A Multi-Stream Recurrent Neural Network for Social Role Detection in Multiparty Interactions

Lingyu Zhang, *Student Member, IEEE* and Richard J. Radke, *Senior Member, IEEE*

***Abstract*—Understanding multiparty human interaction dynamics is a challenging problem involving multiple data modalities and complex ordered interactions between multiple people. We propose a unified framework that integrates synchronized video, audio, and text streams from four people to capture the interaction dynamics in natural group meetings. We focus on estimating the dynamic social role of the meeting participants, i.e., Protagonist, Neutral, Supporter, or Gatekeeper. Our key innovation is to incorporate both co-occurrence features and successive occurrence features in thin time windows to better describe the behavior of a target participant and his/her responses from others, using a multi-stream recurrent neural network. We evaluate our algorithm on the widely-used AMI corpus and achieve state-of-the-art accuracy of 78% for automatic dynamic social role detection. We further investigate the importance of different video and audio features for estimating social roles.**

***Index Terms*—Deep learning, multimodal features, sensor fusion, human conversation analysis, social signal processing.**

I. Introduction

AUTOMATIC human conversation analysis has received continuous attention in the field of social signal processing, opening new avenues for algorithms to be able to understand aspects of group discussion such as intention, mood, personality, leadership, dominance, and persuasiveness [1]–[8]. Multiparty conversation consists of complex, layered visual, audio, and language signals that mutually affect the participants in complex ways. We are particularly interested in estimating the emergent social functional role of each individual in the group, categorized into *protagonist, supporter, neutral, gatekeeper* [9], which has direct implications for participants' leadership, contribution, and productivity. The relationships between participants and their social roles change as the conversation unfolds, making the analysis more challenging.

Key behavior cues extracted from visual, audio, and language data have strong correlations with meaningful social signals. For example, in the language domain, the occurrence of words like "great", "yes", or "bad" convey the sentiment and attitude of the speaker [10], [11]. In the visual domain, gaze behavior is closely related to emotion, personality, and the status of the individual in the group. For example, people perceived as leaders tend to give visual attention more frequently to other participants [12], people with more frequent mutual gaze interactions tend to have higher emotional intelligence [13], and people who smile more often are more likely to score highly in Extraversion [14]. Non-verbal audio features such as prosodic metrics, turn-taking frequency, and silence have also been found to be valuable in analyzing human personality, persuasiveness, and social status in the group [1], [13].

Scenario	Person	Behavior cues
Effect 1: Different orders of the same action pairs can reflect different attitudes.		
1	A	Speaking
	B	Smiling → Frowning
2	A	Speaking
	B	Frowning → Smiling
Effect 2: Different listeners' responses can affect the speaker's behavior differently.		
3	A	Nodding head
	B	Speaking in low voice → Speaking in a confident tone
4	A	Attention focused elsewhere
	B	Speaking in high voice → Speaking in a halting way
Effect 3: Different speakers have different responses based on listeners' behavior.		
5	A	Smiling → Frowning
	B	Speaking → Speaking faster in a tense tone
6	A	Smiling → Frowning
	B	Speaking → Speaking slower with more eye contact

Fig. 1: Three complex scenarios of human interaction that can be captured in our framework.

One major drawback to conventional analysis in this area is that multiple feature modalities are typically (1) computed on an entire-meeting basis, and (2) fused by simply concatenating into one feature vector for correlation or prediction, with little consideration for the temporal dependencies of different behavior cues. This fails to capture important multiparty dynamics. For example, the first row of Figure 1 shows a situation in which the frequency and categories of behavior cues of a listener (Person B) are the same while Person A is speaking. However, the different orders of the same actions actually reflect different attitudes (i.e., in scenario 2 the listener conveys a more positive attitude compared to scenario 1). In the other rows of Table 1, we illustrate that different responses from a listener can change the way a given speaker reacts, or conversely that the same listener response can provoke different responses in different speakers. Clearly, aggregating data on a per-meeting basis would lose these critical temporal dependencies.

In this paper, we build upon the rich set of interdependent, multimodal features that have been shown to be related with key social signals. However, instead of naively fusing frequency-based multimodal features for social signal analysis, we model the sequential order of behavior cues within the

L. Zhang and R.J. Radke are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA (e-mail:zhangl34@rpi.edu; rjradke@ecse.rpi.edu).

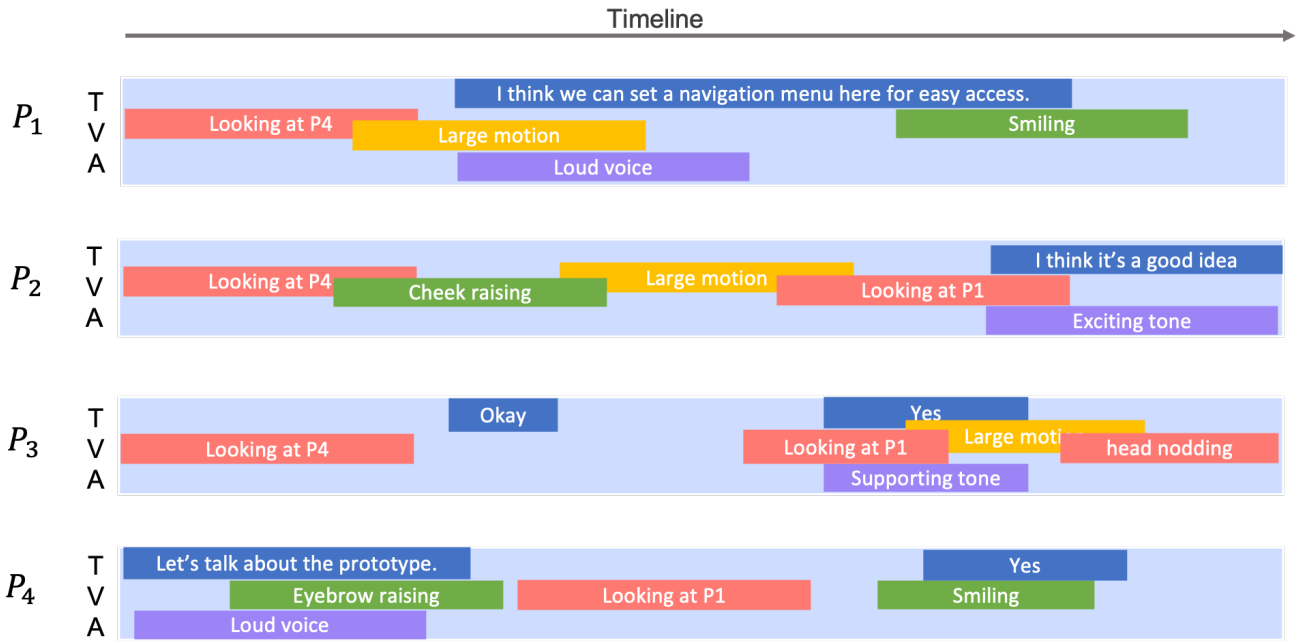


Fig. 2: An illustration of multimodal event streams from multiple people in a group conversation. Each person has three data tracks including text/language (T), visual (V), and audio (A). Within each track, we automatically extract key events such as visual focus of attention given to one participant, large motion intensity, or loud/fast audio.

multiple tracks, temporally fusing the extracted features in a timeline that allows us to better describe the co-occurrences and successive occurrences of different cues. We believe this provides a better basis for studying how a person's behavior and social status are dynamically shaped by the responses received from his/her conversation partners.

In particular, we propose a unified end-to-end framework to encode human interaction dynamics and automatically estimate social roles in a four-person group meeting scenario. As illustrated in Figure 2, group conversation involves multiple behaviors from multiple people; categorizing these cues into visual, audio, and text/language modalities results in 12 simultaneous tracks. These tracks are the input for our double-pyramid-shaped deep learning network. This consists of a multi-stack temporal information loading module for individuals' visual, audio, and language data and a frame-level temporal fusion module for intermediate results extracted from multiperson data. These are followed by a group-level interaction dynamics encoder for generating a representation of the four participants' behavior during the meeting, and a quadruple branch with fully connected layers for jointly estimating the four participants' social roles. This network allows us to explore interpersonal and intermodality effects in group discussion. The key contributions of our work include:

- **Comprehensive temporal modelling.** We model the interaction order in a multimodal multiparty meeting, considering the timing, speed, and interleaving of co-occurrences and successive occurrences of multiple social behavioral cues.
- **Interaction-based group behavior analysis.** When investigating the social role of the target subject, we consider both the behavior of the subject him/herself and the response from others interacting with him/her, which helps us to more accurately estimate the dynamics of the social roles.
- **Unified, extensible framework.** We develop an end-to-end framework using a meeting representation encoder for integrating behavior cues with temporal dependencies for multiperson data. While we study the problem of social role estimation in this paper, the framework would be directly extensible to other social signal analysis such as meeting highlight detection, meeting sentiment analysis, and emergent leadership/dominant contributor detection.
- **Interpretable model.** We visualize the variable-level importance of the trained model to interpret important features, providing a way to select the most relevant feature sets for future study.

