Modeling multiple time series annotations as noisy distortions of the ground truth: An Expectation-Maximization approach

Rahul Gupta, *Member, IEEE,* Kartik Audhkhasi, *Member, IEEE,* Zach Jacokes, *Member, IEEE,* Agata Rozga, *Member, IEEE,* and Shrikanth Narayanan, *Fellow, IEEE*

Abstract—Studies of time-continuous human behavioral phenomena often rely on ratings from multiple annotators. Since the ground truth of the target construct is often latent, the standard practice is to use ad-hoc metrics (such as averaging annotator ratings). Despite being easy to compute, such metrics may not provide accurate representations of the underlying construct. In this paper, we present a novel method for modeling multiple time series annotations over a continuous variable that computes the ground truth by modeling annotator specific distortions. We condition the ground truth on a set of features extracted from the data and further assume that the annotators provide their ratings as modification of the ground truth, with each annotator having specific distortion tendencies. We train the model using an Expectation-Maximization based algorithm and evaluate it on a study involving natural interaction between a child and a psychologist, to predict confidence ratings of the children's smiles. We compare and analyze the model against two baselines where: (i) the ground truth in considered to be framewise mean of ratings from various annotators and, (ii) each annotator is assumed to bear a distinct time delay in annotation and their annotations are aligned before computing the framewise mean.

Index Terms—Time series modeling, Expectation Maximization (EM) algorithm, Multiple annotators, Behavioral signal processing

1 Introduction

TRACKING the evolution of a time series over a continuous variable is a problem of interest in several domains such as social sciences [1], [2], economics [3], [4] and medicine [5], [6]. However, often times the variable of interest may not be directly observable (such as in behavioral time series of psychological states) and judgments from multiple annotators are pooled to estimate the target variable. A classic example is tracking affective dimensions in the study of emotions [7]–[9] where ratings from multiple annotators are used to determine the hidden affective state of a person from audio-visual data of emotional expressions. The general practice in these behavioral domains is to infer the hidden variable by using human annotation. These studies often use heuristic metrics such as mean over the annotator ratings or select annotators based on confidence intervals for the true estimate (the ground truth) of the unobserved variable. However, these metrics may not provide an accurate representation for the ground truth. Apart from assuming a definite relation between the ground truth and the annotator ratings, several factors such as individual differences between the annotators and annotator reliability are not accounted for.

Recent research has addressed a few of these problems. For instance, Nicolaou et al. [10] assume that there is a latent space shared by annotator ratings and identify it using dynamic probabilistic Canonical Correlation Analysis (CCA) model with time warping. Another model proposed by Mariooryad et al. [11] aligns the annotator ratings by adjusting delays identified using mutual information between features and every annotator's ratings. Along the lines of the proposals by Nicolaou et al. [10] and Mariooryad et al. [11], we present a new model which assumes that the ground truth can be computed using a set of low level features based on a "feature mapping function". Furthermore, the annotators process this (latent) ground truth based on annotator specific "distortion functions" to provide their ratings. Our model is inspired from multiple annotator modeling proposed by Raykar et al. [12], and Figure 1 provides an intuitive summary of the model. Similar to Mariooryad et al. [11], our model relies on both annotator ratings as well as features to identify the latent ground truth and is, in fact, a generalization of their model. This design assumption is inspired from the classic channel transfer function estimation in communication theory [13], [14] wherein the channel (annotator) corrupts the true signal based on a transfer function (distortion function). These annotator specific distortion functions, apart from allowing model evaluation on annotator ratings themselves, also provide a window to an annotator's hidden perceptual and cognitive processes.

The proposed model specifically targets the class of problems where the ground truth can not be observed, but judgments from multiple annotators are obtainable/available. We approach this problem using an Expectation Maximization (EM) [15] class of algorithms, a framework widely used under similar circumstances involving an unobserved/hidden variable. We assume specific structures for the feature mapping function and the distortion functions and present an EM algorithm involving iterative execution of an expectation step (E-step) and a maximization step (M-step). The E-step

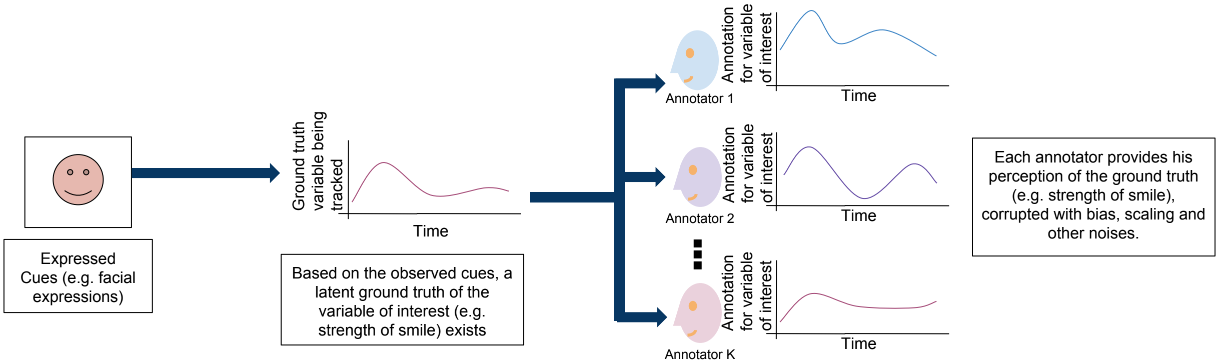
Expressed Cues (e.g. facial expressions)
Ground truth variable being tracked
Time
Based on the observed cues, a latent ground truth of the variable of interest (e.g. strength of smile) exists
Annotation for variable of interest
Annotator 1
Annotator 2
Annotator K
Time
Each annotator provides his perception of the ground truth (e.g. strength of smile), corrupted with bias, scaling and other noises.

Fig. 1. A figure providing the intuition of proposed model, inspired from Raykar et al. [12]

estimates the ground truth based on the values of model parameters at hand and the M-step recomputes the model parameters based on the ground truth obtained in the E-step. We demonstrate the effectiveness of the proposed algorithm in a study involving prediction of time continuous confidence ratings of smile intensity in a video dataset involving toddlers engaging in a brief play interaction with an adult. A set of 28 annotators provide their confidence ratings of the child's smile by looking at a video of the face recorded during the interaction. We present a brief data description and statistics on annotator ratings followed by experimental details of testing various baselines and the proposed model on this dataset. Our results show that our model outperforms baseline models that assume ground truth to be the mean of all annotator ratings as well as the model proposed by Mariooryad et al. [11]. We present our analysis on the distortion functions and compare the structural patterns in the estimated ground truth, annotator ratings and the mean over all annotator ratings. Finally, we also observe the impact of removing a few annotators and record performance changes over each annotator by the proposed and the baseline models.

To summarize, the major contributions of this paper include: (i) designing a system to jointly model time-continuous annotations from multiple annotators (ii) proposing an EM based algorithm to train the system and, (iii) applying and interpreting of the system on a specific case study involving estimating confidence ratings of smile intensity.

