Exploiting observers' judgements for nonverbal group interaction analysis

Gokul Chittaranjan[1,2], Oya Aran[1] and Daniel Gatica-Perez[1,2]
Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne, Switzerland
{gthatta, oaran, gatica}@idiap.ch

Abstract— **Incorporating annotators' knowledge into a machine-learning framework for detecting psychological traits using multimodal data is an open issue in human communication and social computing. We present a model that is designed to exploit the subjective judgements of multiple annotators on a social trait labeling task. Our two-stage model first estimates a ground truth by modeling the annotators using both the annotations and annotators' self-reported confidences. In the second stage, we train a classifier using the estimated ground truth as labels. We also define ways to verify the consistency of our model and validate it using annotations and nonverbal cues for a dominance estimation task in a group interaction scenario on the publicly available DOME corpus, in addition to synthetically generated data. Our models give satisfactory results, outperforming the commonly used majority voting as well as other approaches in the literature.**

I. Introduction

In many studies conducted in the areas of affective and social behavior, researchers are interested in finding psychological traits. Traits are defined as habitual patterns of behavior, thought, and emotion [9] that are often identifiable in group interactions [11]. Whether a subject has a trait or not is identified using human judgements from either external observers or from the subjects themselves. In the former case, the trait is known to be "perceived by others" and in the latter, to be "self-perceived". In the recent past, there have been many studies on automated methods for detecting important traits in face-to-face group interactions such as dominance [14], [1], [8], leadership styles [7], and other traits related to personality [12]. In the studies that involve the detection of traits perceived by others, multiple annotators are often asked to label the presence (or absence) of these traits using either direct questions or standard psychology questionnaires. The availability of online annotation resources (e.g. Mechanical Turk) opens up the possibility of obtaining multiple human judgements for each data point. These judgements have to be handled with care [3]. In the standard psychology literature, the quality of these annotations is determined via inter-rater agreements, often measured using metrics such as Cohen's Kappa values [4]. If the inter-rater agreement is sufficiently high, a common approach is to use the majority agreement of annotators as the ground truth labels for further analysis (e.g. for classification tasks). However, the removal of no-agreement cases leads to the shrinkage of data sets that are, more often than not, moderately sized due to the existing technological means used to record group interactions with cameras and microphones. Furthermore, relying on majority agreement has its disadvantages. This scheme weights each annotator equally, whereas in reality some annotators might do better than others or may express more confidence than others, due to their experience with the task, interest, or even their personality. Therefore, the main challenge in estimating the "true" label from these annotations is that the expertise of the annotators is not readily quantifiable. Moreover, given that they are all human judgements, the "perfect" ground truth might not be available for even a subset of the data, making validation of the estimated labels impossible.

In this paper, we introduce models that circumvent these problems and have the potential of being applied to many instances of affective and social interaction analysis. We use the knowledge provided by the annotators, the annotations and their confidences in the form of weights, to estimate final class labels and then use these to train classifiers. These annotation weights can be easily procured through an additional question to the annotator. For experimental verification, we apply it to the case of identifying dominant people in small group (i.e. groups of size 3 to 6) conversations [8] from nonverbal cues extracted from audio and video. To show the importance of using annotator weights (in addition to their decisions) in modeling annotators, we benchmark our model in the presence and absence of this information on a recent, publicly available corpus of group interactions (DOME) [2]. Finally, we compare it to the single-step model introduced by Raykar et. al [13]. Although we discuss the results with dominance data, the models can be used in other domains requiring ground truth estimation from multiple annotators.

The paper is organized as follows. Section II presents the related work. The objective of our work is given in Section III. Section IV describes the proposed two stage model along with two other models that are used to show the motivations behind our model. We describe the audio and visual nonverbal cues used to computationally characterize dominance in group interactions in Section V. Results on synthetic data and the DOME corpus are presented in Section VI. Lastly, we conclude with ideas for future work in Section VII.

This work is supported by Swiss NSF project "Sensing and Analysing Organizational Nonverbal Behaviour (SONVB)" and EU FP7 MC IEF project "Automatic Analysis of Group Conversations via Visual Cues in Nonverbal Communication (NOVICOM)".