To the best of our knowledge, this is the first attempt to design a unified framework for frame-level temporal fusion of multimodal data, jointly learning about multiple participants and their mutual effects on each other in group conversation. We evaluate our model on the widely-used AMI corpus [15] and demonstrate that we achieve better accuracy than competing state-of-the-art algorithms.

II. Related Work

A. Features for behavior analysis

Many multimodal features have been proposed for automatic group behavior analysis. Frequently used features in the vision domain include visual focus of attention (VFOA)

estimated from head pose and eye gaze [12], [13], [16], body movement [6] estimated by motion energy images (MEI) and motion history images (MHI) [17], and facial expressions encoded by the Facial Action Coding System [18], [19]. Based on facial muscle movements or basic facial expression classes like smiling or laughter [20], texture features extracted from the face region have also been used to estimate apparent personality traits [21] for a job screening process.

In the audio domain, prosodic features to describe tone, stress, or rhythm including pitch, energy, zero-crossing rate, and mel frequency cepstral coefficients (MFCC) [1], [13], as well as structural features including binary speaking status and speaking length [1], [22], have been proposed. In the language domain, the occurrences of key semantic words like "great" or "yes" are counted using Linguistic Inquiry and Word Count (LIWC) [22], and word embedding models are applied to convert the spoken words into a multi-dimensional vector [23].

Typically, meeting-wide features are computed in a statistical form based on per-frame cues, e.g., the mean and standard deviation of the average motion intensity during the meeting [2], [3], the percentage of time that large motion is observed [1], [6], the percentage of time that the target subject is looking at or being looked at by a group member [16], or the percentage of time that two group members have mutual gaze interaction [12], [13]. These aggregated features are then used in further correlation analyses and/or prediction models, e.g., to estimate personality traits or emergent leadership.

B. Social signal estimation algorithms

Given a set of meaningful multimodal features, various machine learning algorithms have been developed for automatic behavior analysis and social signal prediction. Sanchez-Cortes et al. [6] used the statistics of head/body motion and binary speaking status during a meeting along with audio/visual features to infer the emergent leader in group meeting using Support Vector Machines (SVMs) and the iterative classification approach (ICA). Staiano et al. [16] computed acoustic features and VFOA features at each video frame and applied Naive Bayes, Hidden Markov Models, and SVMs to predict the personality traits of the group members. In our group's previous work [12], [13], we applied Pearson correlation analysis and linear prediction to uncover the relationship between individual features and emergent leadership, personality, and contribution. We also considered the timing of the occurrences between participants' body, hand, and face movements for personality trait prediction [24]. Several end-to-end convolutional neural networks (CNNs) have also been developed for specific estimation purposes including a deep residual network for first impression analysis based on videos [25] and an ensemble regression framework for analyzing personality traits in single-person short video sequences [26].

Our work differs from these existing approaches in that instead of using frequency-based feature measurements, we perform a temporal fusion of the multimodal features to better model the feature dynamics over time, enabling us to take informative interpersonal interactions into consideration.

C. Multimodal fusion algorithms

Various fusion methods have been proposed to integrate multiple modalities. With respect to the level of the fusion layer, the early fusion method was proposed for feature-level combination [27]–[29], in which different modalities are combined into a single feature vector before being fed into any learning model. This approach is able to maintain the concurrence and correlation between different dimensions of the features, although problems may arise if the features in different modalities have different sampling rates or are hard to synchronize. Late fusion methods were proposed for decision-level combination [30]–[33], in which an individual classifier for each modality is trained and the final prediction is determined by the average or majority vote of all the individual classifiers. Late fusion can be more flexible in terms of the selection of the specific individual prediction model for each modality compared with early fusion, but it could lose important interaction information in the early stages. Hybrid fusion [34]–[36] is designed to combine the benefits of the two mechanisms, in which multiple fusion layers are designed and entangled to obtain the final prediction. Besides direct feature concatenation in early fusion or score combination in late fusion, learning-based methods for intermediate-level fusion have been proposed [37], [38]. In particular, multiple kernel learning [38] has been used to investigate which kernel function is better for modeling the inter-modality interaction. A neural network fusion layer [39] is designed for taking hidden representation features in two or three modalities as input and learning a fusion model for generating a joint representation.

In terms of aggregation strategies when feeding the multiple inputs to the neural network fusion layer, concatenating the intermediate features [40] is one of the most common methods. This provides a way to exploit the dependencies between different dimensions of the intermediate features and allows flexibility in the selection of the optimal neural network architecture. While concatenation can lead to extremely large vectors, Ortega et al. [41] proposed additive aggregation schemes for multiple modalities and then trained a neural network fusion model on top of the summation. Vielzeuf et al. [42] proposed that the combination of a classifier trained on unimodal features and a classifier trained on the weighted sum of hidden representations in different modalities can boost performance. Similarly, a multiplicative method [39] was proposed to help reduce the model size while maintaining the important correlations between different modalities.

In our work, we focus on neural network-based fusion methods for multiple modality integration and explore different aggregation strategies for the modalities to be fed into the neural network.

III. Social Role Annotation in the AMI Corpus

We conduct our research on the AMI corpus [15], in which participants have a natural group discussion about a new design project and are recorded by multiple sensors. Figure 3 shows a snapshot of one meeting in the AMI corpus. Four participants sitting on two sides of the table were given a topic

to discuss, and each individual was recorded by a frontal-facing camera and a headset microphone.

Fig. 3: In the AMI corpus, four participants sitting on two sides of a table were given a group discussion topic and recorded by individual frontal-facing cameras and headset microphones.

According to the annotation by Sapru and Bourlard [22], while the social role of each individual varies throughout the meeting, it can be modeled as constant during a thin meeting slice. In their annotation, 59 meetings with an average length of 0.57 hours were selected. Each meeting was cut into several thin slices and 37.7% of the data was annotated for dynamic social roles. In every meeting slice, each of the four participants was classified into a social role based on the following definition:

Protagonist: A participant who takes the floor and leads the discussion.

Neutral: A participant who acts as a listener and accepts others' ideas passively.

Gatekeeper: A participant who encourages communication and acts as a mediator.

Supporter: A participant who acts in a cooperative way and provides support to others' ideas.

Attacker: A participant who deflates others' status.

Different participants in a meeting slice can have the same social role. The annotated data contains 1714 meeting slices in total with an average length of 27.16 seconds and a maximum length of 51.04 seconds. The meeting slice length distribution is illustrated in Table I.

TABLE I: Lengths of the thin AMI corpus meeting slices containing social role annotation by Sapru and Bourlard [22].

Slice length (seconds)	Counts (percentage)
0-10	0.7%
10-20	0.9%
20-30	73.9%
30-40	16.5%
40-50	5.0%
50-60	3.0%

IV. Proposed Approach

As shown in Figure 4, our proposed approach consists of a set of multimodal feature descriptors and a double-pyramid-shaped network. The network contains modules for multi-stack information loading and alignment, temporal fusion, encoding group-level interaction dynamics, and deciding individuals' social roles. In the following sections we describe the different modules in more technical detail.

A. Multimodal feature descriptors

As shown in Figure 5, instead of feeding raw image or audio data into the framework, we use a set of off-the-shelf feature descriptors that have been shown to be closely related to social behaviors.

1) Visual features: For each video frame, we consider 4 different dimensions of visual features including facial expressions described by facial action units, head pose and eye gaze angles, visual focus of attention (VFOA), body movement, and physical facial appearance.

Facial expressions. We use the OpenFace toolkit [43] to compute the intensity value ranging from 1 to 5 for 17 action units such as inner brow raiser, cheek raiser, nose wrinkler, lip corner puller, or lid tightener.

Head pose and eye gaze behavior. We compute the head pose roll, pitch and yaw angles and the eye gaze direction in azimuth and elevation with respect to the individual recording camera using OpenFace, resulting in a 5-dimensional vector for each video frame. The angle vector is then fed into our previously proposed VFOA estimation model [13] trained on the AMI corpus to identify the visual target for each participant in the group. The possible targets can be "group member straight ahead", "group member to the left/right", or "group member in the diagonal direction", as well as non-person object classes including "table", "slide screen", "whiteboard" and "other".

Body movement. We calculate the motion intensity and the number of independently moving parts for each participant based on the motion template MEI [44], [45] over a short time window. The motion intensity and the number of moving parts are normalized by their maximum value during the meeting to eliminate the effects of environmental factors, thus reflecting the degree of movement relative only to the current meeting.

Physical facial appearance. We extract appearance-based facial features using OpenFace, consisting of 34 non-rigid shape parameters and the scale, rotation, and translation values based on the point distribution model (PDM).

The visual feature set for each participant is computed at every video frame and concatenated into a 65-dimensional feature vector. We use V_{pi} to denote the visual feature set of participant $p \in \{1, 2, 3, 4\}$ at the i-th video frame of the meeting slice.