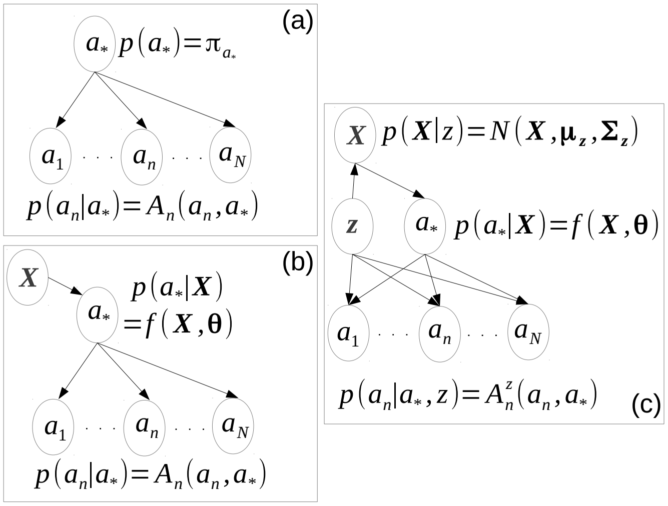
(a)
a_* $p(a_*)=\pi_{a_*}$
a_1 . . . a_n . . . a_N
$p(a_n|a_*)=A_n(a_n,a_*)$
(b)
$\boldsymbol{X}$
a_* $p(a_*|\boldsymbol{X})$ $=f(\boldsymbol{X},\boldsymbol{\theta})$
a_1 . . . a_n . . . a_N
$p(a_n|a_*)=A_n(a_n,a_*)$
$\boldsymbol{X}$ $p(\boldsymbol{X}|z)=N(\boldsymbol{X},\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)$
z
a_* $p(a_*|\boldsymbol{X})=f(\boldsymbol{X},\boldsymbol{\theta})$
a_1 . . . a_n . . . a_N
$p(a_n|a_*,z)=A_n^z(a_n,a_*)$
(c)

Fig. 2. Graphical models for schemes previously proposed to model discrete label problems. (1a) Maximum likelihood estimation of observer error-rates using the EM algorithm [16] (1b) Supervised learning from multiple annotators/experts [12] (1c) Globally variant locally constant model [17].

2 Background

Several previous works have addressed a range of multiple annotator problems involving discrete class labels. Figure 2 shows a few schemes for the discrete class modeling problem, each with a specific set of assumptions. Dawid et al. [16] provided one of the earlier models for the problem as shown in Figure 2(a). a_* represents an unobserved reference label for a given training example, drawn from a probability distribution such that $P(a_*) = \pi_*$. Given a set of N annotators, the n^{th} annotator provide his judgment of the example based on a reliability matrix A_n. Raykar et al. [12] extended the above model to train a discriminative classifier as shown in Figure 2(b). The model first estimates the probability of reference label given a set of features $\boldsymbol{X}$ based on a function $f(\boldsymbol{X}, \boldsymbol{\theta})$ ($\boldsymbol{\theta}$ is the set of function parameters). Each of the annotators provides his/her judgment assuming a similar strategy as the first model. Audhkhasi et al. [17] presented a further modification assuming variable feature reliability as shown in Figure 2(c). The data is assumed to be generated based on the parameter z, which also affects the judgment of each annotator. The probability of a_* is obtained based on the features $\boldsymbol{X}$, through a discriminative maximum entropy model. Similar multiple annotator models have also been proposed by Bachrach et al. [18], Yan et al. [19] and Welinder et al. [20]. However these models have not been generalized to continuous time series annotations, despite covering a range of multiple annotator problems.

Apart from multiple annotator models, other schemes that handle noisy distortion of data include matrix factorization techniques [21], [22], wavelet based methods [23] and other matrix recovery methods [24].

On the other hand, several studies have also focused on modeling time series data. A classic example is modeling emotional dimensions (e.g. valence, dominance, arousal) during human interaction [9], human-computer interaction [8], [25] as well as in music [26], [27]. These studies use multiple annotators to derive the ground truth reference and use heuristic metrics over the annotator ratings as a proxy for the latent emotional dimension. For instance, all the studies listed above use mean over annotator ratings as the ground truth. Other human interaction modeling examples that represent time series of discrete events capturing a hidden internal human state include characterizations of client and counselor behaviors during psychotherapy [6], [28], couples therapy [29] and human-machine spoken dialogs [30]. These studies either substitute ground truth using annotations from a single annotator or use majority voting over multiple annotator ratings at every sample. These approximations of the ground truth are rather crude as they do not account for annotator specific traits such as their proficiency, subjective references as well as motor and cognitive delays in task performance.

Recent research studies have addressed a few of these problems in aggregating annotator ratings using novel methods to account for annotator disparities. For instance, Nicolaou et al. [10] assume that each annotator's ratings could be factored into individual factors and a warped shared latent space representation. They perform this factorization using a Dynamic Probabilistic CCA (DPCCA) model. In later versions of their model [31], they proposed further extensions where features from the data are assumed to be generated conditioned on the latent shared space (Supervised-Generative DPCCA) as well as a discriminative model where the features determine the latent shared space (Supervised-Discriminative DPCCA). In its formulation, the Supervised-Discriminative DPCCA is similar to the proposed model. The model uses CCA and dynamic time warping to address the fact that Raykar's model [12] does not account for temporal correspondences between annotation samples. On the other hand, our model uses a distortion function which operates on the latent ground truth to provide annotator ratings. The distortion function provides proxies for biases and delays estimated for each annotator, which we further interpret in the experiment of our interest (Nicolaou et al. [31] provide other interpretations such as ranking and filtering annotations). Also, Nicolaou et al. [31] evaluate model performance based on how well the features predict the latent ground truth. Although this evaluation is appropriate, the model should also be evaluated on predicting the observed data (i.e., the annotator rating themselves), which is not trivial to obtain using this model. Mariooryad et al. [11] proposed another approach where they first identify annotator specific delays based on mutual information between the annotator ratings and the data stream. The final aggregation is computed as a frame-wise mean of annotator ratings after accounting for delays. Note that this model uses the data feature stream in computing the annotator delays and it is possible to compute (and hence evaluate on) the individual annotator ratings from the ground truth by reintroducing those delays. Our model is an extension to the model proposed by Marioordad et al. [11] wherein instead of only estimating a constant delay, we estimate a more general Finite Impulse Response (FIR) filter which can not only account for delays but also scaling and bias introduction in annotator ratings.

Generally, our work is inspired from the models on discrete class labels and is modified to be applicable on continuous annotations. In the next section, we first describe the general framework for our model. We then describe the data set used for evaluating our model and also discuss the baseline models in comparison to the proposed model. Finally, we interpret the model parameters obtained on the data set and analyze the findings.

3 Distortion based multiple annotator time series modeling

We propose a distortion-based modeling scheme similar in structure to Raykar et al. [12] to model time series annotations from multiple annotators. Given a session s drawn from a set of sessions S, we assume that the ground truth is conditioned on the session features $\boldsymbol{X}^s$. Furthermore the annotator ratings are assumed to be noisy modifications of the hidden ground truth, determined by annotator specific functions. We describe these two assumptions behind our model in detail below.

(i) First, we assume that the ground truth ratings for the session s, $\boldsymbol{a}_*^s = [a_*^s(1), .., a_*^s(t), .., a_*^s(T^s)]^T$ are conditioned on a set of session features $\boldsymbol{X}^s = [\boldsymbol{x}^s(1), .., \boldsymbol{x}^s(t), .., \boldsymbol{x}(T^s)]$. T^s is the number of data frames in s, $a_*^s(t)$ is the ground truth value at the frame index t and $\boldsymbol{x}^s(t)$ is a K-dimensional column feature vector also at the frame index t. $\boldsymbol{a}_*^s$ is a column vector representation of the time series $\{a_*^s(1), .., a_*^s(T^s)\}$. Equation (1) shows the relation between the ground truth time series $\boldsymbol{a}_*^s$ and $\boldsymbol{X}^s$ based on a feature mapping function g. $\boldsymbol{\theta}$ represents the set of mapping parameters for the function g.

$$\boldsymbol{a}_*^s = \mathrm{g}\big(\boldsymbol{X}^s, \boldsymbol{\theta}\big) \tag{1}$$

(ii) Next, we assume that the ratings provided by each annotator are distortions of the ground truth. For the session s, ratings from the n^{th} annotator are represented as a column vector $\boldsymbol{a}_n^s = [a_n^s(1), .., a_n^s(t), .., a_n^s(T^s)]^T$, $a_n^s(t)$ being the rating at the t^{th} frame. We obtain $\boldsymbol{a}_n^s$ based on a distortion function $\mathtt{h}$ operating on $\boldsymbol{a}_*$ as shown in (2). For the n^{th} annotator, $\boldsymbol{D}_n$ represents the set of parameters for $\mathtt{h}$.