II. Related Work

Rienks and Heylen [14] were among the first to study dominance in group conversations using computational means. More recently, Jayagopi et al. introduced a set of audio-visual nonverbal features that could be used for detecting the most dominant person in meetings [8] on a subset of the popular Augmented Multi-Party Interaction (AMI) multimodal corpus [5], for which multiple observers provided dominance judgements. This study had several shortcomings. Firstly, only meetings with majority and full agreement annotations were used. This reduced the original data set to a smaller size. Secondly, no analysis was done to detect whether there were more than one dominant person in the meeting or to define ways to detect them. Furthermore, the classifiers used were not able to show complex relationships that could have existed between different audio-visual features. As an extension to this work, Aran and Gatica-Perez [1] presented results on 10 hours of data using the same annotation procedure, but their method had similar limitations mentioned above. They presented the use of rank and score based fusion for multimodal fusion of audio-visual features to infer the most dominant and least dominant person in the meeting. While this method gives an improved performance, it does not directly take into account all the knowledge provided by the annotators and again, no analysis was done to determine the number of dominant participants in a meeting.

While researchers working on computational tasks with subjective annotations have mainly used the Kappa statistic as a measure of annotation reliability [4], and estimated the class labels via majority voting, some studies have attempted to model multiple human judgements to estimate the underlying ”true” label. Smyth et al. [15] proposed an EM based model to estimate the reliability of the annotators on an image labeling task. Raykar et. al. [13] recently introduced a method for modeling annotators to obtain estimates of the true class label while jointly modeling a classifier. Their method jointly modeled a classifier using the features and the annotations. The method was validated against several data sets, with a main focus on datasets involving the detection of tumors. Whitehill et al. proposed another EM-based approach for modeling multiple annotators on an image labeling task, to jointly infer the image label, the expertise of each annotator, and the difficulty of the image [16].

To our knowledge, no such methods have been applied to modeling annotations for social interaction data. Additionally, the single-step approach of jointly modeling the classifier and the annotations [13] does not incorporate the confidence expressed by the annotators. In the area of human interaction analysis, one of the objectives is to find new features that can be used for further analysis (e.g. detecting dominance, personality traits, etc). This means that the introduction of a new feature requires re-modeling the annotators, since there is a close coupling between the input features and the annotator knowledge in this model. Although there are other works in the literature that attempt to first estimate the labels and then learn a classifier [15], their approach to estimate the ground truth does not take the annotator confidences into account. Therefore, our proposed two-step system poses several advantages. Firstly, it allows to include annotation weights into the framework to explicitly model the dependency of the annotations on the “difficulty” of the task, as defined by the annotation weights. Secondly, it allows to train classifiers that use audio-visual features in a separate step, enabling us to compare the results of classifier with a “derived ground truth”. Moreover, the separate classification step gives the flexibility of using any classification method, without a need to directly connect it to the annotator model.

III. The Task

Our objective in this work is to estimate binary dominance level labels of participants in small group conversations using nonverbal audio and visual cues, while at the same time incorporating annotators knowledge into the estimation framework. We use the publicly available DOME corpus [1] [2], which includes dominance annotations on five-minute meeting segments selected from the AMI corpus [5]. Each meeting has four participants, and is recorded with multiple cameras and microphones. We define our dominance estimation task as a binary classification task: Whether a participant in a meeting is dominant or not. However, the models we present in this paper are not limited to a specific group size or to the task of dominance estimation.

The DOME corpus contains 125 meeting segments, each annotated by three annotators. Each annotator filled a questionnaire and ranked the participants between 1 and 4, according to their level of perceived dominance (1 being the most dominant). Besides the dominance ranks, the annotators were also asked to give a dominance weight to each participant. They distributed 10 points among the participants reflecting their impression of relative dominance displayed during the meeting, with more units signifying higher dominance in this setting. Two or more people may have the same weight.