2) Non-verbal audio features: We measure the **loudness**, **noisiness**, **brightness**, **timbre**, **pitch** and **rhythm** of the sound signals. We also use an audio signal analysis toolkit [46] to extract **short-term metrics** for additional audio features including zero-crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, MFCCs, Chroma coefficients and the standard deviation of Chroma.

When computing the audio features, the window size is set to be 50 milliseconds and the window step is set to be 40 milliseconds, resulting in 25 sets of features per second, which exactly aligns with the video frame sampling rate of the AMI corpus. Concatenated with the binary speaking status "speaking" or "silent" at the current video frame, we compute a 35-dimensional audio feature set A_{pi} for participant $p \in \{1, 2, 3, 4\}$ at the i-th video frame of the meeting slice.

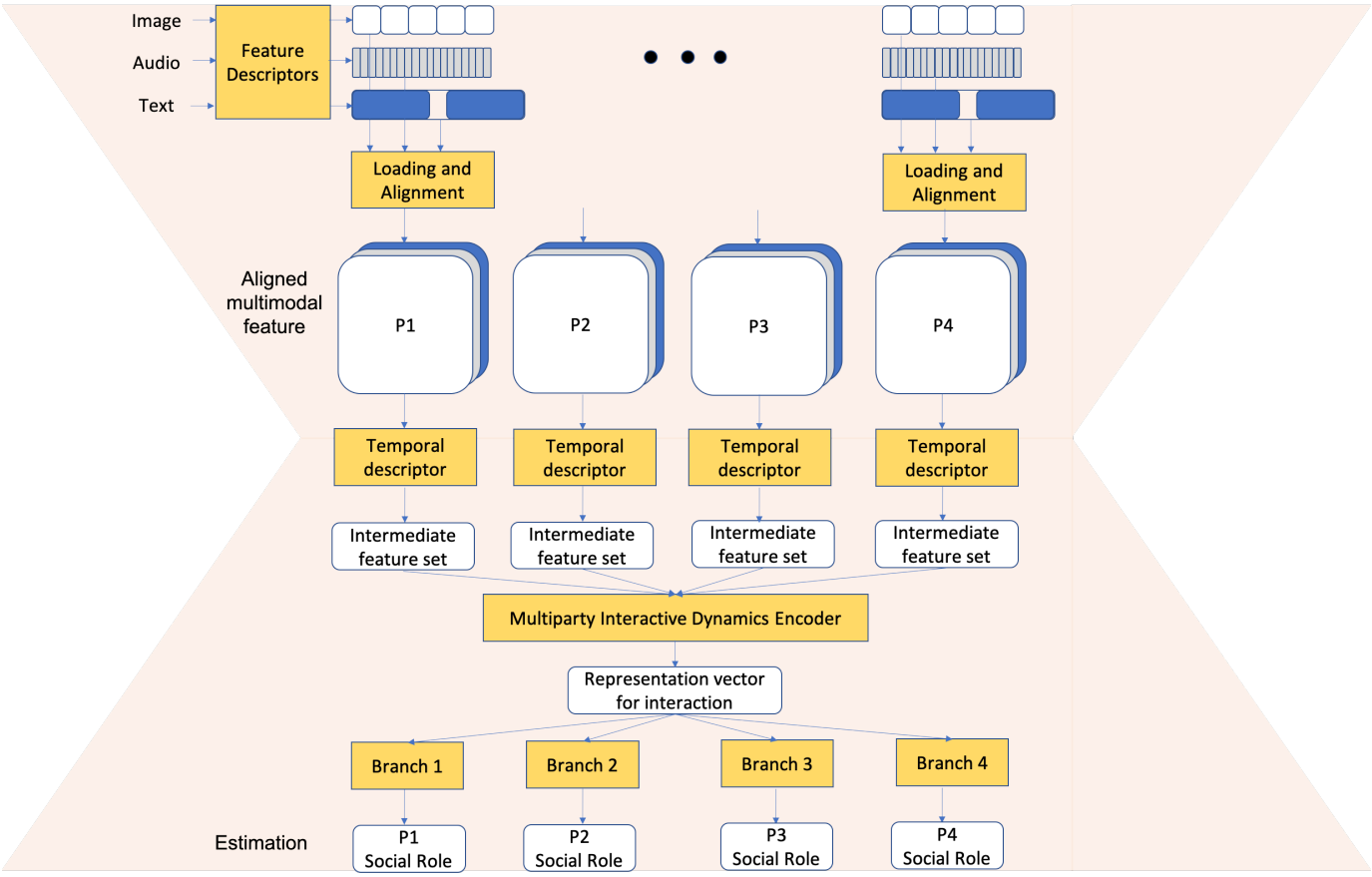


Fig. 4: Our framework contains a set of feature descriptors and a double-pyramid-shaped deep learning network that includes modules for multi-stack information loading and alignment, temporal description, encoding group interaction, and determining individuals' social roles.

3) Language content embedding: We apply the GloVe word embedding model with a vocabulary size of 1.2M trained on Twitter data [47] to convert each word in the group discussion transcript into a **100-dimensional word vector**. For each vectorized word W, we record the start time t_0 and end time t_1 and the corresponding start video frame index n_0 and end video frame index n_1 for further alignment with visual and audio features. We convert contractions into multiple words and split the corresponding time window equally for the resulting multiple words. For example, the word *we'll* with time window $[n_0, n_1]$ is converted into *we* with time window $[n_0, \frac{(n_0+n_1)}{2}]$ and *will* with time window $[\frac{(n_0+n_1)}{2}, n_1]$. There are several unknown words in the AMI transcripts with no sentiment or attitude meaning such as *button* or *wheel*; we assign the average value of all the vectors in the GloVe model as their text vector.

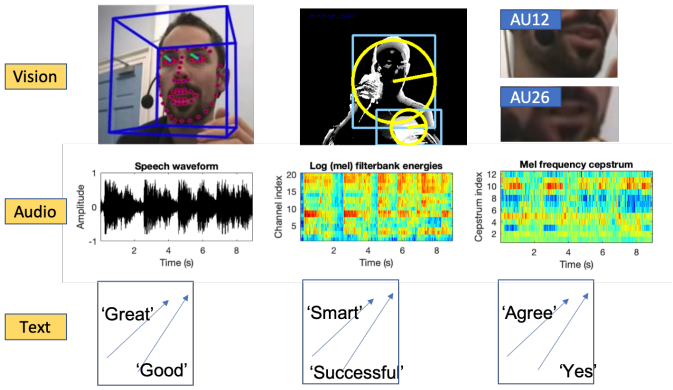


Fig. 5: A set of feature descriptors are used to extract video and audio features, and to convert each word in the transcript into a vector.

B. Multi-stream data alignment and loading module

Different modalities have different sampling rates. In particular, the audio and visual feature sets are extracted at a frame rate of 25 frames per second while the language content is more sparse. We align the sparse text vector to the dense audio/visual ones by duplicating it during the appropriate time window. Specifically, suppose that from video frame n_0 to n_1 participant p is speaking the word W. To align the text vector W with the audio and visual feature sets, we copy the vector W from n_0 to n_1 resulting in a text feature set $T_{pi} = W, i \in \{n_0, n_1\}$ for participant $p \in \{1, 2, 3, 4\}$ at the i-th video frame of the meeting slice.

In this way, we extract aligned visual, audio, and text feature

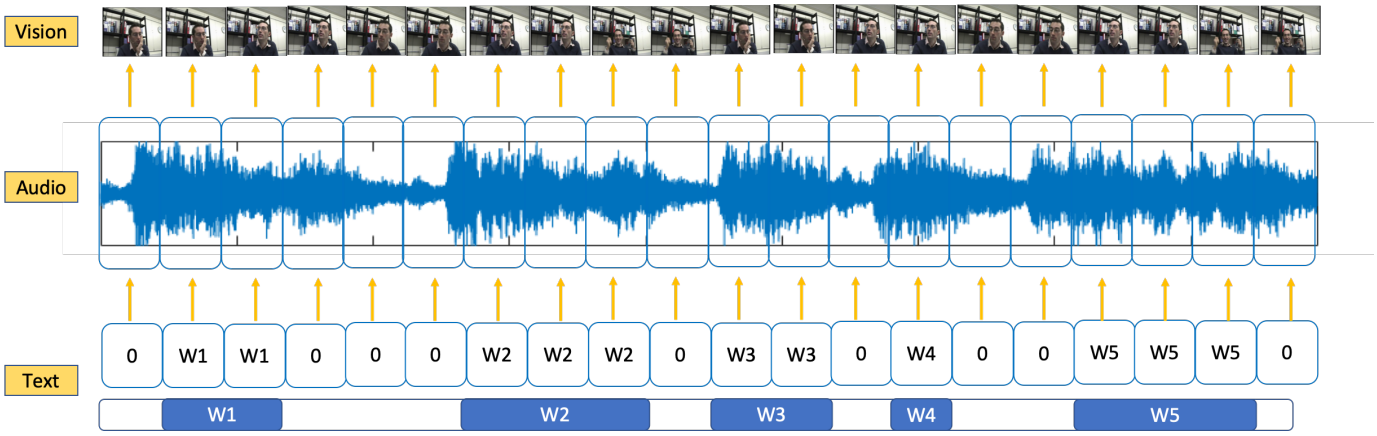


Fig. 6: Multimodal features are aligned based on the video frame sampling rate.

sets at 25 frames per second, which are then aligned with the annotation output label for the thin meeting slice.