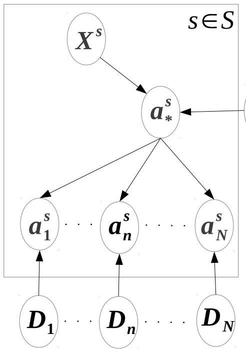

Fig. 3. Graphical model for the proposed framework. $\boldsymbol{X}^s$ represents the features, $\boldsymbol{a}_*^s$ represents the ground truth. $\boldsymbol{\theta}$ and $\langle \boldsymbol{D}_1, .., \boldsymbol{D}_N \rangle$ are the set of parameters for feature mapping function and distortion functions, respectively.

$$\boldsymbol{a}_n^s = \mathtt{h}(\boldsymbol{a}_*^s, \boldsymbol{D}_n);\ n = 1, 2, .., N \tag{2}$$

Figure 3 shows the Bayesian network for the proposed scheme. All session specific variables are located inside the plate. The conditional dependencies (direction of edges) are determined based on the equations (1) and (2). $\boldsymbol{a}_*^s$ can be determined based on $\boldsymbol{\theta}$ and $\boldsymbol{X}^s$, hence the two variables are set to be the parents of $\boldsymbol{a}_*^s$. Similarly, $\boldsymbol{D}_n$ and $\boldsymbol{a}_*^s$ are parents of $\boldsymbol{a}_n^s$.

3.1 Choices for the feature mapping function and the distortion function

In this work, we chose linear functions with additive noise terms as the representations for the functions g and h. Linear representations lead to better interpretability and easier parameter learning but the model can be extended to more complicated representations. The additive noise terms account for factors that can not be captured by linear modeling and is a commonly used component in various regression and classifier learning schemes [32]. We describe our choices in detail below.

Feature mapping function: We choose a linear mapping between the features $\boldsymbol{X}^s$ and $\boldsymbol{a}_*^s$ as shown below.

$$\boldsymbol{a}_*^s = \mathtt{g}\big(\boldsymbol{X}^s, \boldsymbol{\theta}\big) = \begin{bmatrix} \boldsymbol{X}^s \\ \mathbf{1} \end{bmatrix}^T \boldsymbol{\theta} + \boldsymbol{\psi}^s \tag{3}$$

In the equation above, $\boldsymbol{\theta}$ is a $K+1$ dimensional vector, $\boldsymbol{\psi}^s = [\psi^s(1), .., \psi^s(t), .., \psi^s(T^s)]^T$ is a random noise vector with noise variable $\psi^s(t)$ added at the t^{th} frame. $\mathbf{1}$ represents a vector of ones and appends a bias term to feature vector at each frame. In effect, ground truth at frame t, $a_*^s(t)$ is obtained from (3) as

$$a_*^s(t) = \begin{bmatrix} \boldsymbol{x}^s(t) \\ 1 \end{bmatrix}^T \boldsymbol{\theta} + \psi^s(t) \tag{4}$$

We assume the noise vector $\boldsymbol{\psi}^s \sim \mathcal{N}(\mathbf{0}, \sigma_\psi \times \boldsymbol{I}_{T^s})^1$. Given the affine transformation in (3), $\boldsymbol{a}_*^s$ follows the distribution given by

$$\boldsymbol{a}_*^s \sim \mathcal{N}\Bigg(\begin{bmatrix} \boldsymbol{X}^s \\ \mathbf{1} \end{bmatrix}^T \boldsymbol{\theta}, \sigma_\psi \times \boldsymbol{I}_{T^s} \Bigg) \tag{5}$$

Similar assumptions on noise distribution are made in several regression and classification models [33], [34]. The Gaussian noise distribution allows for easy computation, however, can be replaced with other noise distributions as done is several previous works [35], [36].

Distortion function: An annotator may modify the ground truth based on his/her perception. We aim to capture this annotator specific modification using a distortion function operating on the ground truth. We assume that the n^{th} annotator's ratings $\boldsymbol{a}_n^s$ for the session s are obtained after distorting the ground truth based on a linear time invariant (LTI) filter with additive bias and noise terms. Although a linear operation, LTI filters can account for scaling and time delays introduced by the annotators. We assume a filter of length W with coefficients $\boldsymbol{d}_n = [d_n(0), .., d_n(W-1)]$ along with an additive bias term d_n^b. The noise random vector is represented by $\boldsymbol{\phi}_n^s = [\phi_n(1), .., \phi_n(t), .., \phi_n(T^s)]^T$ where $\phi_n(t)$ is noise random variable for t^{th} frame. The set of parameters $\mathbf{D_n}$ for the distortion function h as represented in (2) are the filter coefficients $\boldsymbol{d}_1, .., \boldsymbol{d}_N$ and the bias terms $d_1^b, .., d_N^b$. Based on the filter coefficients, the bias term and the noise vector, $\boldsymbol{a}_n^s$ is given as shown in (6).

$$\boldsymbol{a}_n^s = \mathtt{h}(\boldsymbol{a}_*^s, \boldsymbol{d}_n) = (\boldsymbol{d}_n * \boldsymbol{a}_*^s) + (d_n^b \times \mathbf{1}^s) + \boldsymbol{\phi}_n^s \tag{6}$$

In (6), $\mathbf{1}^s$ represents a vector of ones with as many entries as the number of frames in the session s. The operator $*$ represents the convolution operation between the time series a_*^s and annotator specific filters $\boldsymbol{d}_n$. Further, we assume $\boldsymbol{\phi}_n$ to be a zero mean Gaussian noise with a covariance matrix of the form $(\sigma_\phi \times \boldsymbol{I}_{T^s})$, where $\boldsymbol{I}_{T^s}$ represents an identity matrix with dimensions (T^s, T^s). Since $\boldsymbol{\phi}_n \sim \mathcal{N}(\mathbf{0}, \sigma_\phi \times \boldsymbol{I}_{T^s})$, we can state the following given the affine transformation in (6)

$$p(\boldsymbol{a}_n^s | \boldsymbol{a}_*^s, \boldsymbol{d}_n) \sim \mathcal{N}\Big((\boldsymbol{d}_n * \boldsymbol{a}_*^s) + (d_n^b \times \mathbf{1}^s), \sigma_\phi \times \boldsymbol{I}_{T^s} \Big) \tag{7}$$

4 Training Methodology

We use data log-likelihood maximization technique for training the proposed model. Based on the definitions of the functions h and g, we maximize the likelihood of the observed data (i.e., the annotator ratings) to obtain the parameters $\boldsymbol{d}_n, d_n^b$ and $\boldsymbol{\theta}$. Also note that in the

[1] We use the notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ to represent a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\sigma}$. In $\mathcal{N}(\mathbf{0}, \sigma_\psi \times \boldsymbol{I}_{T^s})$, $\mathbf{0}$ represents a zero mean vector and $\sigma_\psi \times \boldsymbol{I}_{T^s}$ is a diagonal covariance matrix with all entries equal to σ_ψ. In this case, the operator × implies multiplication of a scalar value to all entries of a matrix/vector.