IV. The Model

We first introduce the two stage model (Model I) that we propose for use with social interaction data. Next, we describe a model that does not use annotation weights to estimate the label values (Model II), to study the importance of having additional information from the annotators. Finally, we also present the model introduced by Raykar et al. [13] (Model III) for validating our experiments.

A. Model for Dominance Detection using Annotators' Knowledge (Model I)

For our task, we use a two-stage model. The flowchart of the model is given in Fig. 1. The first stage involves the modeling of annotators (hereafter called the A-model). The A-model is used to obtain the class label estimates, that can be used as ground truth for further analysis. We consider two kinds of information in the annotations: (i) The label

[1] http://www.idiap.ch/scientific_research/resources/dome

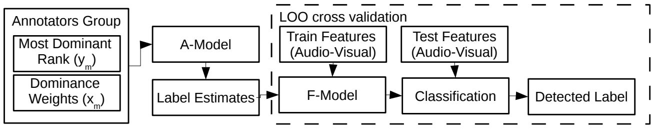

Annotators Group
Most Dominant Rank (y_m)
Dominance Weights (x_m)
A-Model
Label Estimates
LOO cross validation
Train Features (Audio-Visual)
Test Features (Audio-Visual)
F-Model
Classification
Detected Label

Fig. 1: The proposed two-stage model.

indicating the choice of the annotator for the most dominant person (i.e. the person having rank 1), with other participants (having lower ranks) in the meeting being labeled as not dominant; and (ii) the relative weight for the participants in the meeting. For example, we could have the following annotation by three annotators:

$$A = \left[\begin{array}{cccc|cccc|cccc} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right], \quad (1)$$

$$W = \left[\begin{array}{cccc|cccc|cccc} 5 & 3 & 1 & 1 & 4 & 3 & 2 & 1 & 4 & 4 & 1 & 1 \end{array}\right], \quad (2)$$

where A, W are the binary valued annotation ranks and the annotation (dominance) weights, respectively, and the vertical lines partition the annotations by the three annotators.

The A-model provides the estimates of labels, which is then used in the second stage of our model, in order to train a classifier in a supervised manner, using audio-visual features (the F-model).

1) A-Model: Consider R annotators and N data points. Let y be the "true" label and y^j be the label assigned by the j-th annotator. The A-Model is defined in terms of the sensitivity α_j and specificity β_j, defined as:

$$\alpha_j = \text{Pr}[y^j = 1|y = 1], \beta_j = \text{Pr}[y^j = 0|y = 0]. \quad (3)$$

Further, we can model the influence of the annotation weights on the output label as a logistic sigmoid as:

$$\text{Pr}[y = 1|\mathbf{x}, \mathbf{w}] = \sigma(\mathbf{w}^T \mathbf{x}), \quad (4)$$

where $\mathbf{x} \in \mathbb{R}^{R \times 1}$ is the set of dominance weights obtained from all the R annotators and $\mathbf{w}$ is the set of weights for the logistic regression model. The logistic sigmoid function is defined as $\sigma(z) = 1/(1 + e^{-z})$.

We can also set a beta prior on α_j and β_j, and a Gaussian prior on $\mathbf{w}$ to get:

$$\begin{aligned} \text{Pr}[\alpha_j|a_1^j, a_2^j] &= \text{Beta}[\alpha_j|a_1^j a_2^j], & (5) \\ \text{Pr}[\beta_j|b_1^j, b_2^j] &= \text{Beta}[\beta_j|b_1^j b_2^j], & (6) \\ \text{Pr}[\mathbf{w}] &= \mathcal{N}(\mathbf{w}|0, \Gamma^{-1}), & (7) \end{aligned}$$

where $a_1^j, a_2^j, b_1^j, b_2^j$, and Γ are the hyperparameters for α, β, and $\mathbf{w}$ respectively. For our work, we choose the hyperparameters of α and β to be 1, since we have no reason to favor one annotator over the others. Γ was chosen empirically. The MAP estimate of the parameters is obtained as:

$$\hat{\theta}_{MAP} = \arg\max_{\theta}\{\ln \text{ Pr}[\mathcal{D}|\theta] + \ln \text{ Pr}[\theta]\} \quad (8)$$