The seating arrangement in a group meeting can also reflect and affect the social status and relationship between group members [48], and the VFOA class in the visual feature set is also relative to seat position. Thus, we encode each participant's spatial position in the data loading module in the order *lower right person, upper left person, lower left person, upper right person.*

For each meeting slice, we thus populate the data loading module with 12 tracks of feature sets in a specific seat order for further fusion and analysis. The whole process is illustrated in Figure 6.

C. Multiparty interaction representation learning module

After the multiple tracks of feature data are loaded into the temporal and multiparty fusion modules, we perform three levels of temporal feature fusion to learn a joint representation for multiparty interaction cues as described below.

Intra-modality temporal fusion. Within one track x^t of the data stream, we apply bi-directional long short term memory (LSTM) [49], [50] units to encode the individual temporal information for behavior cues. Taking the visual track of one participant as an example, in the forward direction, the visual feature x^t at timestamp t consists of facial expression e^t, head pose and eye gaze behavior g^t, movement m^t, and physical shape appearance s^t.

$$x^t = [e^t, g^t, m^t, s^t] \tag{1}$$

The input gate i^t at timestamp t is determined by the current input feature vector x^t and the past hidden state h^{t-1}:

$$i^t = \sigma\left(W_i[h^{t-1}, x^t] + b_i\right) \tag{2}$$

where W_i is the weight matrix, b_i is the bias vector, and σ is the non-linear activation function. For the backward direction, the states are updated using the same equations with reversed inputs.

The multiple dimensions of the visual track are mixed at this step by being multiplied by the weight matrix W_i, thus capturing the co-occurrences for different dimensions in the visual modality, e.g., having large body movement while looking at others.

The forget gate f_t is determined by the current input feature x^t and the past hidden state h^{t-1} with weight matrix W_f and bias vector b_f:

$$f^t = \sigma(W_f[h^{t-1}, x^t] + b_f) \tag{3}$$

The combination of the input feature x^t and the information from past hidden state h^{t-1} are multiplied by the weight matrix W_c and added to the bias vector b_c, and fed into a tanh function to get a new candidate cell $\tilde{C}^t$:

$$\tilde{C}^t = \tanh(W_C[h^{t-1}, x^t] + b_C) \tag{4}$$

Next, the new state cell C^t is updated based on the old state cell C^{t-1} and the new candidate state cell $\tilde{C}^t$:

$$C^t = f^t * C^{t-1} + i^t * \tilde{C}^t \tag{5}$$

Finally, the hidden state at the current timestamp t is computed based on the new state cell C^t and the output gate o^t determined by the weight matrix W_o and bias vector b_o.

$$o^t = \sigma\left(W_o[h^{t-1}, x^t] + b_o\right) \tag{6}$$

$$h^t = o^t * \tanh C^t \tag{7}$$

In this way, the output hidden units for timestamp t are dependent on the information of timestamp $t-1$, thus encoding the historical information along the timeline and enabling us to capture successive occurrences across multiple dimensions within one modality, e.g., smiling after head nodding.

Inter-modality feature fusion. For the visual, audio and text tracks of a single participant, in the time window $[1, t]$, we aggregate the intermediate results $V_{fh_{1:t}}, A_{fh_{1:t}}, T_{fh_{1:t}}$ of the hidden units in the forward direction as well as the hidden states $V_{bh_{1:t}}, A_{bh_{1:t}}, T_{bh_{1:t}}$ in the backward direction after the bi-directional LSTM layers via channel concatenation and then feed them into the neural network fusion module to encode inter-modality information for each person.

Figure 7 shows the inter-modality dependency capturing mechanism using the neural network. In this way, multiple

feature modalities during time window $[1, t]$ are fused, thus capturing the mutual effects of different tracks, e.g., saying "Great!" with an excited tone while giving visual attention to others and head nodding, or showing lip corner depressor after saying "Really" with a disagreeable tone.

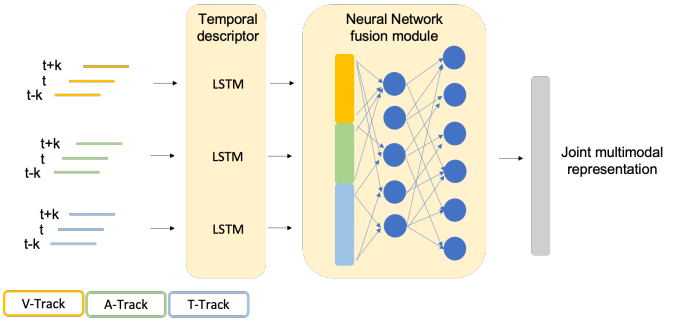


Fig. 7: Neural network-based inter-modality fusion.

Inter-personal feature fusion. Similarly, we include inter-personal effects in the concatenated layer and combine them into the neural network fusion model for joint consideration. Hidden states after the LSTM layer in all visual, audio, and text tracks among the four participants are mixed and fed into the batch normalization layer, followed by a set of fully connected layers to obtain the encoded representation vector χ for the four participants' dynamic interaction cues. Specifically, we use K to denote the input to the neural network encoder, and the joint representation χ is computed as

$$\chi = \sigma\left((W^n)^t H^n + b^n\right) \tag{8}$$

where $H^l, l \in \{1, 2, ..., n\}$ represents the intermediate result in one fully connected layer in the neural network

$$H^l = \sigma\left((W^{l-1})^t H^{l-1} + b^{l-1}\right) \tag{9}$$

$$H^{l-1} = \sigma\left((W^{l-2})^t H^{l-2} + b^{l-2}\right) \tag{10}$$

$$\begin{aligned} &... \\ H^1 &= \sigma\left((W^1)^t K + b^1\right) \\ &= \sigma\left((W^1)^t [K_1, K_2, K_3, K_4] + b^1\right) \end{aligned} \tag{11}$$

where

$$\begin{aligned} K_p = [&V_{fh_{1:t}}, A_{fh_{1:t}}, T_{fh_{1:t}}, \\ &V_{bh_{1:t}}, A_{bh_{1:t}}, T_{bh_{1:t}}] \end{aligned} \tag{12}$$

$$p \in \{1, 2, 3, 4\}$$

Therefore, through learning the mutual effects with the neural network, we are able to capture inter-personal dependencies during the group meeting, e.g., while person 1 is speaking, person 2 is smiling at first, but as the speech continues, person 2 seems to become confused with visual attention given to the paper on the table and brow going lower. Person 1 then changes her speaking speed to be slower and her tone becomes softer.

D. Decision-level branching module

Based on the encoded interaction dynamics vector, we then design a quadruple branching module for more dedicated inference for each person. The meeting dynamics representation vector is fed into individual fully connected layers in each branch. A one-hot label for the four possible social role classes for each participant is constructed. To jointly predict the dynamic social role for the four participants in the group, we use the cross-entropy loss for each branch in the network, and our final loss function is defined as the sum of losses on all four participant branches:

$$L(\hat{Y}, Y) = \sum_{p=1}^{4} L_p = \sum_{p=1}^{4} \left(-\sum_{c=1}^{4} y_{o,c} log(\hat{y}_{o,c}) \right) \tag{13}$$

where $c \in 0, 1, 2, 3$ is the class index, o is the sample index, $y_{o,c}$ is the binary ground truth indicator for sample o to be in class c, and $\hat{y}_{o,c}$ is the predicted probability for sample o to be in class c.

E. Network Structure and Implementation Details

Figure 8 illustrates the details of our network structure. Two LSTM layers are designed for each of the visual, audio, and text tracks; the input dimensions for the LSTM layers are 65, 35, and 100, which correspond to the input feature dimension for each modality of data. The aggregated multimodal multiperson data is then fed into fully connected layer to encode the meeting representation vector containing the information about multiparty interaction dynamics. Each of the quadruple branches consists of 4 fully connected layers with a ReLu activation function for non-linearity modeling. Cross-entropy loss is applied for each of the branches and the Adam optimizer is applied to minimize the loss function during the training process.

During training, we select the batch size to be 56, the initial learning rate to be 0.0001 with a decay factor of 0.1, and the drop out rate to be 0.5. The model is trained on an Nvidia RTX 2080 Ti machine and it takes about 1.5 hours for one round of the training process to converge.