multiple annotator experiments under consideration, the ground truth $\boldsymbol{a}_*^s$ is not directly observable. Therefore the Expectation-Maximization (EM) algorithm [15] is a suitable candidate for maximum likelihood estimation. The data log-likelihood $\mathcal{L}$ is defined on the observed annotator ratings $\langle \boldsymbol{a}_1^s, .., \boldsymbol{a}_N^s \rangle$ given the feature values $\boldsymbol{X}^s$ and model parameters $\boldsymbol{\Pi} = \langle \boldsymbol{d}_1, .., \boldsymbol{d}_N, d_1^b, .., d_N^b, \boldsymbol{\theta} \rangle$ over all the sessions $s \in S$ as shown below.

$$\mathcal{L} = \sum_{s \in S} \log p(\boldsymbol{a}_1^s, .., \boldsymbol{a}_N^s | \boldsymbol{\Pi}, \boldsymbol{X}^s) \tag{8}$$

The above expression is equivalent to the marginalized log-likelihood over the hidden ground truth variable $\boldsymbol{a}_*^s$ as given below.

$$\mathcal{L} = \sum_{s \in S} \log \int_{\boldsymbol{a}_*^s} p(\boldsymbol{a}_1^s, .., \boldsymbol{a}_N^s, \boldsymbol{a}_*^s | \boldsymbol{\Pi}, \boldsymbol{X}^s) \, \partial \boldsymbol{a}_*^s \tag{9}$$

A complete derivation of the EM algorithm for the model in Figure 3 based on the structural assumptions for the distortion and feature mapping functions is given in Appendix 1. Below, we briefly summarize the model training using the EM algorithm and the criteria to evaluate the model.

4.1 EM algorithm implementation

• **Initialize** filter coefficients $\langle \boldsymbol{d}_1, ..., \boldsymbol{d}_N \rangle$, bias terms $\langle d_1^b, .., d_N^b \rangle$ and mapping function parameter $\boldsymbol{\theta}$.
• **While** the data-log likelihood converges, perform:

- *E-step*: In this step, we obtain the ground truth estimate $\bar{\boldsymbol{a}}_*^s$. Based on the Gaussian distribution functions defined in (5) and (7), we arrive at the optimization problem shown in (10). $||.||_2$ represents the L_2 vector norm.

$$\begin{aligned} \bar{\boldsymbol{a}}_*^s = \arg\min_{\boldsymbol{a}_*^s} \sum_{n=1}^{N} \big|\big|(\boldsymbol{a}_n^s) - (\boldsymbol{d}_n * \boldsymbol{a}_*^s + d_n^b \times \mathbf{1}^s)\big|\big|_2^2 \\ + \big|\big|(\boldsymbol{a}_*^s) - \begin{bmatrix} \boldsymbol{X}^s \\ \mathbf{1} \end{bmatrix} \boldsymbol{\theta} \big|\big|_2^2; \;\; \forall s \in S \end{aligned} \tag{10}$$

- *M-step*: In the M-step, we estimate the model parameters based on the Gaussian distribution functions defined in (5) and (7). A detailed derivation of this estimation is shown in Appendix 1 and it turns out that we can estimate filter coefficients $\langle \boldsymbol{d}_1, .., \boldsymbol{d}_N \rangle$, the bias terms $\langle d_1^b, .., d_N^b \rangle$ and parameter $\boldsymbol{\theta}$ by operating separately on the two constituent terms. The optimization problem to obtain the distortion function parameters is given below.

$$\boldsymbol{d}_n, d_n^b = \arg\min_{\boldsymbol{d}_n, d_n^b} \sum_{s \in S} \sum_{n=1}^{N} \big|\big|(\boldsymbol{a}_n^s) - (\boldsymbol{d}_n * \bar{\boldsymbol{a}_*^s} + d_n^b \times \mathbf{1}^s)\big|\big|_2^2 \tag{11}$$

The above optimization to obtain $\boldsymbol{d}_n$ and d_n^b can be carried out jointly by using a matrix formulation. Optimization problem to obtain $\boldsymbol{\theta}$ is stated below.

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \sum_{s \in S} \big|\big|(\bar{\boldsymbol{a}_*^s}) - \begin{bmatrix} \boldsymbol{X}^s \\ \mathbf{1} \end{bmatrix} \boldsymbol{\theta} \big|\big|_2^2 \tag{12}$$

• **End while**

In the next section, we describe our evaluation criteria on a given test set after training the model using the EM algorithm.

4.2 Evaluation criteria

We chose two evaluation criteria for our model: (i) accuracy of the feature mapping function in predicting the ground truth, and (ii) accuracy in prediction of annotator ratings themselves. We discuss these two criteria below.

4.2.1 Eval1: Accuracy of the feature mapping function in predicting the ground truth

In our first criterion, we estimate the latent ground truth $a_*^{\hat{s},\text{true}}$ for a test session $\hat{s}$ using the annotator ratings only based on the optimization problem stated in (13). Then, we make ground truth predictions $a_*^{\hat{s},\text{pred}}$ from the feature mapping function as shown in (14). The Eval1 criterion is given as the correlation between the estimated ($a_*^{\hat{s},\text{true}}$) and predicted ($a_*^{\hat{s},\text{pred}}$) ground truths. This evaluation criterion was also adopted by Nicolaou et al. [10] where they compute the ground truth based on annotator ratings and use features to predict the estimated ground truth. They motivate this evaluation criteria by arguing that a better ground truth can be better predicted using the low level features. Similarly, Mariooryad et al. [11] first compute the ground truth after accounting for lags from annotator ratings and later use features from the data to predict sufficient statistics of the estimated ground truth such as its mean.

$$\boldsymbol{a}_*^{\hat{s},\text{true}} = \arg\min_{\boldsymbol{a}_*^{\hat{s}}} \sum_{n=1}^{N} \big|\big|(\boldsymbol{a}_n^{\hat{s}}) - (\boldsymbol{d}_n * \boldsymbol{a}_*^{\hat{s}} + d_n^b \times \mathbf{1}^s)\big|\big|_2^2 \tag{13}$$

$$\boldsymbol{a}_*^{\hat{s},\text{pred}} = \begin{bmatrix} \boldsymbol{X}^{\hat{s}} \\ \mathbf{1} \end{bmatrix} \boldsymbol{\theta} \tag{14}$$

4.2.2 Eval2: Accuracy in predicting the annotator ratings

Since the ground truth is a latent variable in the problems of interest, we also evaluate our model directly on the observed data, i.e., the annotator ratings themselves. An accurate prediction of observed ratings would imply that the model is able to capture the inherent relationship between the features, ground truth and annotator ratings. We report the correlation coefficient (ρ) between the true and predicted ratings per annotator which also allows for observing the performance for each annotator separately. The annotator ratings are obtained using the following two steps: (i) we first predict the ground truth $a_*^{\hat{s}}$ on a test session $\hat{s}$ using the feature mapping function as stated in (14) (ii) next, we compute $\boldsymbol{a}_1^{\hat{s}}(t), .., \boldsymbol{a}_N^{\hat{s}}(t)$ from $\boldsymbol{a}_*^{\hat{s}}$ and $\boldsymbol{d}_1, .., \boldsymbol{d}_N$ using the operation shown below.

$$\boldsymbol{a}_n^{\hat{s}} = \boldsymbol{d}_n * \boldsymbol{a}_*^{\hat{s}} + d_n^b \times \mathbf{1}^s \quad (15)$$

Note that these estimates of $\boldsymbol{a}_*^{\hat{s}}$ and $\boldsymbol{a}_n^{\hat{s}}$ are the means of Gaussian probability distribution functions stated in (5) and (7), hence also the maximum likelihood estimates. In the next section, we describe the experimental evaluation and our dataset of choice.

