on any outside annotations.

Discussions were conducted in English, and self-reporting statistics indicate that 45% of the participants were White, 35% were Asian, 10% were Hispanic/Latino, and 10% were Black. Additionally, 39% of the participants were female, and the ages of the participants ranged from 18 to 38 years old with an average age of 22 years old and a median age of 20 years old.

C. Utterance segmentation

Before annotation, each participant's speech was automatically separated into discrete utterances by algorithmically segmenting the signal during periods when the signal energy and spectral centroid dropped below a certain threshold (a pause in speech). If a segment of speech was within 50 ms of the prior speech segment, it was considered to be part of that prior utterance. If not, a new utterance was instantiated. After this segmentation process, the average utterance was 7.0 s long, whereas the shortest utterance was 1.3 s and the longest utterance was 30 s.

Given the nature of the discussion, there were very few instances when more than one participant was talking. Each participant's audio was individually recorded with lapel microphones of sufficient directionality, for example, to render the speech of participants 2–4 virtually inaudible to the microphone of participant 1. Therefore, the audio of two participants engaged in cross talk could be annotated separately.

D. Dataset annotation

The dataset was manually annotated by two individuals using a graphical user interface (GUI), illustrated in Fig. 2, in 10 min sessions. The audio of each utterance was played in a random order and then annotated in two ways by each annotator: first, categorically, and then on the 3D scale. For the category-based annotations, slightly different emotion categories were selected than those in the "basic six" proposed by Ekman *et al.*, thought to be more appropriate for the tenor of the discussions that took place. These emotions are anger, excitement, fear, frustration, happiness, neutral, and sorrow (but with the option to choose "other").

For the 3D approach, the annotators ranked each emotion on a scale of 1–5 along axes of activation (arousal-nonarousal), dominance (dominance-submissiveness), and valence (pleasantness-unpleasantness). Commonly, these three axes are represented pictorially using discrete "manikins" for each scale (see Fig. 2), originally created to quickly and easily track personal response to an affective stimulus.[42] The method has been shown to have a low standard deviation and high agreement between evaluators.[43] The method is also useful because it avoids inevitable differences between each evaluator's understanding of purely linguistic emotion labels, making the approach intuitive and efficient.

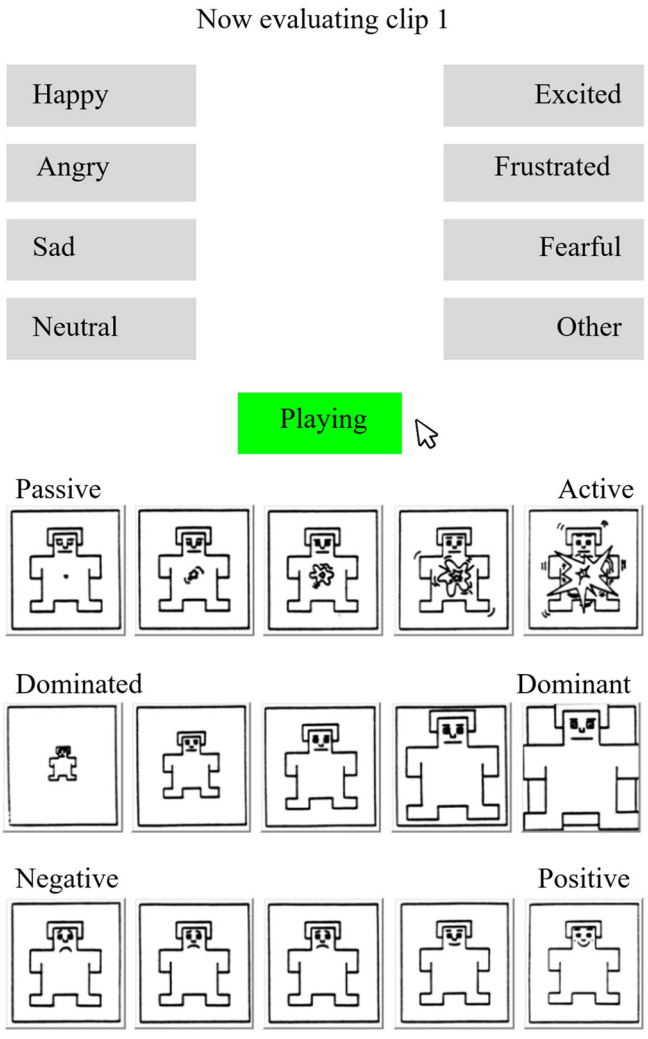

FIG. 2. (Color online) Pictural representation of the graphical user interface (GUI). (Top) GUI for the categorical audio segment annotation. (Bottom) GUI for the 3D activation-dominance-valence annotation using the non-verbal pictural assessment technique, Self-Assessment Manikin (SAM; Ref. 42).