$$\text{where Pr}[D|\theta] = \prod_{i=1}^{N} \text{Pr}[y_i^1, ..., y_i^R|\mathbf{x}_i, \theta]$$

$$\begin{aligned} = \prod_{i=1}^{N} \text{Pr}[y_i^1, ..., y_i^R|y_i = 1, \alpha]\text{Pr}[y_i = 1|\mathbf{x}_i, \theta] + \quad (9) \\ \text{Pr}[y_i^1, ..., y_i^R|y_i = 0, \beta]\text{Pr}[y_i = 0|\mathbf{x}_i, \theta]. \end{aligned}$$

Here, $\mathcal{D} = \{\mathbf{x}_i, y_1^1 ..., y_i^R\}_{i=1}^N$. The parameter set, $\theta = \{\alpha, \beta, \mathbf{w}\}$, can be estimated using a combination of EM and Newton-Raphson update [13]. This in turn can be used to obtain probabilistic estimates of the labels, $\{\mu_i\}_{i=1}^N$, where

$$\mu_i = \text{Pr}[y_i = 1|y_1^1, ..., y_i^R, \mathbf{x}_i, \theta]. \quad (10)$$

This procedure is similar to the method proposed by Raykar et al. [13], with the difference that annotator confidences are used in place of features for estimating the labels.

2) F-Model: The F-Model uses the estimated label values, $\{\mu_i\}_{i=1}^N$ to train a classifier using the audio-visual features $\{\mathbf{f}_i\}_{i=1}^N$ as inputs. Since we use a separate stage for using the audio visual features, we first threshold the labels into binary class labels (either 0 or 1) and then use a classifier for classification. In our experiments, we use a Support Vector Machine (SVM) with radial basis kernel to achieve this.

B. Model with Annotator Labels Only (Model II)

In order to emphasize the importance of the use of annotation weights estimated by the annotators in the model, we also altered our original model to use only the annotator labels, to get the A-model. Since for our original model, we are choosing a uniform prior for the sensitivity/selectivities, for this model, we can compute the ML estimate instead of the MAP estimate for comparison. We therefore maximize

$$\hat{\theta}_{ML} = \arg\max_{\theta}\{\ln \text{ Pr}[\mathcal{D}|\theta]\}, \quad (11)$$

where $\mathcal{D} = \{y_1^1 ..., y_i^R\}_{i=1}^N$ and $\theta = \{\alpha, \beta\}$ is the parameter set. $\text{Pr}[\mathcal{D}|\theta]$ is defined as:

$$\text{Pr}[\mathcal{D}|\theta] = \prod_{i=1}^{N} \frac{1}{2}(p_i + q_i), \quad (12)$$

$$p_i = \text{Pr}[\{y_i^r\}_{r=1}^R|y_i = 1, \alpha] = \prod_{j=1}^{R} [\alpha^j]^{y_i^j}[1 - \alpha^j]^{1 - y_i^j}, \quad (13)$$

$$q_i = \text{Pr}[\{y_i^r\}_{r=1}^R|y_i = 0, \beta] = \prod_{j=1}^{R} [\beta^j]^{1 - y_i^j}[1 - \beta^j]^{y_i^j}. \quad (14)$$

We get the update equations as follows:

$$\text{E-Step: } \mu_i = \text{Pr}[y_i = 1|\{y_i^r\}_{r=1}^R, \theta] = \frac{p_i}{p_i + q_i} \quad (15)$$

$$\text{M-Step: } \alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}, \beta_j = \frac{\sum_{i=1}^N (1 - \mu_i)(1 - y_i^j)}{\sum_{i=1}^N (1 - \mu_i)} \quad (16)$$

The resulting model is similar to the model proposed in [15].

C. Integrated Single-Step Model (Model III)

As a third model, we model the annotators and audio-visual features jointly, similar to the work of Raykar et al. [13]. The objective of this is to compare the selectivity and sensitivities of annotators as well as the estimated labels with that of our original model.