5 Experimental evaluation

We evaluate the proposed framework on ten sessions of a dyadic child-clinician interaction dataset, the Rapid-ABC dataset [37]–[39] focusing on perceived ratings of the strength of a child's smile. The data were collected to computationally investigate behavioral markers of psychological and cognitive health conditions such as Autism Spectrum Disorders; the patterns of smiles are hypothesized to be an important cue [40]. Each session is approximately three minutes long and involves natural interaction between an adult and a child between the ages of 15 and 30 months. The interaction elicits verbal as well as non-verbal behaviors (e.g, smile, laughter, grins). The overarching goal of this data collection was to understand various aspects of child-adult interaction including social response, joint attention and child engagement.

For the purpose of our study, a set of 28 annotators later independently viewed a video from each session that captured the child's face during the interaction. They provided ratings on the strength of a child's smile (using a joystick arrangement), recorded at a frame rate of 30 samples/second over a dynamic range of 0-500. The corresponding audio included both psychologist and child speech. The annotators underwent an extensive initial training in rating the smile confidences. During this training, the annotators would rate a file and their ratings were discussed with the data collectors (third and fourth authors of this paper). The discussion points included disagreements with the data collectors and other annotators, the offset and onset of smile confidence annotations and other factors such as the annotator's consistency. After multiple rounds of this training procedure, they were assigned the 10 sessions used in this study to code by themselves with no feedback. We show the inter-rater agreements using the correlation coefficient (ρ) between every pair of annotators as the metric in Figure 4. These ρ values are computed over frames from all the 10 sessions. The annotator indices are assigned based on agreement with the first annotator; where the last index is assigned to the annotator having least agreement with annotator 1.

From the figure, we observe that the ρ values are in the range of 0.35 to 0.80 for most of the annotator pairs. However the ρ values of annotator 27 and 28 with other annotators are particularly low. This is indicative of a lower quality of ratings from these two annotators. Therefore, apart from initially testing our models by including all the annotators, we also conduct a follow up evaluation after removing these two annotators and analyze the results. Evaluation including annotators 27 and 28 helps us to interpret their impact on the model by analyzing the parameters corresponding to these annotators. On the other hand, evaluation without annotators 27 and 28 provides an insight into the impact of removing noisy annotator on the predictive capability of the model. In order to evaluate our model, we perform a 10 fold cross-validation, where 8 sessions are used for training, 1 as development set and 1 for testing. In the next section, we describe the features $\boldsymbol{X}^s$ used in this work.

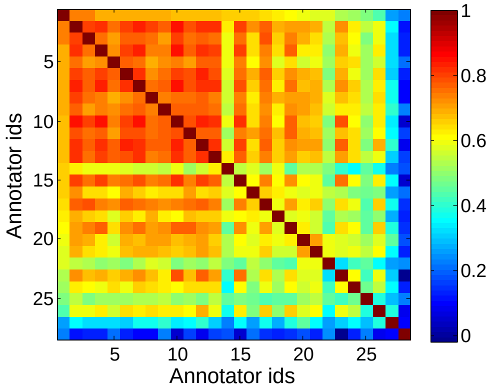

Fig. 4. Correlation coefficient (ρ) between every pair of annotators represented as an image matrix. Colorbar on the right indicates the value of the correlation coefficient. Due to indexing based on agreement with annotator 1, annotators with lower indices have a higher ρ with annotator 1. Annotator 27 and 28 have a very low agreement with several of the annotators.

5.1 Feature set

Smile is a visual phenomenon and previous research has used several visual features for smile detection [41], [42] and analysis [43]. We use a set of similar video based features in our study. The video features are computed per video frame (30 frames/second) and are synchronized with the annotator ratings. We describe the features below.

Facial landmarks: We use the CSIRO Face analysis SDK [44] to track facial landmarks on the child's face. We fit 66 landmark points to the face at every frame. Figure 5 shows a video frame from the database with landmark points marked on the face. Based on these landmark points, we compute two sets of features: (1) velocity of the head based on the nose-tip landmark point, and (2) distance and velocity of all other landmark points with respect to the nose tip landmark point.

Local binary patterns (LBP) based features: LBP features [45] are well known for describing facial expressions. During the computation of this feature, every pixel's intensity is compared to its neighbors and a

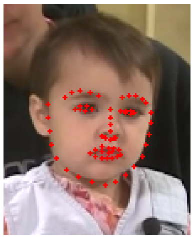

Fig. 5. Facial landmark points tracked on the children's face during interaction.

binary vector is returned. LBP descriptor is a histogram over these binary patterns.

We combine the facial landmark features and the LBP features to obtain a feature vector with dimensionality $K = 387$ for every video frame. For more details on the features, please refer to [44], [45]. In the next section, we provide a description of the baseline models.

5.2 Baseline models

We use two baseline models to compare against the proposed model. In the first baseline model the ground truth is assumed to be a frame-wise mean over all the annotator ratings and the second baseline is borrowed from the work by Mariooryad et al. [11]. We discuss these baselines below.

5.2.1 Baseline 1: Frame-wise mean of annotator ratings

We use a baseline model, where the ground truth at a given frame is assumed to be the mean over ratings from all the annotators at that frame. Several previous works [8], [9], [46] have used this assumption in obtaining the ground truth from multiple annotators on similar time series modeling problems. This scheme assigns equal weight to each annotator and does not account for individual differences. In the baseline case, the relation between the ground truth and the annotator ratings is presumed before hand and can be represented by the following operation in (16). $\boldsymbol{I}_{T^s}$ represents an identity matrix of dimensionality (T^s, T^s).

$$\boldsymbol{a}_*^s = \frac{1}{N} \underbrace{\left[\boldsymbol{I}_{T^s} | \boldsymbol{I}_{T^s} | \ldots | \boldsymbol{I}_{T^s}\right]}_{\text{N-times}} \begin{bmatrix} \boldsymbol{a}_1^s \\ \vdots \\ \boldsymbol{a}_N^s \end{bmatrix} \tag{16}$$

We incorporate the assumption in (16) in the framework of our model. We obtain the mapping parameter $\boldsymbol{\theta}$ based on $\boldsymbol{a}_*^s$ (obtained as in (16)) using the MMSE criteria in (12). However, instead of obtaining filter coefficients using EM algorithm, they have to be computed based on equation (16). We use two different methods to compute the filter coefficients using the hard coded ground truth $\boldsymbol{a}_*^s$ from (16) as listed below.

Baseline 1(a): In the first baseline model, the filter coefficients are computed using the MoorePenrose pseudoinverse (Pinv) [47] operation on the set of identity matrices in (16) as shown in (17). As per (17), the multiplication of $\boldsymbol{I}_{T^s}$ to a_*^s to obtain a_n^s implies that the filters are inferred to be unit impulse response filters with no delay. Hence the filter coefficient $\boldsymbol{d}_n$ is a unit Kronecker delta function. The bias terms d_n^b are all estimated to be 0.

$$\begin{bmatrix} \boldsymbol{a}_1^s \\ \vdots \\ \boldsymbol{a}_N^s \end{bmatrix} = \text{Pinv}\Big(\frac{1}{N} \underbrace{\left[\boldsymbol{I}_{T^s} | \boldsymbol{I}_{T^s} | \ldots | \boldsymbol{I}_{T^s}\right]}_{\text{N-times}}\Big) \boldsymbol{a}_*^s = \begin{bmatrix} \boldsymbol{I}_{T^s} \\ \vdots \\ \boldsymbol{I}_{T^s} \end{bmatrix} \boldsymbol{a}_*^s \tag{17}$$

Baseline 1(b): In this case, we set $\boldsymbol{a}_*^s$ to the value shown in (16). Then, we compute the filter coefficients $\boldsymbol{d}_n$ and the bias terms d_n^b using the MMSE criteria listed in (11). The filter length parameter W is tuned on the development set.