In the scheme in Fig. 2, each of the five illustrations per dimension convey a progression from one end of the spectrum (1) to the other (5). In this work, "1" will be used to indicate the most passive, submissive, and negative (lowest activation, dominance, and valence) utterances while "5" will be used to denote the utterances with the highest activation, dominance, and valence. Therefore, an utterance with an annotation (3,5,1) indicates "neutral" activation (neither calm nor excited), high dominance, and very low valence (very negative).

Despite all of the precautions, these annotations are subjective assessments by their very nature; to decrease ambiguity, only prototypical speech segments, i.e., segments for which both annotators agreed on the annotation, were considered. A negligible number of utterances were classified in the "anger," "fear," and other categories and not included in the study. The evaluators were university students who were fluent English speakers. The top rows of Tables I and II show the percentages of the total database that each emotion category comprises for both the categorical annotations and affective dimensional annotations.

III. COMPUTATIONAL METHODS

A. Data pre-processing

To prepare the database for neural network classification, each audio segment was first downsampled to 16 kHz and segmented into 8-s-long clips. Segments shorter than 8 s were padded to 8 s (with clips shorter than 1 s not considered for network training). At this stage, each audio clip can be represented as a 128 000-sample vector. For some of the neural network architectures, the clips were then turned into log-mel spectrograms, representing the short-term power spectrum of the audio. A fast Fourier transform (FFT) window length of 2048 and a hop length of 512 were chosen, resulting in spectrograms with 128 mel frequency bins and 251 temporal frames.

B. Data augmentation

Before network classification of the lunar task dataset could begin, the severe class imbalance and relatively small amount of data needed to be taken into account. In addition

TABLE I. Percentage (%) of 8 s categorical annotations for the lunar task database and the IEMOCAP database.

	Anger	Excitement	Frustration	Happiness	Neutral	Sorrow
Lunar task	0.0	3.5	4.0	12.5	78.0	2.0
IEMOCAP	15.0	14.0	24.5	9.5	22.5	14.5

https://doi.org/10.1121/10.0003433

TABLE II. Percentage (%) of 8 s 3D annotations—activation, dominance, valence—for the lunar task database and IEMOCAP database.

		1	2	3	4	5
Lunar task	Activation	8.0	26.0	49.0	15.0	2.0
	Dominance	7.0	20.0	46.0	28.0	2.0
	Valence	1.0	7.0	78.0	12.0	2.0
IEMOCAP	Activation	1.0	28.0	46.0	23.0	2.0
	Dominance	1.0	16.0	45.0	32.0	6.0
	Valence	3.0	33.0	40.0	22.0	2.0

to oversampling the underrepresented classes with a weighted loss function, data augmentation techniques were introduced to increase the generalizability of the model. Specifically, slight random horizontal shifts and flips of the spectrogram images were introduced using the Keras ImageDataGenerator class.[44] The test images were also augmented using test time augmentation (TTA).[44] With TTA, a prediction is output for each of ten augmented versions of the same image. Then, these predictions are averaged together and taken as the final estimate. By averaging the predictions on randomly modified images, the errors are also averaged.

C. Deep learning architectures

1. CNN

For the first deep learning architecture, a four-layer CNN was constructed, taking the 2D spectrogram images as input. The first 2 layers had 64 filters, whereas the latter 2 layers had 128 filters. A kernel of three was used for all four layers. A batch normalization layer and a max pooling layer follow each convolutional layer. For the first convolution step, both the kernel and stride of the max pooling layer were two while for all other layers, a kernel and stride size of four was used. Stochastic gradient descent (SGD) was used as an optimizer for the categorical task, and Adam was used for the dimensional task.