V. Experimental Results

We evaluated our algorithm on 59 annotated meetings in the AMI corpus. Since the meeting is conducted in a group discussion scenario, as suggested in [22], we filter out the slices with the role annotation of *Attacker*, which account for less than 1% of the data, and use the rest of the data for training and evaluation. Following the same train/test data splitting scheme in [22], we do a K-fold validation to test the effectiveness of our network with $K = 22$. We randomly partition all of the meetings into 22 groups. During this process, we strictly follow the rule that the same participant is never in both the training and testing set at the same time, and that thin slices belonging to the same meeting will never be in both the training and testing set at the same time.

Within each round of validation, we select one group as the testing set and the remaining 21 groups as the training set.

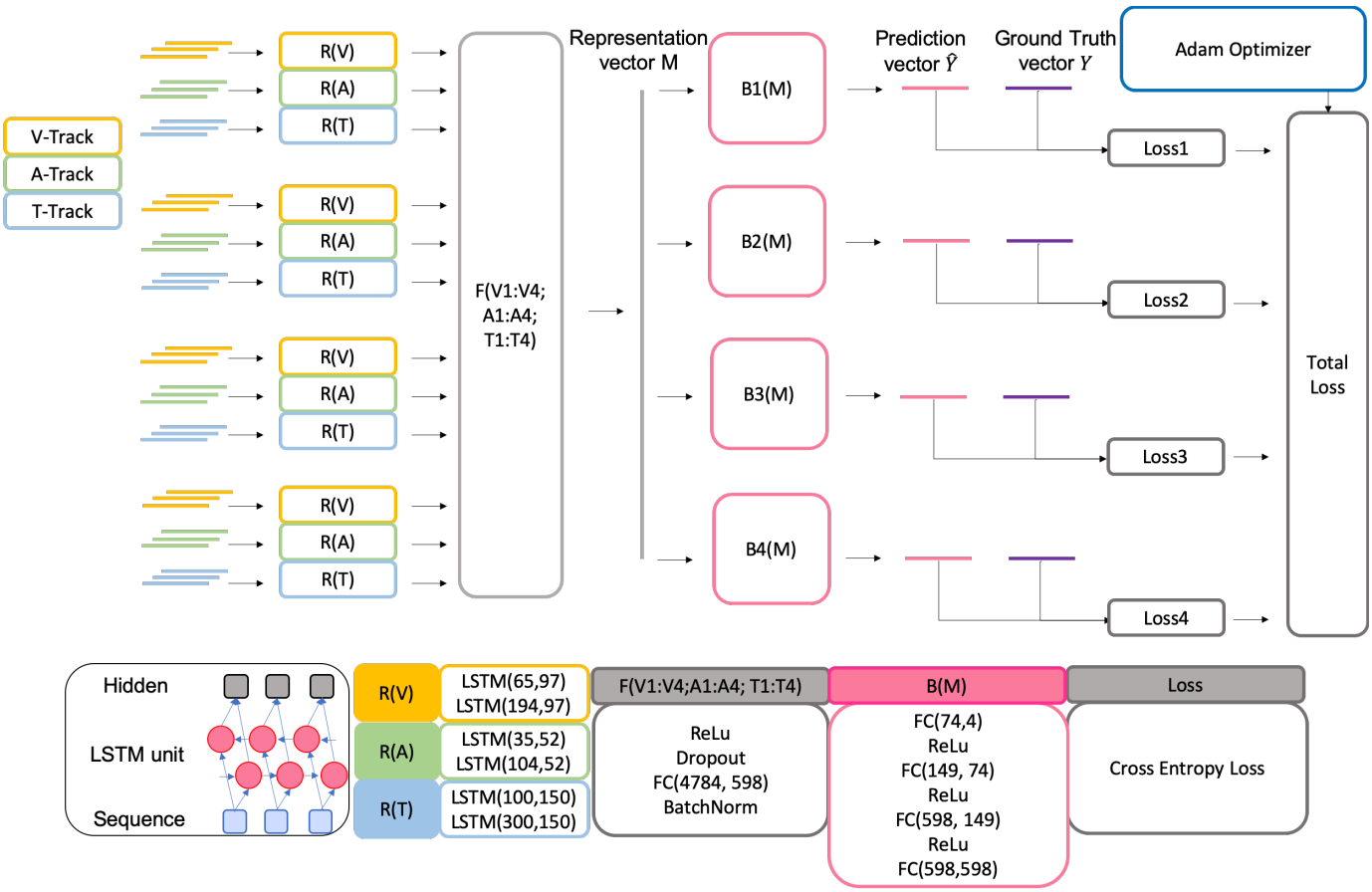


Fig. 8: Multiple tracks of feature data are fed into the LSTM units for temporal dependency extraction and fused together to obtain a unified representation vector. A quadruple branch is designed for jointly estimating individuals' social roles.

We run the validation for 22 rounds and compute the average value of the testing set accuracy in the 22 rounds as the final testing accuracy.

We compute both the F-measure and overall accuracy for our social role classification results. Specifically, for each meeting slice, the predicted result is denoted by $\{\hat{y}_{1i}, \hat{y}_{2i}, \hat{y}_{3i}, \hat{y}_{4i}\}$ and the ground truth label is denoted by $\{y_{1i}, y_{2i}, y_{3i}, y_{4i}\}$. The accuracy for validation round k is calculated as

$$Acc_k = \sum_{i=1}^{4N} \frac{\sum_{s=1}^{4}(\hat{y}_{rsi} = y_{rsi})}{4N} \tag{14}$$

where $i \in [1, N]$ is the index for the meeting slice, N is the total number of slices in the testing set, and s is the index of the participant in the group conversation. The F-measure of role r for validation round k is calculated as

$$F_k = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{15}$$

The final average accuracy for K-fold validation is

$$Acc_{final} = \frac{\sum_{k=1}^{K} Acc_k}{K} \tag{16}$$

A. *Feature Importance Analysis*

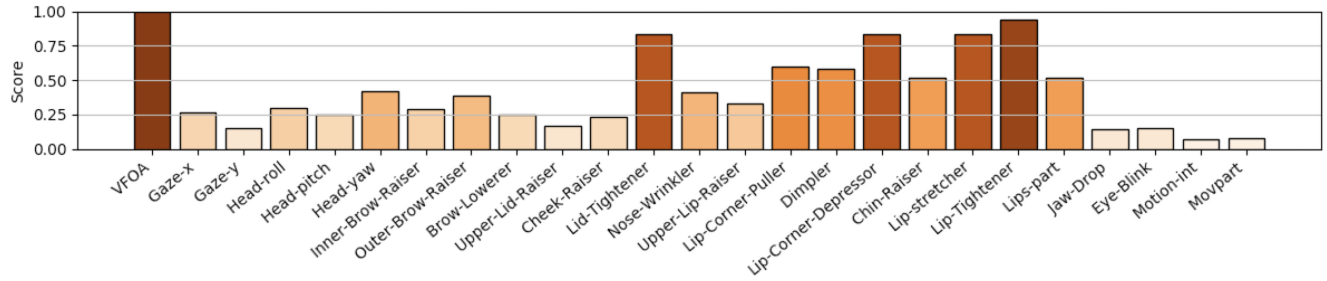
We first explore the importance of the features within each track. Following the procedure in [51], [52], we block out different dimensions of the variables and evaluate the change in the loss function. For each feature dimension to be blocked out, we perturb the original value with additive white Gaussian noise, apply the original model to the perturbed feature set, and compute the prediction loss. A larger change in the loss function means that an important variable has been blinded out. Specifically, the original input to the network Ψ contains the multimodal feature X for each of the participants $p \in \{0, 1, 2, 3\}$, where $X = [X_0, X_1, ..., X_{37}]$ contains 38 different features including 25 visual features, 12 audio features, and 1 text feature. The original prediction for input X is

$$\hat{Y} = \Psi(X) \tag{17}$$

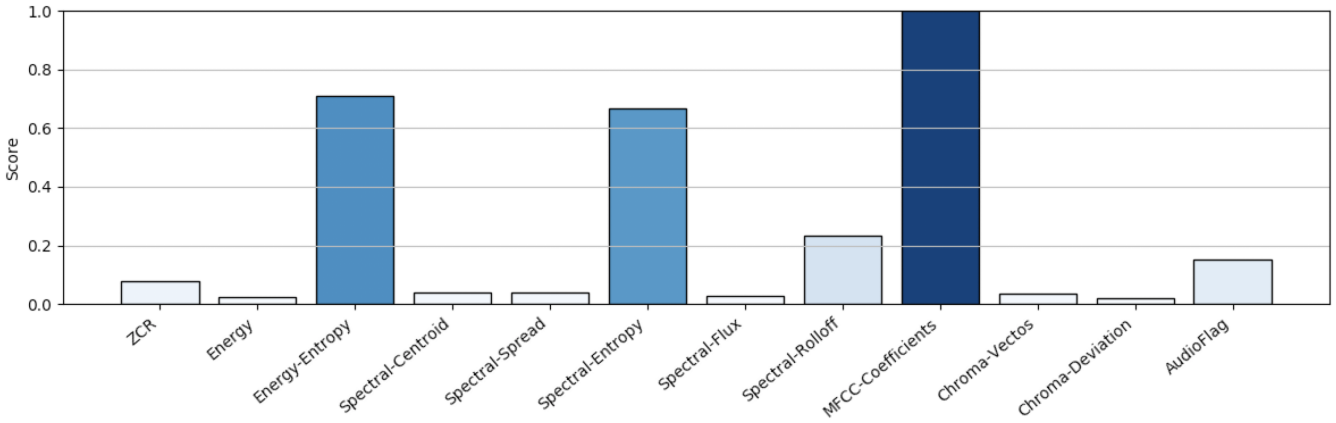
For the m^{th} feature to be blocked out, the variance of the additive white Gaussian noise ϵ_m for variable X_m is modeled as

$$\sigma^2 = (\alpha s)^2 \tag{18}$$

where s is the maximum amplitude of X_m and $\alpha = 0.1$ in our experiment.