5.2.2 Baseline 2: Lag compensated aggregation of annotator ratings

Our second baseline is borrowed from the work by Mariooryad et al. [11] where we first estimate the lags per annotator with respect to the features obtained from the data stream. The lags per annotator are computed by introducing a delay in the ratings per annotator till his/her ratings have the maximum mutual information with the frame-wise features. Note that this formulation is a special case of the proposed model when the distortion function is constrained to be a unit impulse response filter with a constant delay ($\boldsymbol{d}_n$ in (6) is set to a Kronecker delta function with the delay corresponding to the n^{th} annotator). The bias terms d_n^b are set to 0 in this formulation. After compensating for the annotator delays calculated on the training set, a_*^s for every data partition is computed as the frame-wise mean of the aligned annotator ratings (also the solution to the optimization in (13)). We obtain the mapping parameter θ from the computed $\boldsymbol{a}_*^s$ using the MMSE criteria in (12). In order to compute back the annotator ratings for the Eval2 criterion, individual annotator ratings on the test set are computed as per the convolution stated in (15). In essence, the convolution operation reintroduces the estimated delays in the ground truth to compute each annotator's ratings. For more details regarding this baseline, please refer to section 4 in [11].

5.3 Results

Using the stated cross validation split, we train the baseline and proposed models. For the proposed model and the baseline model 1(b), the filter length parameter W is tuned on the development set. Note that W is tuned globally over all the annotators, as tuning a W for each annotator is computationally expensive and the filter characteristics are expected to be robust to small changes in the length W. Table 1 shows the correlation coefficient ρ of feature mapping prediction with the estimated ground truth (Eval1 criterion). Note that results are the same for baselines 1(a) and 1(b) due to the common ground truth computation criteria, i.e., frame-wise means of annotator ratings. The Eval1

	Baseline 1a/1b	Baseline 2	Proposed Model
Eval1 criteria, correlation coefficient with the ground truth	0.28	0.30	0.34

TABLE 1
Correlation coefficient ρ between the estimated ground truth and the predictions from the feature mapping function. A higher ρ implies that the estimated ground truth is better estimated using the low lever features. The improvement over the closest baseline using the proposed model is significant based on the Fisher z-transformation test [48] (p-value $<$ .001, z-value = 6.1, number of samples equals the number of analysis frames: $\sim$37k).

criterion correlation of the proposed model is better than the baseline using the Fisher z-transformation test [48] considering value at each frame to be a sample. Figure 6 shows the ρ in predicting the observed annotator ratings (Eval2 criterion). For the Eval2 criterion, the proposed model is significantly better than all the baselines for 20 annotators (Fisher z-transformation test, p-value $<$ 10%, number of samples is the number of analysis frames: $\sim$37k). This excludes the noisy annotators 27 and 28 as observed in Figure 4. The Cohen's D [49] comparing the proposed model against each baseline yields a values of .31 (baseline 1a), .11 (baseline 1b) and .33 (baseline 2). The Cohen's D is computed using correlation coefficients for each annotator as the sample values. These values indicate a small improvement effect over baseline 1b and medium improvement effect over baselines 1a and 2.

5.4 Discussion

The performance results in Table 1 are in the expected order. The naive baseline of computing the ground truth as frame-wise mean of the annotator ratings could not be well modeled by the features at hand and thus performs the worst. Adjusting for annotator specific delays and then aggregating the annotator rating performs better than baselines 1(a) and 1(b). However factors such as differences in annotator biases, range of annotation and context in annotation can not be modeled by imposing a constant delay assumption on the distortion functions. These factors are accounted for in the proposed model by allowing the distortion function to be an LTI filter, thereby providing the best performance. For the Eval2 criterion in predicting the annotator ratings, the proposed model performs the best for most of the annotators. Performance is particularly low in predicting the ratings for annotator 27 and 28. This indicates that these annotators are noisy and hard to model, an observation consistent with the inter-rater correlations shown in Figure 4. The performances of baseline 1(a) and 2 are comparable for the Eval2 criterion. This stems from the fact that the distortion functions for both these baselines are constrained to be unit response filters (with additional delay allowed for baseline 2), and thus carry low modeling strength in predicting back the annotator ratings. Baseline 1(b) still allows for the distortion function to be an LTI filter which can account for a longer temporal context in predicting annotator ratings from the ground truth (even though the ground truth is a naive frame-wise mean of annotator ratings). In the following section, we make a few more observations regarding the model parameters, the inferred ground truth and effect of removing a few annotators. We note that the interpretation of these parameters only offers a window to the complex cognitive factors.

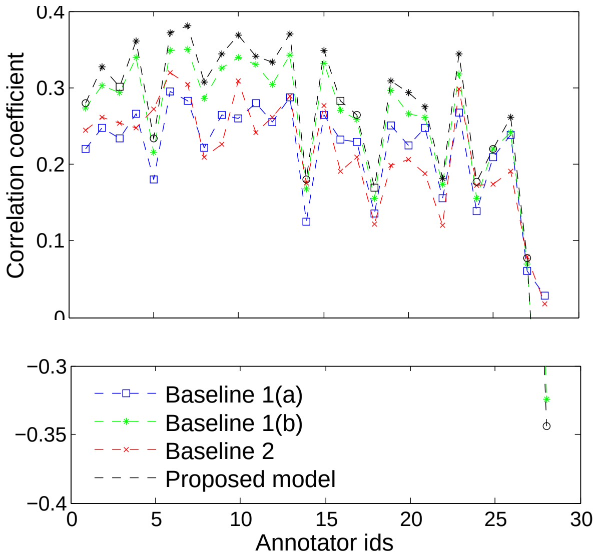

Fig. 6. Correlation coefficients ρ between the true and predicted annotator ratings. A higher ρ implies that the model is better able to model the dependencies between low level features and the annotator ratings. The ρ values of proposed model significantly better (at least at 5% level using Fisher z-transformation test) than all the baseline are marked with $*$. Annotators 3,16 and 18 are significant only at 10% level (marked with a $\Box$) and annotator 1, 5, 14, 17, 24, 25, 17 and 28 are either not significantly better or worse than at least one of the baselines.

5.4.1 *Interpreting the distortion function parameters*

In this section, we plot and interpret various parameters of the distortion function. Figure 7 shows the LTI filter coefficient values for the 28 annotators, obtained using model training over all the 10 sessions. The bias term in the filter is shown as a stem plot in Figure 8. From the filter coefficients in Figure 7, we can make several observations to compare an annotator with others. For instance, the filter coefficients of Annotator 1 are such that the a_*^s samples in the past are weighted higher in convolution to obtain a_1^s. The opposite is true for annotator 6 as a_*^s samples closer to the current frame carry higher weight than the samples in the past. A phase delay analysis of filters from these two annotators suggests that the filter from annotator 1 introduces a greater delay in the ratings than that of annotator 6. Another observation is that the filter coefficients for