In the case of the categorical task, for all of the networks, a fully connected layer was added before the softmax function was employed to output the final emotion prediction. In the case of the 3D affective dimension regression task, there are instead three nonconsecutive fully connected layers, each with a linear activation function used to output a predicted value for activation, dominance, and valence. Instead of a categorical accuracy metric, the mean-squared error and the concordance correlation coefficient (CCC) were used to evaluate the network performance during training. In both cases, 15 training epochs occurred.

2. LSTM network

For the second network, 2 LSTM layers, consisting of 512 and 256 units, were added in sequence, followed by a dropout layer. For temporal processing, the unprocessed audio waveform input vector is first reshaped into an 80×1600 matrix before being fed into the network. Because the audio data were downsampled to 16 kHz, this reshaping represents dividing the waveform into eighty 100 ms time steps, each containing 1600 features. For the categorical task, SGD was used as the optimizer, whereas for the dimensional task, Adam was used. In both cases, 15 training epochs occurred.

3. CNN-LSTM network

For the third network, the CNN front-end is identical to that described previously. Following this, a LSTM layer with 256 units was appended. The optimizer Adam was used for both the classification and regression tasks. In both cases, 15 training epochs occurred.

4. Fusion network

For the 1D segment of the fusion network, the audio waveform input vector is passed through two 1D convolutional layers, the first layer with 64 filters and the second with 128 filters. The first layer had a kernel size of eight while the second layer had a kernel size of four. A pooling layer with kernels of 10 and 20 was added after each convolution.

For the 2D segment of the fusion network, the spectrogram image is again passed through two 2D convolutional layers with 64 and 128 filters each. The first had a kernel size of eight while the second had a kernel size of four. A max pooling layer with a kernel and stride of two was added after each convolution step.

The resulting output vectors from both segments of the network were then concatenated and fed into two LSTM layers, each with 128 units. A dropout layer follows before categorical prediction or regression, depending on the task. For the classification task, 15 training epochs are used while 25 epochs are used for the regression task. In both cases, Adam is used for the optimizer.

A summary of all four neural network architectures, as well as the two different output structures for each of the two tasks, can be found in Fig. 3.

D. Transfer learning

With transfer learning, information learned from one classification task can be leveraged for a similar task, making the resulting neural network less prone to overfitting. Since the lunar task dataset is small compared to other speech emotion datasets, we predict that pretraining these neural networks on a larger speech emotion database may result in higher classification accuracies. The IEMOCAP corpus was selected for this task, given the similarity in database construction and annotation format.

The IEMOCAP corpus was created to study how emotive human communication occurs. Beyond the audio data, the corpus also contains video footage and motion capture markers of conversing pairs of actors (six male, six female). As mentioned above, the emotion is both simulated and elicited with actors participating in both scripted and improvised emotional scenes. In all cases, a minimum of three