(a) Visual feature importance heatmap



(b) Audio feature importance heatmap

Fig. 9: Feature importance heatmaps created by blocking out individual dimensions.

The prediction for the perturbed input $\tilde{X} = [X_0, X_1, ..., (X_m + \epsilon_m), X_{m+1}, ..., X_{37}]$ is

$$\tilde{Y} = \Psi(\tilde{X}) \tag{19}$$

The importance score for the m^{th} feature is defined as

$$\zeta(m) = |L(\hat{Y}, Y) - L(\tilde{Y}, Y)| \tag{20}$$

Since there is no semantic meaning to the individual dimensions of the 40-dimensional physical appearance/deformation parameters or the 100-dimensional embedded word vector, the blocking-out strategy is not applied for them here.

For the 22-fold validation, we calculate the loss changes on the 22 different testing sets and average them to get the final feature importance score for each variable. We then visualize these importances in the heatmaps shown in Figure 9, in which the color depth indicates the level of importance.

We observe that among the visual features, VFOA plays an important role in recognizing social interaction dynamics, which corroborates the existing literature that VFOA has strong correlations with personality traits and perceived leadership score [13]. Among the facial action units, "Lid-Tightener", "Lip-Tightener", "Lip-stretcher", and "Lip-Corner-Depressor" are the most important. The number of independently moving parts is more important than the overall movement intensity. The head yaw has the largest importance across the eye gaze and head behavior.

For the audio feature set, we observe that MFCC coefficients have the largest importance, indicating that the rhythm of the speech plays a crucial role in social interaction signaling. The entropy of energy also has a large weight, indicating that the loudness of the speech helps shape social functional roles to some degree. The flag for whether the participant is speaking also has a reasonably high weight, which is in line with existing findings [1], [22] that metrics related to speaking status such as speaking length and interruptions are closely related to group behavior. For the spectral dimension of the audio signal, spectral entropy and spectral rolloff are more important, indicating that the brightness of the speaker's voice can also affect the social interaction status.

B. Network Structure Analysis

We now further investigate the effectiveness of our model architecture by removing the intra-modality fusion, inter-modality fusion, or inter-personal feature fusion structures so that the multimodal multi-feature and multiparty co-occurrences or successive-occurrences cannot be captured in the baseline models. The details are as follows:

SMM: Single features, multiple tracks, and multiple participants. This baseline is designed to verify the importance of the intra-modality co-occurrence mechanism. For each of the visual and audio tracks, there is only a single feature within that track. Based on the analysis in the previous section about

the importance of subfeatures in the visual and audio tracks, the visual track only contains the action units features and the audio track only contains the MFCC features. We do not remove any features in the text track since it only contains the GloVe word embedding features for the transcript.

MSM: multiple features, single track, and multiple participants. To verify the importance of fusion among the multiple modalities, we constructed a baseline without inter-modality co-occurrences. The input variable for each participant only contains a single track, either visual (MSM-V), audio (MSM-A) or text (MSM-T).

MDM: multiple features, double tracks, and multiple participants. We further investigate whether each pair of modalities could be enough for representing interaction cues and whether there are any redundant modalities for social role estimation. This baseline involves limited inter-modality co-occurrences for any two tracks including audio-visual (A+V), audio-text (A+T) or visual-text (V+T).

MMS: Multiple features, multiple tracks, and single participant. For this baseline, we consider the participants in the same conversation separately. That is, the interaction between participants is ignored and we predict each user's social role using their own single-user behavior.

Figure 10 shows the K-fold validation results for the comparisons. Our proposed model is denoted by **MMM**, representing the full model with multiple features, multiple tracks and multiple participants. First, the comparison of our proposed model (MMM) with the baseline models SMM, MSM, MDM, and MMS shows that the prediction accuracy is decreased when we remove the fusion mechanism from our model, indicating that full joint consideration of modalities and co-occurrences is effective and important for social role prediction.

We note that the model with a single text track (MSM-T) achieves much higher accuracy than the model with single visual track (MSM-V) or a single audio track (MSM-A), showing that the transcript can reveal more information about social role relationships between group members. Additionally, the combination of audio and visual tracks (MDM-A+V) has better performance than the audio or visual track alone, indicating that while they have similar prediction accuracies, the two modalities provide complementary information.

The full triple-track model (MMM) surpasses all the double-track models, indicating that the combination of the visual, text and audio modalities is necessary for providing comprehensive interaction cues for the task. The single feature model (SMM) shows that co-occurrent or successive-occurrent features within a single track can be useful for describing complex behavior patterns. Furthermore, the single-participant model (MMS) can only reach 67% accuracy. while integrating behaviors and interactions between multiple group members improves the accuracy to 78%. This reflects the intuitive idea that one's social role is affected by others' behavior and responses.

C. Effectiveness of Quadruple-Branching Structure

We further verify whether the quadruple branching module on top of the joint representation vector is necessary to model the social roles of the four participants. Specifically, we feed the meeting dynamics representation vector into one set of neural network layers to predict a multi-output vector and use the cross-entropy loss for the unified classification task on 4 participants. We found that the quadruple branching module achieved an accuracy of 78% while the multi-output approach only achieved an accuracy of 73%.

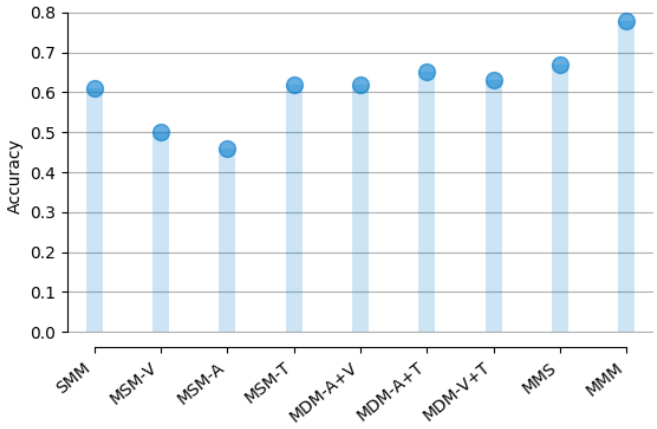


Fig. 10: Comparing the full model against various baselines with aspects removed.

D. Comparison with End-to-End Deep Learning

We next implemented two deep learning methods to investigate whether top-performing end-to-end models in video classification could achieve better results on our task compared with our structure, which learns a deep learning model on top of well-defined features.

C3D: Convolutional 3D network [53]. This baseline utilizes a 3D convolution module to help preserve temporal information between consecutive frames. Spatio-temporal features are extracted that capture human appearance, human action, and human-object interaction. We uniformly sample each input video to generate a set of images that represent the video clip compactly. The image set is then fed into 8 convolutional layers with 5 pooling layers and 2 fully connected layers to get the final classification result.

3DRes: 3D Residual network [54]. This baseline employs residual blocks on a convolutional 3D network containing 5 convolutional layers and 1 fully connected layer. Similar to C3D, the original video slice is uniform sampled to get a set of images with fixed length. From the second to the fifth convolutional module, shortcut connections are created by residual networks to bypass the signal at the top of the module, and the signals are summed from the top to the tail.

Since the C3D and 3DRes models require an abstraction parameter l, the length of the set of images for representing the original video clip, we vary this parameter and compare the best-obtained results with our proposed model (MMM). According to the results reported in Table II, despite their good performance on large-scale action classification, the C3D and R3D models performed quite poorly on social role prediction.

This is understandable since these two models take a set of images as input instead of considering features during the whole video clip. Crucial moments for the formation of the social roles are likely left out, and critical information from the audio and text modalities is never considered.

TABLE II: Comparisons with C3D and 3DRes.

Model	Accuracy
C3D	0.18
3DRes	0.09
MSM-V	0.50
MMM	0.78

We noticed that while increasing the number of images for video clip representation does improve the accuracy, this approach is limited since the input vector is fed into the 3D convolutional layer, requiring a huge number of parameters to be learned. Additionally, unlike the relatively unambiguous labels in action classification, social roles are far more subtle and difficult to infer by visual features only, which we showed earlier by looking at the poor accuracy of the baseline model MSM-V.