V. Multimodal Nonverbal Features

Social psychologists have found that dominance is often displayed via audio and visual cues such as speaking time, turns, interruptions, pitch, visual activity, expressions, and gaze [10], [6]. In connection to these nonverbal cues, we extract the following audio and visual features as descriptors of dominance. The exact definitions and the details of the features have been given by other authors (Jayagopi et. al. [8], and Aran and Gatica-Perez [1]).

A. Audio Features

For the audio features, we used recordings from close-talk microphones attached to each participant in the meetings and based on speaker segmentation, we extracted the following features that characterize basic turn taking attributes: speaking length, speaking turns, turns without short utterances, average speaker turn duration, successful interruptions, and speaker floor grabs.

B. Visual Activity Features

We processed the close-up camera video data (4 cameras, one per person) to estimate the total activity of each person in each frame with standard computer vision techniques, such as skin color detection and motion estimation, and extracted the following features: visual activity length, visual activity turns, turns without short movements, average visual activity turn duration, visual activity interruptions, and visual activity floor grabs. These features can be considered analogous to turn taking features in speech.

C. Audio-Visual Features

We also use Audio-Visual (AV) multimodal features, which are defined as the visual activity features of a person while speaking (i.e. all frames when people are silent are not considered). These features are multimodal as both audio and visual modalities are taken into account during the extraction: AV length, AV turns, turns without short movements, average AV turn duration, AV interruptions, and AV floor grabs.

VI. Experimental Setup and Results

We performed our experiments on two datasets: a synthetic dataset and the DOME corpus for dominance estimation. The synthetic dataset is specifically used for validating our model's performance with respect to a golden ground truth, as the golden ground truth does not exist for the dominance estimation task (and often when analysing many other social traits), given that it is based on human judgements.

A. Results on Synthetic Data

To generate the synthetic data, we defined a new task on the DOME corpus as follows: "Determine whether the participant is visually more active than the average group activity". Following the task definition, we use the visual activity length (VL) feature as the basis for generating the ground truth and the ground truth weights. In principle, any other feature could have also been chosen by changing the task definition. To generate the ground truth, the value of VL for each participant was normalized by the total visual activity in the meeting (nF) and participants showing more than 25% of the total activity in the meeting were assigned label $y_i = 1$ as being "visually active":

$$y_i = \begin{cases} 1, \; if \; nF_i > 0.25, \\ 0, \; otherwise. \end{cases} \tag{17}$$

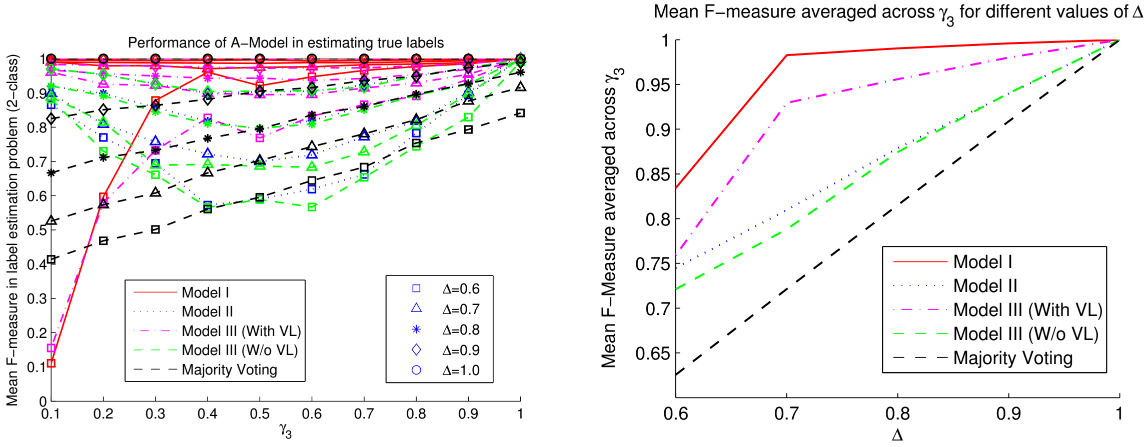
The annotations were then generated for $R = 3$ annotators, from the ground truth by randomly inverting values based on the sensitivity/selectivity of annotators. The weights given by these annotators were computed by adding noise to the ground truth weights, while satisfying two constraints: (i) It should be greater than zero and (ii) Distortion must be more for an annotator with lower value of sensitivity/selectivity. The procedure for doing this is given in Algorithm 1. Note that the annotator weights for this data is real valued as opposed to the discrete valued weights used in real annotations. We simulated three annotators, in order to keep the similarity to the DOME corpus. As the data used to generate the simulation is from the DOME corpus, the amount of data is the same: we have 125 four-person meetings in total.