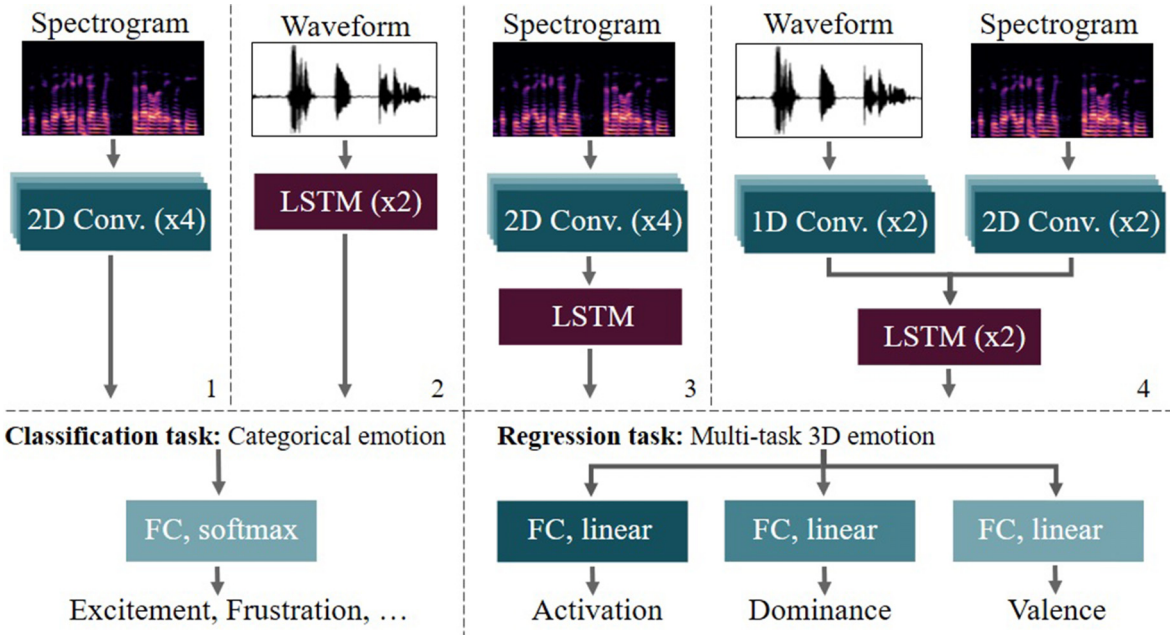

FIG. 3. (Color online) (Top) Details of each of the four neural network architectures. (1) The CNN, (2) LSTM network, (3) CNN-LSTM network, (4) fusion network. (Bottom) Two different output structures, one for the categorical emotion classification task and the other for the dimensional emotion regression task.

evaluators labelled the data. The IEMOCAP corpus annotations consist of the same six emotional categories, anger, excitement, frustration, happiness, neutral, and sorrow, and was also annotated on the same 3D scale. Whereas the lunar task dataset consists of approximately 4500 usable utterances, the IEMOCAP dataset is almost twice that size with roughly 8500 utterances. Comparing the distribution of categorical annotations between the two datasets in Table I, a much more severe class imbalance exists in the former with the majority of utterances considered to be neutral. This is logical, considering the dataset largely consists of previously unacquainted participants having natural, cooperative group discussions that were not specifically engineered to elicit emotional responses. For the 3D annotation format (see Table II), the distribution of the lunar task dataset is similar to that of the IEMOCAP such that in both cases, the neutral middle category captures the majority of the utterances.

For all four neural network architectures and for the two different emotion prediction tasks, performance is compared with and without employing transfer learning. With transfer learning, the network is first trained on the IEMOCAP dataset for 20 epochs.

E. Hyper-parameter optimization

The python library Hyperopt was used to optimize the network parameters for each deep learning architecture. Hyperopt compares the results of training each network on different combinations of parameters using an oriented random search.[45] The number of convolutional layers, filters, adn kernels, pool and stride size, as well as the number of LSTM units, epochs, and optimizers were all selected in this fashion. A summary of the final values for each parameter is shown in Table III.

IV. CLASSIFICATION RESULTS

The entire 14-meeting, 45-participant dataset was used for analysis. All results reported are the average after five-fold cross-validation in which 10% of the dataset is reserved for validation and 10% is reserved for testing.

A. Task 1 results: Categorical emotion classification

Table IV shows the unweighted accuracy (UA) and weighted-by-emotion-category (WA) classification results for each of the neural network architectures both with and

TABLE III. List of hyper-parameter values after optimization. If a hyper-parameter value differed between the categorical and multitask regression classification models, the latter is indicated with brackets.

Network	Epochs	CNN layers	CNN filters	CNN kernel	LSTM units	Optimizer
CNN	15	4	64 (×2), 128 (×2)	3 (×4)	—	SGD (Adam)
LSTM	15	—	—	—	512, 256	SGD (Adam)
CNN-LSTM	15	4	64 (×2), 128 (×2)	3 (×4)	256	Adam
Fusion	15 [25]	2	64, 128	8,4	128, 128	Adam

https://doi.org/10.1121/10.0003433

TABLE IV. Mean unweighted accuracy (UA) and weighted-by-category (WA; %) for fivefold cross-validation for each of the four neural network architectures on the lunar task dataset when no pretraining is implemented and when each network is first trained on the IEMOCAP database for 20 epochs.

	No pretrain		Pretrain	
Network	UA	WA	UA	WA
CNN	31.1	23.8	34.1	32.2
LSTM	26.5	26.0	26.6	24.5
CNN-LSTM	32.1	30.0	**34.2**	**32.6**
Fusion	**32.3**	**30.4**	31.2	29.4

without pretraining on the IEMOCAP database. Without pretraining, the fusion network outperforms all other network architectures but only by a slim margin. The performance of the CNN-LSTM network, in particular, differs by less than half of a percent for both UA and WA. The CNN, while performing nearly comparably when evaluated with the UA, severely misclassifies the sorrowful utterances (typically as neutral utterances), resulting in WA results that are almost 10% lower. Interestingly, this issue appears similar to that seen in the documented classification results for the IEMOCAP dataset in which happy utterances are almost entirely misclassified as a similar emotion, excitement.[46] For the lunar task dataset, because the discussion was more naturalistic, the utterances annotated as being sorrowful are likely fairly subtle, resulting in their appearance being similar to neutral utterances. Finally, the LSTM, overall, performed the worst likely because it accepts only a 1D input, and deep learning networks tend to favor more input data not less.