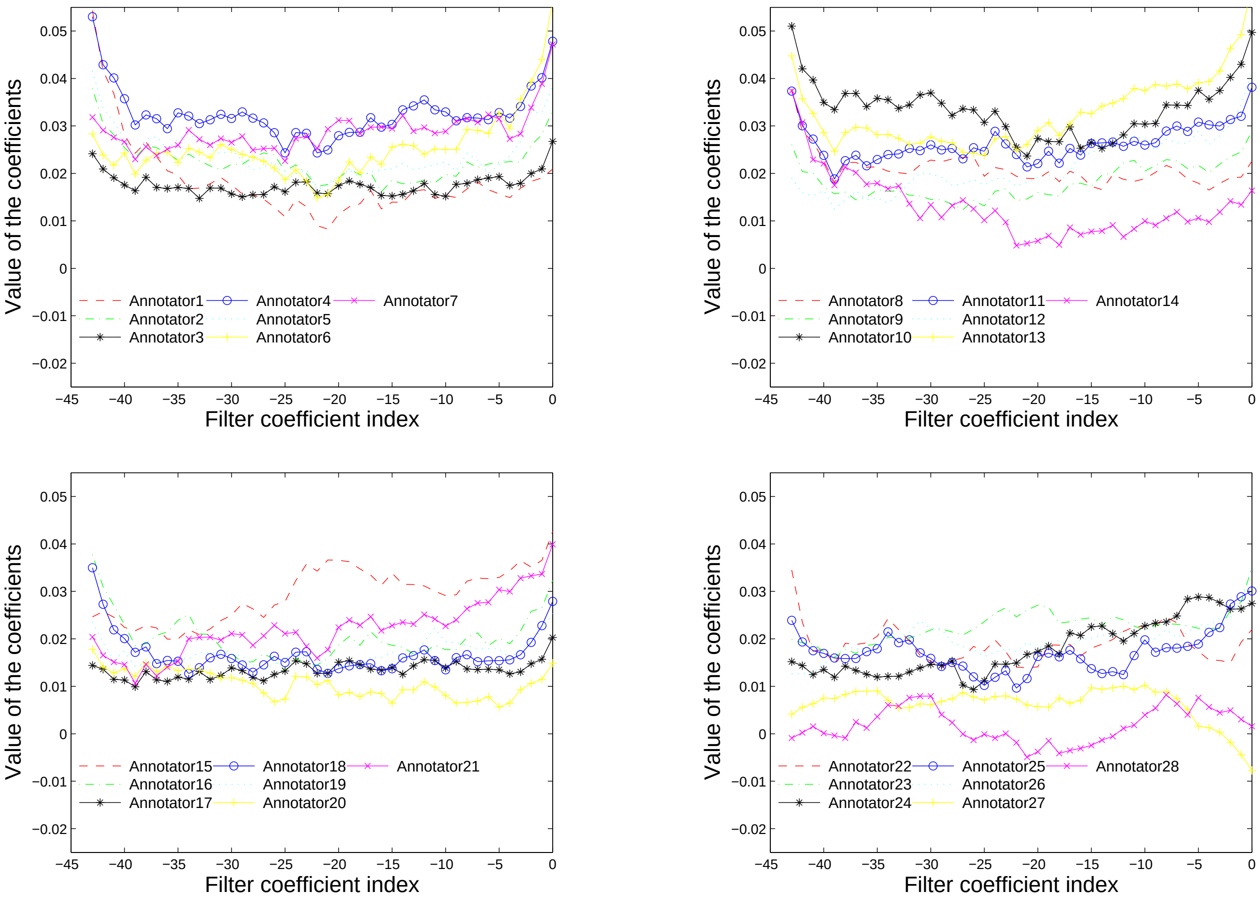

Fig. 7. Filter coefficients estimated by the proposed EM algorithm for each of the annotators. The filter are plotted as $d(-(W-1)),..,d(-1),d(0)$ used during convolution as: $a_n^s(t) = \sum_{w=0}^{W-1} a_*^s(t-w) \times d_n(-w)$. A higher value for the coefficients towards the left in the figure implies a higher emphasis on the past samples.

annotator 27 and 28 have lower absolute values. Thus, the ground truth ratings are attenuated to obtain annotations for the annotator. On the other hand, ratings for annotator 15 is obtained after amplification of the ground truth. Overall, the shape of LTI filter co-efficients varies across annotators (e.g. annotators 4 and 10 have a U-shaped filter and annotator 17 and 20 have a more flat filter shape). We note that these filters coefficients are obtained in a data driven fashion and their phase and magnitude responses provide an ad-hoc quantification of the complex annotation behavior.

From the bias terms shown in Figure 8, we observe that annotator 14 and 28 have a high positive annotation bias term and annotator 10 has a high negative bias term. These terms are added to the ground truth to obtain the respective annotator ratings. The group of annotators 6, 7, 11 and 24 have a relatively low bias term. We also plot the annotator delays estimated using the baseline 2 in Figure 9. Annotators 1, 14, 18 and 28 are estimated to have the longest delays. This observation is fairly consistent with the filter coefficient estimates shown in Figure 7, where the filter coefficients in the past are estimated to carry higher value thereby introducing a larger phase delay.

We note that interpretation of these parameters only offers a window to the complex cognitive factors during annotation. The parameters of annotator bias, delay and distortion are estimates obtained as per the model assumptions. They are further influenced by other factors such as the overall interaction dynamics between the child and the psychologist as well as other latent annotator states (such as their mood and the environment). These factors are not accounted for by our model and can be the subject of a future study.

5.4.2 *Inferred ground truth from the annotator ratings*

We compare the estimated ground truth for an arbitrary segment of the data, from the various baselines and the proposed model in Figure 10. As expected, we observe that the ground truth estimate from baseline 2 has a phase lead over that estimated from the baseline 1(a)/(b) (compare the peaks in the plot). For the proposed model, a lead is again observed when compared to baseline 1(a)/(b), but not as large as baseline 2. Also, the dynamic range for the segment is higher for the baseline estimated from the proposed model.

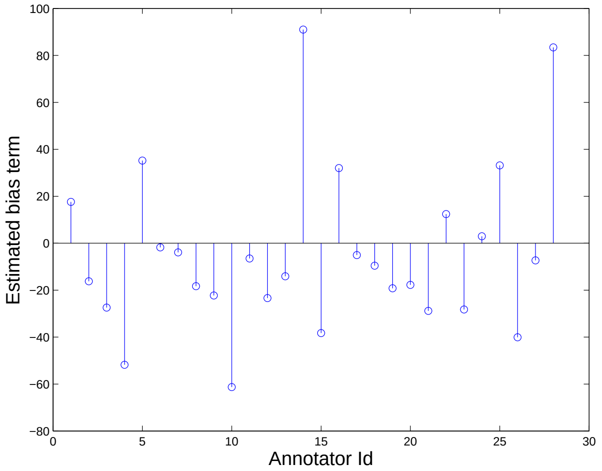

Fig. 8. Annotator bias d_n^b estimated using the proposed model.

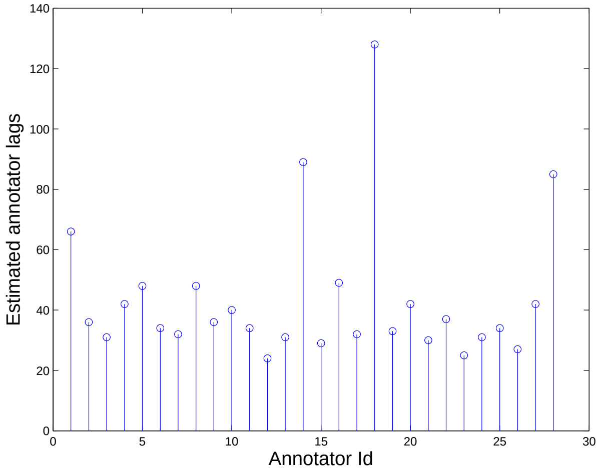

Fig. 9. Annotator delays estimated using the baseline 2 proposed in Mariooryad et al. [11].

This results from the capability of the proposed model to be able to account for annotator bias as well as amplifying/attenuating their ratings, as discussed in the previous sections. Furthermore, high frequency components in the features get added during the ground truth computation using the proposed model (equation (33)). The features are otherwise not used during framewise aggregation in the baseline models.