Algorithm 1 Algorithm to generate artificial annotations

```
Inputs:
M : # of meetings, G : # of participants/meeting
R : # of annotators, $\{(\alpha_j = \beta_j) = \gamma_j\}_{j=1}^{R}$
$GT \in \mathbb{R}^{M \times G}$ : Ground truth
$nF \in \mathbb{R}^{M \times G}$ : Ground truth weights
Output:
$\{\mathcal{Y}_j \in \mathbb{R}^{M \times G}\}_{j=1}^{R}$ : Labels given by annotators
$\{\mathcal{X}_j \in \mathbb{R}^{M \times G}\}_{j=1}^{R}$ : Weights given by annotators
for $j = 1$ to $R$ do
    $\mathcal{Y}_j = GT$
    $\mathcal{X}_j = nF + \min(nF) \times (1 - \gamma_j) * (\text{rnd}(M, G) - 0.5)$
    $T \leftarrow \text{rnd}(M, G) \in \mathbb{R}^{M \times G}$
    for $i = 1$ to $M$ do
        for $k = 1$ to $G$ do
            if $T(i, k) >= \gamma_j$ then
                $\mathcal{Y}_j(i, k) \leftarrow !\mathcal{Y}_j(i, k)$
```

To study the effect of different values for annotator sensitivity and selectivity, we varied these parameters and calculated the accuracy of the estimated labels and the Root Mean Square (RMS) error of the estimated parameters with respect to the ground truth. The estimated labels are calculated using the A-model from Models I and II. For simplicity, we assumed that for each annotator j, the sensitivity and selectivity values are equal ($\alpha_j = \beta_j = \gamma_j$). The results for all our experiments were based on the mean value obtained for 10 independent trials.

For our experiments, we kept two annotators constant (by setting $\gamma_1 = \gamma_2 = \Delta$ and varied the parameter γ_3 for the third annotator. The experiments were repeated for $\Delta = \{0.6, 0.7, 0.8, 0.9, 1\}$. The mean F-measure for the label estimation task was calculated by averaging 10 independent trials for different parameter values. The results are given in Fig. 2(a). It is clear from the figures that Model I outperforms Models II and III in all cases. For a simpler presentation, we also show the mean F-measure averaged across all γ_3 values for a given Δ, for the three models and also the majority



(a)

(b)

Fig. 2: Mean F-measures for estimating true labels with A-models for synthetic data. (a) Mean F-measure for varying γ_3 and Δ values. (b) Mean F-measure averaged across γ_3 for a given Δ.

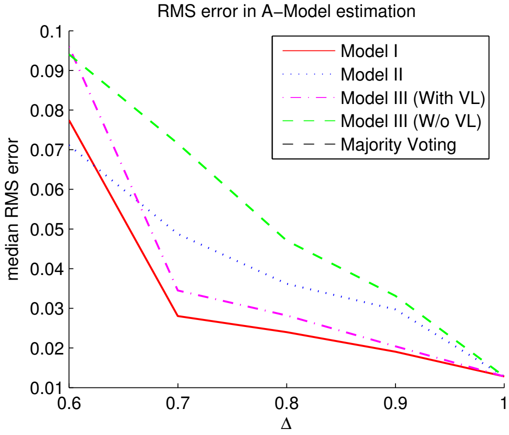


Fig. 3: Median of RMS error in annotator model estimation for different Δ.

voting case, in Fig. 2(b). An interesting result we see is that the exclusion of VL when training Model III (which uses the features) causes the label estimation accuracy to drop below that of Model II, but its inclusion improves it beyond that of Model II. This is an expected result as the ground truth has been derived from the VL feature. In either case Model I outperforms other models, and all models outperform the majority voting case.