After pretraining each network on the IEMOCAP database for 20 epochs, the 2 neural network architectures that do not feature 1D inputs improve significantly despite the different nature of the emotional utterances contained in each database (acted/elicited vs natural). For the CNN architecture, performance improves by 3% and almost 9% for UA and WA, respectively. Thus, pretraining on the IEMOCAP database was able to ameliorate the inability of the CNN to classify one particular emotion, balancing out classification performance and reducing overfitting. The CNN-LSTM accuracies improved by a sufficient margin, thus, making this architecture the highest performing for the classification task. Specifically, the network reaches accuracies roughly 2% higher than the second highest performing architecture, the fusion network with no pretraining. The LSTM network and fusion network both suffer from slight accuracy decreases after pretraining, suggesting that the 1D training data are less generalizable from one dataset to the next.

B. Task 2 results: 3D emotion regression

For the second task, each network was trained on each of the 3D emotion attributes: activation, dominance, and valence. Tables V and VI show the CCCs for each of the affective dimensions, as well as the average CCC across all three dimensions, with and without pretraining on the IEMOCAP dataset.

TABLE V. Average CCC across fivefold cross-validation for arousal, dominance, and valence for each of the four neural network architectures on the lunar task dataset (no pretraining).

Network	Arousal	Dominance	Valence	Average
CNN	0.359	0.372	0.187	0.306
LSTM	0.305	0.252	0.0572	0.205
CNN-LSTM	**0.422**	**0.405**	**0.202**	**0.343**
Fusion	0.314	0.277	0.0481	0.213

Without pretraining, the CNN-LSTM network performs better than all of the other network architectures across all dimensions. The average CCC is notably higher than that resulting from the CNN, although the CNN performs far better than the LSTM and fusion networks perform. The LSTM network performs predictably poorly given the results from the classification task but, interestingly, the fusion network is almost as inaccurate. While the fusion network performed well for the classification task, for the regression task, it performs only slightly better than the LSTM network, which relies on a 1D input alone. Overall, across all four network architectures, valence is the most difficult dimension to classify, a result consistent with the majority of the literature publishing similar results.[18,24]

With pretraining, the CNN and CNN-LSTM networks perform nearly identically with the CNN very narrowly and, likely, negligibly achieving a higher average CCC. This outcome was not observed with the first classification task. Indeed, pretraining notably decreases the result for the CNN-LSTM network while delivering modest performance gains to the CNN. However, the CNN-LSTM network does perform better for the latter two dimensions, dominance and valence. Overall, for this second task, pretraining allows this simpler architecture to draw equal with the CNN-LSTM network. Again, the LSTM network performs significantly worse than the two 2D networks as with the categorical classification task. As expected, based on the results of the first task, pretraining does not improve the results from the fusion network enough to place it in competition with the networks trained only on 2D data.

V. CORRELATION AND REGRESSION ANALYSIS

To confirm the utility of constructing an automatic speech emotion classifier for the purposes of studying group dynamics, the degree of correlation between the ground-

TABLE VI. Average CCC across fivefold cross-validation for arousal, dominance, and valence for each of the four neural network architectures on the Lunar Task Dataset when each network is first trained on the IEMOCAP database for 20 epochs.

Network	Arousal	Dominance	Valence	Average
CNN	**0.404**	0.377	0.195	**0.325**
LSTM	0.332	0.248	0.0562	0.218
CNN-LSTM	0.394	**0.381**	**0.197**	0.324
Fusion	0.331	0.296	0.0544	0.227

truth speech emotion annotations and the post- and pre-task target variables of perceived leadership, contribution, and EI were examined for the entire 14-meeting, 45-participant dataset.