E. Classification Accuracy Analysis

We next compare the performance of our model with the reported accuracy in the existing literature.

Mboost: Multiclass Boosting. Sapru et al. reported results on a subset of the AMI corpus containing 5 meetings using a multiclass boosting algorithm [55]. One-level decision trees are utilized as weak learners using a feature set consisting of prosodic, turn-based structural, and lexical features.

DBN: Dynamic Bayesian models. Vinciarelli et al. employed dynamic Bayesian models [56] on the same subset used in [55]. Dependencies between social role and speaker turn-taking patterns, prosody features, and speaking turn duration are integrated into the dynamic Bayesian network whose parameters are estimated using maximum likelihood.

HCRFs: Hidden conditional random fields. The state-of-the-art result for social role classification was reported by Sapru et al. [22] using hidden conditional random fields. This model includes features from both long-term (around 30 seconds) and short-term (around 2 seconds) windows. During the long-term window, acoustic features and structural features including total speech time, total number of turns, and statistics like maximum, minimum, or standard deviation value of those numbers are extracted in conjunction with linguistic features for language style representation. The prediction for the social role is determined by both the relationship between the role and the observed features and the role transition probability between adjacent meeting slices. Parameters are estimated by maximizing the conditional log likelihood of the role sequence.

Table III shows our $K-$fold validation results. We improve the total accuracy from the state-of-the-art result 74% [22] to 78%, demonstrating the effectiveness of our model. Additionally, since the existing approaches only consider clip-level statistics instead of the frame-level interaction cues in our proposed model, the comparison indicates that temporal interaction dynamics and mutual effects across multiple modalities are important.

One difficulty with the dataset is that the class distribution is unbalanced [22], The Neutral and Supporter roles comprise the majority of the labels (49% and 28%), while the Gatekeeper and Protagonist roles account for fewer labels (14% and 9%). Our model's 0.63 F-score for Protagonist demonstrates that in addition to the overall high accuracy, our model is able to correctly identify the small amount of data with this class label.

We note that our F-scores on the Supporter and Gatekeeper roles are lower than HCRFs. We hypothesize that the difference in performance on these two classes is due to the type of features used by the HCRF method, specifically the Linguistic Inquiry and Word Count (LIWC) features [57], [58] for text analysis that include more subjective and psychological aspects of words. These aspects might be more relevant to the Supporter and Gatekeeper roles, which have a more emotional component than the Protagonist and Neutral roles. We also observed that the Gatekeeper and Supporter roles are less likely to change slice-to-slice, making them easier to predict.

F. Strategies for Multimodal Fusion

In our proposed model, intermediate features produced from the bi-directional LSTM layer for different modalities are aggregated via channel concatenation and then fed into a set of neural network layers to learn a joint meeting dynamics representation. Specifically, the intermediate result after the first fully connected layer when learning the joint representation H^1 is computed as

$$H = \sigma\left((W^1)^t K + b^1\right) \tag{21}$$

where K is the aggregated intermediate features.

Here, we further explore other options for aggregation that could improve the performance or reduce the model size compared with concatenation. The details are as follows:

ADD: Proposed NN-based fusion + Additive aggregation. For this baseline model, outputs from the bi-directional LSTM layer are summed up and the neural network fusion module is trained to learn a meeting dynamics vector. The aggregated intermediate feature set for each participant p is described as:

$$K_p = \begin{bmatrix} V_{fh_{1:t}} + A_{fh_{1:t}} + T_{fh_{1:t}} \\ V_{bh_{1:t}} + A_{bh_{1:t}} + T_{bh_{1:t}} \end{bmatrix} \tag{22}$$

Strategy S_1 in Figure 11 shows the additive fusion mechanism. The joint representation vector is then fed into the same quadruple branching module for individual social role prediction.

MULT: Proposed NN-based fusion + Multiplicative aggregation. Multiplicative aggregation is applied to the neural network-based fusion layer for inter-modality feature integration. Element-wise multiplication is performed across the hidden units for the three modalities. Strategy S_2 in Figure 11 shows the multiplicative fusion mechanism.

$$K_p = \begin{bmatrix} V_{fh_{1:t}} * A_{fh_{1:t}} * T_{fh_{1:t}} \\ V_{bh_{1:t}} * A_{bh_{1:t}} * T_{bh_{1:t}} \end{bmatrix} \tag{23}$$

TextADD: Proposed NN-based fusion + Text-assisted additive aggregation. According to the experimental results in

TABLE III: K-fold validation on the AMI corpus for social role detection

Model	Per-role F-measure				Accuracy
	Protagonist	Supporter	Neutral	Gatekeeper	
Mboost [55]	0.74	0.62	0.46	0.37	66%
DBNs [56]	–	–	–	–	68%
HCRFs [22]	0.62	0.79	0.75	0.64	74%
Proposed	**0.63**	0.71	**0.89**	0.60	**78%**

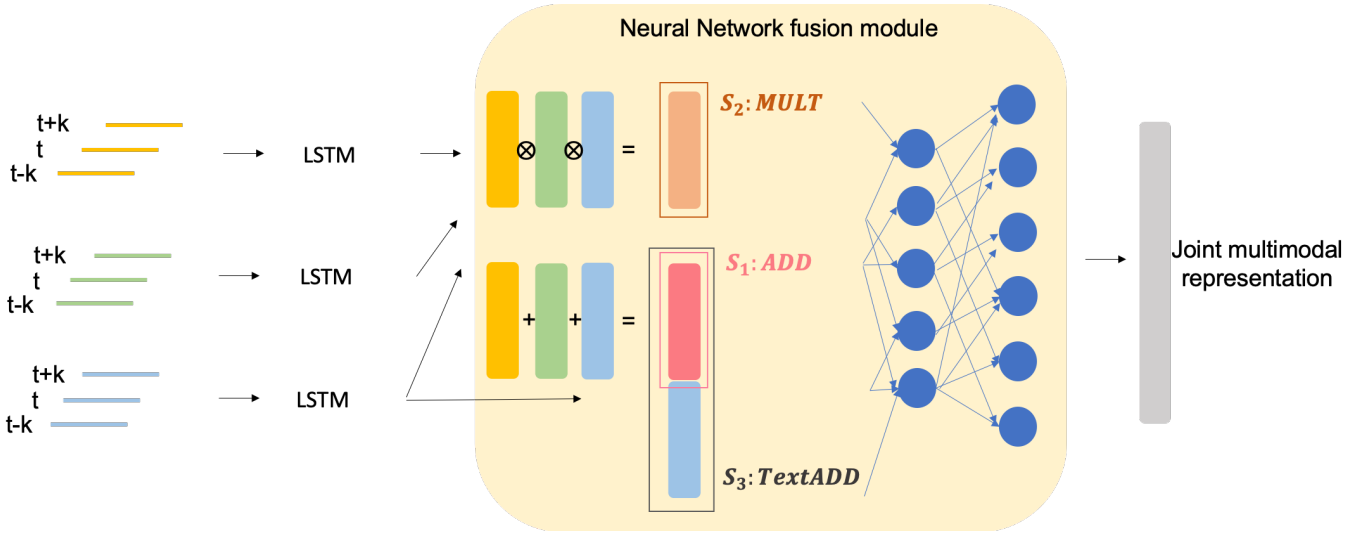


Fig. 11: Neural network-based inter-modality fusion using various aggregation strategies.

Section V-A, the text track contains more information than the audio or visual tracks. Since additive aggregation can lose some delicate information in the individual tracks, inspired by the work in [59], we construct a baseline model in which the intermediate features in the three modality tracks are first summed and then concatenated with the text track features before feeding into the information fusion layer. Therefore, the rich information in the text track can assist the aggregated features via addition as well as reducing the model size compared with full channel concatenation. Strategy S_3 in Figure 11 shows the text-assisted fusion mechanism. Specifically, the aggregated intermediate feature set K_p for each paricipant p can be computed as:

$$K_p = \begin{bmatrix} V_{fh_{1:t}} + A_{fh_{1:t}} + T_{fh_{1:t}} \\ V_{bh_{1:t}} + A_{bh_{1:t}} + T_{bh_{1:t}} \\ T_{fh_{1:t}} \\ T_{bh_{1:t}} \end{bmatrix} \tag{24}$$

Table IV shows the comparison results for different aggregation strategies. Since both the additive (ADD) and multiplicative (MULT) methods are compressing the signals across multiple modalities, aggregating intermediate features via channel concatenation (CONCAT) maintains the dependencies to a greater degree, and our proposed method achieves the highest classification accuracy. While CONCAT also has the largest model size at 39.8MB, ADD has almost half the model size while maintaining a good prediction performance of 77% accuracy. Despite our hypothesis, text assisted channel concatenation (TextADD) did not show significant improvement, achieving 75% prediction accuracy.

TABLE IV: Analysis of aggregation strategy in neural network-based multimodal fusion.