Further, we see that the performance of all models dips around $\gamma_3 = 0.5$. This is an interesting result as $\gamma_3 = 0.5$ corresponds to the case in which approximately 50% of the samples are wrongly annotated. This shows that the models have successfully used or discarded information given by the third annotator for $\gamma > 0.5$ and $\gamma < 0.5$ respectively.

Next, we plotted the RMS error in the A-model parameter estimation averaged across all γ_3 values, for a given value of Δ. The RMS error is computed using the equation:

$$e_K^{(RMS)} = \sqrt{\frac{\sum_{j=1}^{R} (\hat{\alpha_j}^{(K)} - \gamma_{GT})^2 + (\hat{\beta_j}^{(K)} - \gamma_{GT})^2}{2 \times R}}, \qquad (18)$$

where K corresponds to either Model I or Model II, $\gamma_{GT} = 1$ is the ground truth specificity/sensitivity, $\hat{\alpha}_j^{(K)}, \hat{\beta}_j^{(K)}$ are the estimated parameters for annotator j for model K, and R is the number of annotators. The results are given in Fig. 3. We see that Model I outperforms both Models II and III. Also, the errors converge for $\Delta = \gamma_3 = 1$. This is expected as the annotations are very clean in this case and the additional information given in the form of annotation weights or other features is not required. Interestingly, the model errors for Model III were greater than that of Model II, even though Model III trained with VL included performed better than model II in terms of label estimation accuracy. Also, the RMS errors for Model III when VL is not used were higher. This suggests that inclusion of features that do not perfectly contain the same information as VL distorts the model estimation process. An important implication of this result is that an integrated model as suggested by Raykar et.al. [13] works better when the features are not noisy. This further justifies the use of our two stage model for the next experiment on the DOME corpus with real annotators.

B. Results on the DOME corpus

The DOME corpus contains 14 groups of annotators and 26 meetings have been annotated on average by each group. For Model I, we chose a beta prior that corresponds to a uniform distribution (in (5) and (6)) as we have no reason to initially favor one annotator over the other. The value of the prior for weights ($\mathbf{\Gamma}^{-1}$) in (7) was varied and $\Gamma = 0.05$ was found to be good. For Models I and II, we trained one A-model per annotator group in the first stage. In the second stage, we used all the estimated labels together to train the classifier that uses audio-visual cues described in Section V. Since our data set size was moderate ($125 \times 4 = 500$ data samples), we used leave-one-out cross validation to get the average performance. A threshold of 0.5 was used to obtain the class labels for training the SVMs in the second stage.

The results are given in Table I. We report the mean F-measure for Model I and II for different feature sets: audio, visual, audio-visual, feature level fusion of audio and video, and feature level fusion of all feature sets. F-Measure was chosen as a performance metric because of the imbalance in

TABLE I: Performance of F-model on audio-visual features

Feature Set	**Mean F-Measure**	
	Model I	Model II
Audio (A)	0.78	0.77
Video (V)	0.71	0.73
Audio-Visual (AV)	0.77	0.77
A+V+AV	0.43	0.45
A + V	0.64	0.68
Baseline	0.40	0.41

the class sizes as accuracy does not take this into account. The baseline reflects the case where a classifier assigns the label of the largest class to each data sample. Interestingly, we found that both models gave us similar performance. In all cases, the models performed better than random.