The correlation of many different low-level audiovisual metrics with various social-psychological group variables, such as perceived and/or emergent leadership, leadership style, dominance, and extroversion, have been extensively explored.[3,4,6–8,39,47,48] Our speech emotion metrics, both categorical and dimensional, may offer a more intuitive option than fine-grained, low-level metrics derived from visual focus of attention or prosodic acoustic features, for example.

To examine these correlations, two target variables, perceived leadership and perceived contribution, are defined using the post-task questionnaire described above. Because each group member rates the leadership and contribution of all other group members on a five-point scale, an individual's perceived leadership score can be computed as the average of all leadership scores that are received from that group. The perceived contribution score is defined similarly.

Using single variable regression, the Pearson correlation coefficient (ρ) can then be computed to understand the correlation between the speech emotion metrics and the post-task questionnaire variables of leadership and contribution. Using this framework in the categorical space, a lack of sorrowful utterances was found to have a correlation with contribution ($\rho = 0.42, p = 0.01$) and leadership ($\rho = 0.37, p = 0.01$).

In the 3D emotion space, EI was highly correlated with high valence ($\rho = 0.32, p = 0.04$), high activation ($\rho = 0.27$, $p = 0.08$), and low dominance ($\rho = 0.39, p = 0.01$). Additionally, the contribution was correlated with high valence ($\rho = 0.31, p = 0.05$), although there were no significant correlations with leadership. Finally, one of the "Big Five" personality traits, conscientiousness, was correlated with both low valence ($\rho = 0.34, p = 0.02$) and low activation ($\rho = 0.31, p = 0.05$).

We next carried out multiple linear regressions with the entire suite of speech emotion values for each participant against the post-task questionnaire variables to investigate the capability of the extracted metrics to *predict* the leadership and contribution scores for each participant (rather than simply investigating the correlation between the two).

To establish a ground-truth, the participant who received the highest overall leadership/contribution scores for their group was considered to be the perceived leader/major contributor (each group could have more than one perceived leader and major contributor). Because these ground-truth scores are quantized (given that they were reported on a scale from 1 to 5), we also quantized the predicted scores that are derived from the regression coefficients to the nearest actual bin to find the participant(s) with the highest scores. Therefore, an actual/ground-truth perceived group leader is the participant(s) with the highest received leadership score, and a predicted group leader is the participant(s) with the highest predicted quantized leadership score (derived from the linear regression coefficients).

Using this approach, the entire set of categorical emotion distributions for each participant is able to correctly predict 71% of emergent group leaders (i.e., for 10 of the 14 meetings) and 86% of major group contributors (12 of the 14 meetings). The linear coefficients derived from the affective dimensional emotion values are able to correctly predict only 50% of the leaders (7 of the 14 meetings) and 79% of the major contributors (11 of 14 meetings). These results indicate a meaningful relationship between the perception of emergent leadership and contribution and the emotional expressions of the participants even when the visual information and the informational content of the meeting itself is not taken into account.

VI. CONCLUSIONS AND FUTURE WORK

Whereas various low-level visual and acoustic metrics have been shown to correlate with social-psychological group metrics like emergent leadership and productivity, the intuitive reason for these correlations is often unclear. This paper instead proposes to use higher-level, more interpretable speech emotion categories estimated using various deep learning methods to predict emergent leadership and other group metrics.

After comparing four commonly used neural network architectures for two different speech emotion recognition tasks on a new group dynamics dataset, a fusion CNN-LSTM network architecture that combined 1D waveform and 2D spectrogram inputs performed best for a categorical classification task but only when pretraining on a larger speech emotion dataset was not implemented. With pretraining, a CNN-LSTM network architecture which only used 2D spectrogram inputs outperformed the fusion model accuracies (with and without pretraining), suggesting that although transfer learning significantly boosts classification accuracy in most cases, 1D waveform inputs may reduce cross-corpus model generalizability.

For the 3D speech emotion regression task, the same superiority of the fusion model before the implementation of transfer learning was not observed, and the CNN-LSTM network achieves the highest average CCC. However, with transfer learning, the CNN model architecture performed nearly equally with the CNN-LSTM network because for the second task, cross-corpus training did not improve CNN-LSTM network results. This emphasizes that a one-size-fits-all approach is not appropriate for different deep learning tasks, even when conducted on the same dataset.