Model	Accuracy	Model size
ADD	0.77	20.3MB
MULT	0.73	20.3MB
TextADD	0.75	21.6MB
CONCAT	0.78	39.8MB

VI. Discussion and Future Work

We designed a unified framework for multiparty group conversation dynamics to jointly estimate the dynamic social role of four participants. We exploited the temporal information across multiple modalities of the individual feature data and the mutual effects across inter-personal feature data. We evaluated our algorithm on the AMI group meeting corpus and achieved better accuracy than the competing state-of-the-art.

In the short term, it would be worthwhile to integrate LIWC features [57] into our text analysis framework in addition to the GloVe word embedding, with the hope that the more psychological categories would have more bearing on interpersonal dynamics and interesting relationships to gestures or facial expressions in the video. In addition, it would be worthwhile to explore more complex methods for intermodality feature fusion. While the strategies in Section V-F generally treated the modalities as independent streams, clearly the text and

audio streams are more correlated with each other than either is with video, and relationships between the streams may occur across several seconds [48], [60]. While the LSTM in our framework implicitly captures these time-offset relationships, it may be interesting to find ways to explicitly associate video events with key textual or non-verbal events (e.g., gestures that may lead or lag the statement of a key idea) and give these events additional weight.

One promising future direction is to extend this framework to recognize highlights of the meeting in the temporal domain, e.g., identifying the crucial moment at which the social functional statuses of participants in the group are formed/changed, or at which point an insightful idea is proposed that plays a major role in shaping the group. Existing work on multimodal highlight detection is mainly done in the context of sports events, such as Joshi et al. [61] in which crowd cheer, commentator excitement, and player celebration are detected and a combined excitement score is used for determining the important moments in golf games. Xiong et al. [62] proposed an unsupervised approach for sports highlight detection in which video duration is used as a latent signal and features from short video clips are combined with features in long videos for optimized inference. The challenge in detecting highlights in a meeting scenario is that the context information from the environment is limited, and the highlights in the group discussion task are more complex and latent than in a sports game.

In our scenario, identifying meeting highlights could help us to detect "keypoints" in the temporal domain, supporting automatic meeting summarization or abstract generation. It would be interesting to develop a model for generating a representation vector that encodes all the important interaction dynamics for the thin slices of group discussion. This "feature descriptor" for the multiparty meeting slice could then be applied to different social dimension predictions including the joint estimation of leadership scores, perceived contribution, dominance, or Big-Five personality traits.

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation under award IIP-1631674 from the PFI:BIC program.

REFERENCES

[1] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*. ACM, 2015, pp. 15–22.

[2] O. Aran and D. Gatica-Perez, "Cross-domain personality prediction: from video blogs to small group meetings," in *Proceedings of the 15th ACM International Conference on Multimodal Interaction*. ACM, 2013, pp. 127–130.

[3] O. Celiktutan, P. Bremner, and H. Gunes, "Personality classification from robot-mediated communication cues," in *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016.

[4] P. Ekman and W. V. Friesen, "Head and body cues in the judgment of emotion: A reformulation," *Perceptual and Motor Skills*, vol. 24, no. 3 PT 1, pp. 711–724, 1967.

[5] G. Beattie, *Rethinking Body Language: How Hand Movements Reveal Hidden Thoughts*. Routledge, 2016.

[6] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 816–832, 2011.

[7] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[8] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, "Multimodal analysis and prediction of persuasiveness in online social multimedia," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 3, p. 25, 2016.

[9] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," in *Proceedings of the 8th International Conference on Multimodal Interfaces*. ACM, 2006, pp. 28–34.

[10] H. Jang and H. Shin, "Language-specific sentiment analysis in morphologically rich languages," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 498–506.

[11] Z. Teng, D. T. Vo, and Y. Zhang, "Context-sensitive lexicon features for neural sentiment analysis," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1629–1638.

[12] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. F. Welles, and R. J. Radke, "A multimodal-sensor-enabled room for unobtrusive group meeting analysis," in *Proceedings of the 2018 International Conference on Multimodal Interaction*. ACM, 2018, pp. 347–355.

[13] L. Zhang, M. Morgan, I. Bhattacharya, M. Foley, J. Braasch, C. Riedl, B. F. Welles, and R. J. Radke, "Improved visual focus of attention estimation and prosodic features for analyzing group interactions," in *Proceedings of the 2019 International Conference on Multimodal Interaction*. ACM, 2019.

[14] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ACM, 2012, pp. 53–56.

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

[16] J. Staiano, B. Lepri, R. Subramanian, N. Sebe, and F. Pianesi, "Automatic modeling of personality states in small group interactions," in *Proceedings of the 19th ACM International Conference on Multimedia*. ACM, 2011, pp. 989–992.

[17] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 257–267, 2001.

[18] P. Ekman and W. Friesen, *Manual of the Facial Action Coding System (FACS)*. Consulting Psychologists Press, 1978.

[19] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: The manual on CD ROM," *A Human Face, Salt Lake City*, pp. 77–254, 2002.

[20] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*. ACM, 2009, pp. 99–106.

[21] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1660–1663.

[22] A. Sapru and H. Bourlard, "Automatic recognition of emergent social roles in small group interactions," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 746–760, 2015.

[23] L. Boratto, S. Carta, G. Fenu, and R. Saia, "Using neural word embeddings to model user behavior and detect user segments," *Knowledge-based Systems*, vol. 108, pp. 5–14, 2016.

[24] 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 *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[25] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 349–358.

[26] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2017.

[27] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2539–2544.

[28] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 2005, pp. 399–402.

[29] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[30] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 172–179.

[31] E. Morvant, A. Habrard, and S. Ayache, "Majority vote of diverse classifiers for late fusion," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2014, pp. 153–162.

[32] S. T. Strat, A. Benoit, P. Lambert, H. Bredin, and G. Quénot, "Hierarchical late fusion for concept detection in videos," in *Fusion in Computer Vision*. Springer, 2014, pp. 53–77.

[33] M. R. Alam, M. Bennamoun, R. Togneri, and F. Sohel, "A confidence-based late fusion framework for audio-visual biometric identification," *Pattern Recognition Letters*, vol. 52, pp. 65–71, 2015.

[34] B. Rajalingam and R. Priya, "Multimodality medical image fusion based on hybrid fusion techniques," *International Journal of Engineering and Manufacturing Science*, vol. 7, no. 1, pp. 22–29, 2017.

[35] P. M. Portillo, G. P. García, and G. A. Carredano, "Multimodal fusion: a new hybrid strategy for dialogue systems," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, 2006, pp. 357–363.

[36] J. Ni, X. Ma, L. Xu, and J. Wang, "An image recognition method based on multiple BP neural networks fusion," in *International Conference on Information Acquisition*. IEEE, 2004, pp. 323–326.

[37] G. Iyengar and H. J. Nock, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2003, pp. 255–258.

[38] S. Luo, S. M. Alqhtani, and J. Li, "Multiple kernel-based multimedia fusion for automated event detection from tweets," *Machine Learning: Advanced Techniques and Emerging Applications*, p. 49, 2018.

[39] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[40] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6966–6975.

[41] J. D. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," *arXiv preprint arXiv:1907.03196*, 2019.

[42] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

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

[44] J. W. Davis and A. F. Bobick, "The representation and recognition of action using temporal templates," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 928–934.

[45] G. R. Bradski and J. W. Davis, "Motion segmentation and pose recognition with motion history gradients," *Machine Vision and Applications*, vol. 13, no. 3, pp. 174–184, 2002.

[46] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[47] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[48] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[50] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[51] R. M. Steele and C. Jaynes, "Overconstrained linear estimation of radial distortion and multi-view geometry," in *European Conference on Computer Vision*. Springer, 2006, pp. 253–264.

[52] R. Kalaivani and S. Chidambaram, "Additive Gaussian noise based data perturbation in multi-level trust privacy preserving data mining," *International Journal of Data Mining & Knowledge Management Process*, vol. 4, no. 3, p. 21, 2014.

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[54] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3154–3160.

[55] A. Sapru and F. Valente, "Automatic speaker role labeling in AMI meetings: recognition of formal and social roles," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5057–5060.

[56] A. Vinciarelli, F. Valente, S. H. Yella, and A. Sapru, "Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the AMI meeting corpus," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2011, pp. 374–379.

[57] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[58] A. Sapru and H. Bourlard, "Investigating the impact of language style and vocal expression on social roles of participants in professional meetings," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 324–329.

[59] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[60] K. Watanuki, K. Sakamoto, and F. Togawa, "Analysis of multimodal interaction data in human communication," in *Third International Conference on Spoken Language Processing*, 1994.

[61] D. Joshi, M. Merler, Q.-B. Nguyen, S. Hammer, J. Kent, J. R. Smith, and R. S. Feris, "IBM high-five: Highlights from intelligent video engine," in *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 2017, pp. 1249–1250.

[62] B. Xiong, Y. Kalantidis, D. Ghadiyaram, and K. Grauman, "Less is more: Learning highlight detection from video duration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1258–1267.