In order to facilitate further comparisons between these models, we used Model III to obtain label estimates by jointly modeling features and annotations. We only used the audio features for training Model III since they gave the highest performance in our previous experiment. We computed the RMS difference between the A-Model parameters of Model I and Models II and III using:

$$e^{(RMS)}_{K,L} = \sqrt{\frac{\sum_{j=1}^{R}(\hat{\alpha_j}^{(K)}-\hat{\alpha_j}^{(L)})^2+(\hat{\beta_j}^{(K)}-\hat{\beta_j}^{(L)})^2}{2\times R}}, \qquad (19)$$

where K refers to either Model II or Model III, L refers to Model I, and $\hat{\alpha_j}, \hat{\beta_j}$ are the estimated parameters for annotator j and R is the total number of annotators. We computed the median value of the RMS difference (Given in Table II). The RMS difference between models I and II is larger than I and III. The trends in the results tally with that obtained using synthetic data. Table III shows the difference in the estimated labels between the three models. We see that the label estimates of Model II and III are quite similar, and Model I estimates are closer to Model III than Model II.

TABLE II: Median value of RMS difference in annotator model parameters with respect to Model I

	Model II	**Model III**
Median of RMS	0.067	0.033

TABLE III: Difference in estimated labels between different models

	I & II	**I & III**	**II & III**
Difference (%)	4.8	2.8	4.0

VII. Conclusion

In the context of social interaction analysis, we have addressed the open problem of modeling annotators' knowledge to obtain estimates of the ground truth and then training a classifier, in a two stage process. This enables us to use the "derived ground truth" for modeling the data using audio/visual features separately. This is useful in the area of human interaction analysis and more generally in other studies that use annotations from multiple annotators. Also, from our experiments, it is clear that annotation weights, as used here, are quite useful for the estimation of labels.

By using our proposed approach, all the annotated data can be utilized, since there is no requirement for a majority agreement. We also formulated the task as a binary classification task for the dominance of each participant, removing the restriction that there could be only one person showing dominance, that was assumed in previous works [1], [8], [2]. Our best model used audio (A) cues, with a performance measure of 0.78, against a baseline of 0.4.

In the future, we would like to study the relationship between the results obtained for the annotator models and inter-rater agreement values such as Cohen's Kappa value. Further, we would like to learn more parameters that could help us understand the quality of the data and annotators, such as the difficulty of labeling a sample and the dependence of labeling accuracy on characteristics of the data. Further, in this study, we used annotator weights as a representation for annotator confidences. We would like to investigate how different confidence representations can be used to improve label estimates. Applying this method to other forms of social data is another dimension of further investigation.

References

[1] O. Aran and D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in small group conversations," in *Proc. Int. Conf. on Pattern Recognition (ICPR), Istanbul*, 2010.

[2] O. Aran, H. Hung, and D. Gatica-Perez, "A multimodal corpus for studying dominance in small group conversations," in *LREC workshop on Multimodal Corpora (LREC MMC'10), Malta*, 2010.

[3] M. D. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? (pre-print)," *Perspectives on Psychological Science*, In press.

[4] J. Carletta, "Assessing agreement on classification tasks: The Kappa statistic," *Computational Linguistics*, pp. 249–254, 1996.

[5] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Workshop Mach. Learn. for Multimodal Interaction (MLMI'05) Edinburgh, U.K.*, 2005.

[6] J. Hall, E. Coats, and L. LeBeau, "Nonverbal behavior and the vertical dimension of social relations: A meta analysis." *Psychological Bulletin*, p. 1313(6):898–924, 2005.

[7] D. Jayagopi and D. Gatica-Perez, "Discovering group nonverbal conversational patterns with topics," in *Proc. Int. Conf. on Multimodal Interfaces, Beijing*, 2009.

[8] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Trans. on Audio, Speech and Language Processing*, 2009.

[9] S. Kassin, *Psychology*. Prentice-Hall, Inc., 2003.

[10] M. Knapp and J. Hall, *Nonverbal communication in human interaction*. Wadsworth Publication, 7th Ed., 2009.

[11] A. Pentland, *Honest Signals: How They Shape Our World*. The MIT Press, 2008.

[12] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proc. Int. Conf. on Multimodal interfaces, Crete*, 2008.

[13] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, 2010.

[14] R. J. Rienks and D. Heylen, "Automatic dominance detection in meetings using easily detectable features," in *Workshop Mach. Learn. for Multimodal Interaction (MLMI'05), Edinburgh*, 2005.

[15] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in Neural Information Processing Systems (NIPS)*, 1995.

[16] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.