Overall, despite the successful implementation of transfer learning from a less natural speech emotion dataset (IEMOCAP), to improve classification accuracies on a smaller, unbalanced, and more natural speech emotion dataset (lunar task), it is clear that deep learning methods are not yet sufficiently advanced to reliably extract the emotional content of natural speech without manual intervention. In particular, more training data of subtle, natural, emotional speech is needed.

However, the strong correlation between the ground-truth emotion annotations and perceived emergent leadership and contribution provide strong motivation to continue

this line of inquiry, even though the predicted emotion annotations (categorical and dimensional) were not in complete alignment with the observed ground-truth group dynamics trends.

In particular, sorrowful utterances were negatively correlated with both leadership and contribution, suggesting that participants with a more positive outlook were thought of as exerting more influence over the discussion, although conversely, sorrow could be a consequence of that participant being consistently overruled. EI was also highly correlated with high valence and activation, as well as with low dominance, indicating that participants who stayed engaged and positive and who did not dominate the conversation achieved higher EI scores.

Additionally, the entire set of categorical and dimensional speech emotion metrics were capable of correctly predicting the emergent leader and contributor for a majority of the group meetings in the dataset. This result advances the goal of using deep-learning-facilitated speech emotion as a tool for estimating the suite of social-psychological metrics that influence group meeting productivity.

To improve the algorithm performance for similar tasks, a multimodal neural network could be constructed with two new inputs: frames from the frontal-facing cameras and meeting transcripts obtained through natural language processing. With the building accuracy and reliability of speech-to-text transcription software, the informational content of such meetings could be included during training. Adding such transcripts, common visual focus of attention metrics (head pose, eye gaze, etc.) extracted using vision-based deep learning methods and the raw video frames themselves would result in an audiovisual model that could improve emotion classification accuracy significantly, achieving results impossible for a network incorporating only a single modality.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIP-1631674 and a Rensselaer Humanities, Arts, and Social Sciences Fellowship. Thank you to Jon Mathews, Lingyu Zhang, and Michael Foley for assistance in room instrumentation and carrying out the experiments, Christoph Riedl and Brooke Foucault Welles for experimental design, and Rahul Jain for ground-truth data annotation.

[1]G. De Souza and H. Klein, "Emergent leadership in the group goal-setting process," Small Group Res. **26**(4), 475–496 (1995).
[2]A. Pescosolido, "Informal leaders and the development of group efficacy," Small Group Res. **32**(1), 74–93 (2001).
[3]C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Multi-task learning of social psychology assessments and nonverbal features for automatic leadership identification," in *Proc. 19th ACM Int. Conf. on Multimodal Interact.* (ACM, New York, 2017), pp. 451–455.
[4]C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino, "Detecting emergent leader in a meeting environment using nonverbal visual features only," in *Proc. 18th ACM Int. Conf. on Multimodal Interact.* (ACM, New York, 2016), pp. 317–324.
[5]D. Sanchez-Cortes, O. Aran, and M. S. M. D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," IEEE Trans. Multimedia **14**(3), 816–832 (2012).
[6]I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. Foucault Welles, and R. Radke, "A multimodal-sensor-enabled room for unobtrusive group meeting analysis," in *Proceedings of the 2018 International Conference on Multimodal Interaction* (ACM, New York, 2018), pp. 347–355.
[7]L. Zhang, M. Morgan, I. Bhattacharya, J. Braasch, C. Riedl, B. F. Welles, and R. Radke, "Visual focus of attention estimation and prosodic features for analyzing group interactions," in *Proceedings of the International Conference on Multimodal Interaction* (ACM, New York, 2019), pp. 385–394.
[8]L. Zhang, I. Bhattacharya, M. Morgan, M. Foley, C. Riedl, B. Welles, and R. Radke, "Multiparty visual co-occurrences for estimating personality traits in group meetings," in *IEEE WACV* (Aspen, CO, 2020), pp. 2085–2094.
[9]P. Ekman, E. Sorenson, and W. Friesen, "Pan-cultural elements in facial displays of emotions," Science **164**, 86–88 (1969).
[10]J. Russel and A. Mehrabian, "Evidence for a three-factor theory of emotions," J. Res. Pers. **11**, 273–294 (1977).
[11]M. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Commun **116**, 56–76 (2020).
[12]F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology* (2005), pp. 1517–1520.
[13]C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," J. Lang. Resour. Eval. **42**(4), 335–359 (2008).
[14]M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE Int. Conf. on Multimedia* (IEEE, New York, 2008), pp. 865–868.
[15]C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, and G. Richard, "Safe corpus: Fear-type emotions detection for surveillance application," in *Proc. of LREC* (2006), pp. 1099–1104.
[16]F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *EmoSPACE in Proc. IEEE Face and Gestures* (2013).
[17]C. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," Artif. Intell. Rev. **43**(2), 155–177 (2015).
[18]G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiri, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP* (2016), pp. 5200–5204.
[19]C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation," in *INTERSPEECH: Speech, Music and Mind* (IEEE, New York, 2018).
[20]Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, New York, 2018), pp. 1771–1775.
[21]A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)* (IEEE, New York, 2017), pp. 1–5.
[22]Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**(11), 1675–1685 (2019).
[23]J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2020), pp. 6474–6478.
[24]Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," in *Interspeech 2018* (ISCA, 2018), pp. 3092–3096.
[25]R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," IEEE Trans. Affective Comput. **8**(1), 3–14 (2017).

[26]J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York 2017), pp. 2746–2750.
[27]S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge—AVEC '17* (ACM, New York, 2017), pp. 19–26.
[28]R. Milner, M. Jalal, R. Ng, and T. Hain, "A cross-corpus study on apeech emotion recognition," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop* (IEEE, New York, 2019), pp. 385–394.
[29]P. Song, X. Zhang, S. Ou, J. Liu, Y. Yu, and W. Zheng, "Cross-corpus speech emotion recognition using transfer semi-supervised discriminant analysis," in *Proceedings of the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (IEEE, New York, 2016), pp. 385–394.
[30]J. Hall and W. Watson, "The effects of a normative intervention on group decision-making performance," Hum. Relat. **23**(4), 299–317 (1970).
[31]S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb, "The 'Reading the Mind in the Eyes' Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism," J. Child. Psychol. Psychiatry **42**(2), 241–251 (2001).
[32]J. Chang, T. Sy, and J. Change, "Team emotional intelligence and performance: Interactive dynamics between leaders and members," Small Group Res. **43**(1), 75–104 (2012).
[33]D. Chrusciel, "Considerations of emotional intelligence (EI) in dealing with change decision management," Manag. Decis. **44**(5), 644–657 (2006).
[34]B. Rammstedt and O. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German," J. Res. Pers. **41**(1), 203–212 (2007).
[35]S. Kichuk and W. Wiesner, "The big five personality factors and team performance: Implications for selecting successful product design teams," J. Eng. Technol. Manage. **14**(3-4), 195–221 (1997).
[36]B. Barry and G. Stewart, "Composition, process, and performance in self-managed groups: The role of personality," J. Appl. Psychol. **82**(1), 62–78 (1997).
[37]P. Curşeu, R. Ilies, D. Virgă, L. Marticuţoiu, and F. Sava, "Personality characteristics that are valued in teams: Not always 'more is better'?," Int. J. Psychol. **54**(5), 638–649 (2018).
[38]R. Lord, "Functional leadership behavior: Measurement and relation to social power and leadership perceptions," Administ. Sci. Quart. **22**, 114–133 (1977).
[39]D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez, "Linking speaking and looking behavior patterns with group composition, perception, and performance," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (ACM, New York, 2012), pp. 433–440.
[40]R. Lord, R. Foti, and C. D. Vader, "A test of leadership categorization theory: Internal structure, information processing, and leadership," Organ. Behav. Hum. Perform. **34**(3), 343–378 (1984).
[41]S. Kozlowski, "Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations," Organ. Psychol. Rev. **5**(4), 270–299 (2015).
[42]M. Bradley and P. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994).
[43]M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Automatic Speech Recognition and Understanding Workshop* (2005), pp. 381–385.
[44]F. Chollet, "keras," available at https://github.com/fchollet/keras (Last viewed 7/13/2020).
[45]J. Bergstra, D. Yamins, and D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proc. 12th Python in Sci. Conf.* (Python, 2013), pp. 13–20.
[46]J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomed. Signal Process. Control **47**, 312–323 (2019).
[47]C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," IEEE Trans. Multimedia **20**(2), 441–456 (2018).
[48]I. Bhattacharya, "Unobtrusive analysis of group interactions without cameras," in *Proceedings of the 2018 on International Conference on Multimodal Interaction* (ACM, New York, 2018), pp. 501–505.