5.4.3 Performance after removing annotators 27 and 28

Finally, we observe the impact on the performance of the model after removing annotators 27 and 28. We observed that annotators 27 and 28 had the lowest inter-rater correlation with annotators in the Figure 4. We remove these annotators during model training and testing. The correlation coefficient ρ of feature mapping prediction with the estimated ground truth (Eval1 criterion) is shown in Table 2. From the results for Eval1 criterion in Table 2, we observe that the performances of all the models are better after removing annotators 27 and 28. Also, the increase in absolute performance is the highest for the proposed model. This indicates that the ground truth estimation in case of the proposed model benefits the most after removing noisy annotators.

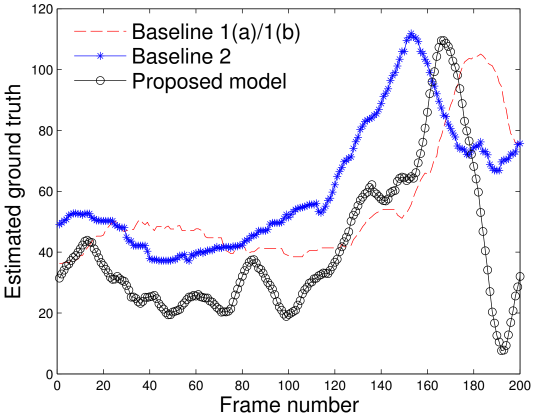

Fig. 10. Ground truth $\boldsymbol{a}_*^s$ as estimated by various baseline and proposed models on an arbitrary section of the data.

Eval1 criteria,	Baseline	Baseline	Proposed
correlation with	1a/1b	2	Model
the ground truth	0.29	0.31	0.36

TABLE 2
Correlation coefficient ρ between the estimated ground truth and the predictions from the feature mapping function after removing annotators 27 and 28 from training. The proposed model is significantly better than the closest baseline model (baseline 2) based on the Fisher z-transformation test [48], considering value at each frame to be a sample (p-value $<$ 0.001, z-value = 7.7).

Figure 11 show the ρ between the predicted and true annotator ratings (Eval2 criterion). After removing the annotators 27 and 28, the proposed model performs significantly better than all the other baselines for 21 out of 26 annotators (at p-value $<$ 10% level). In this experiment, we obtain Cohen's D values of .95, .30 and .97 when comparing the correlation coefficient samples obtained using the proposed method against baseline 1a, 1b and 2, respectively. This indicates a medium improvement effect size over baseline 1b and strong improvement effect sizes over baselines 1a and 2. The improvement in Cohen's D is primarily obtained due to discounting of annotators 27 and 28, which otherwise lead to an increase in standard deviation of obtained correlation coefficients, as presented in Figure 6. Note that the annotators 27 and 28 were poorly modeled by the proposed model as seen in Figure 6 and therefore their removal helps the proposed model in both, estimating a better ground truth as well as modeling other annotators better.

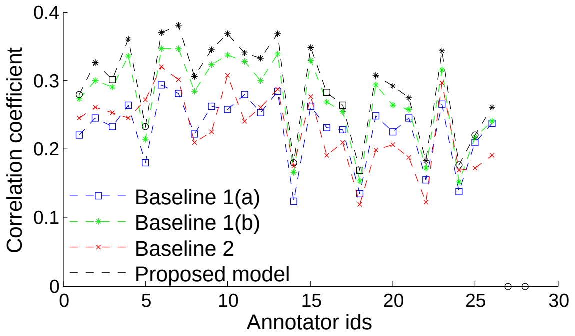

Fig. 11. Correlation coefficients ρ between the true and predicted annotator ratings based on model trained after removing annotators 27 and 28. A higher ρ implies that the model is better able to model the dependencies between low level features and the annotator ratings. For the correlation coefficients obtained using proposed model, $*$ indicates a significant improvement with p-value $<$ 5%, □ indicates significant improvement with p-value $<$ 10% but greater than 5%.

6 Conclusion

Several studies employ multiple annotators to model time series over a continuous hidden (unobserved) variable. The ground truth is often substituted by heuristic measures over the available ratings, which are later used for training and evaluating the model. In this work, we present a novel scheme to model the ratings from multiple annotators using an EM algorithm. Our algorithm infers the hidden ground truth based on a feature mapping function and learns a distortion function for each annotator. This distortion function is used by the annotator to provide his perception of the ground truth. Evaluation on smile confidence ratings from 28 annotators on the Rapid-ABC dataset demonstrates that the proposed model outperforms the baseline cases that substitute ground truth by computing means over annotator ratings or only compensate for delays in the annotator ratings. We further analyze the model parameters and identify annotator specific traits such as annotator bias and delay.

Our model can be further improved by using schemes similar to those proposed in multiple annotator modeling problems over discrete labels [17]–[19]. In this work, we have assumed a specific structure for the feature mapping and distortion functions but other formulations can be tested. The distortion functions from each annotator can also be investigated to study factors such as annotator similarity and reliability. Similarly investigations on feature mapping functions may reveal features best suited for the study. Furthermore, as we pointed out previously, there are several other complex factors that determine factors such as annotator bias and delay (e.g. interaction dynamics in the dyadic conversation, environmental settings). Our model does not account for such factors and they can be a subject for future studies to further understand the dynamics of annotation. Finally, this study may be extended to cases involving multidimensional time series, involving joint modeling over each dimension.

Acknowledgment

The authors would like to thank NSF for their support.

References

[1] Richard McCleary, Richard A Hay, Errol E Meidinger, and David McDowall, *Applied time series analysis for the social sciences*, Sage Publications Beverly Hills, CA, 1980.

[2] Dean K Simonton, *"Sociocultural context of individual creativity: a transhistorical time-series analysis.," Journal of personality and social psychology*, vol. 32, no. 6, pp. 1119, 1975.

[3] Marianne Baxter and Robert G King, *"Measuring business cycles: approximate band-pass filters for economic time series," Review of economics and statistics*, vol. 81, no. 4, pp. 575–593, 1999.

[4] Stephen J Taylor, *"Modelling financial time series,"* 2007.

[5] Niels K Rathlev, John Chessare, Jonathan Olshaker, Dan Obendorfer, Supriya D Mehta, Todd Rothenhaus, Steven Crespo, Brendan Magauran, Kathy Davidson, Richard Shemin, et al., *"Time series analysis of variables associated with daily mean emergency department length of stay," Annals of emergency medicine*, vol. 49, no. 3, pp. 265–271, 2007.

[6] Rahul Gupta, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan, "Predicting clients inclination towards target behavior change in motivational interviewing and investigating the role of laughter," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

[8] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic, *"Avec 2014–3d dimensional affect and depression recognition challenge,"* 2013.

[9] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth S. Narayanan, *"Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, Feb. 2013.

[10] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic, "Dynamic probabilistic cca for analysis of affective behaviour," in *Computer Vision–ECCV 2012*, pp. 98–111. Springer, 2012.

[11] Soroosh Mariooryad and Carlos Busso, *"Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," Affective Computing, IEEE Transactions on*, vol. 6, no. 2, pp. 97–108, 2015.

[12] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy, *"Learning from crowds," The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[13] Dimitris G Manolakis, Vinay K Ingle, and Stephen M Kogon, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, vol. 46, Artech House Norwood, 2005.

[14] Bernard Widrow and Samuel D Stearns, *"Adaptive signal processing," Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985, 491 p.*, 1985.

[15] Arthur P Dempster, Nan M Laird, and Donald B Rubin, *"Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society. Series B (Methodological)*, 1977.

[16] Alexander Philip Dawid and Allan M Skene, *"Maximum likelihood estimation of observer error-rates using the em algorithm," Applied statistics*, pp. 20–28, 1979.

[17] Kartik Audhkhasi and Shrikanth Narayanan, *"A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels," Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 769–783, 2013.

[18] Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver, *"How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing," arXiv preprint arXiv:1206.6386*, 2012.

[19] Yan Yan, Rómer Rosales, Glenn Fung, Mark W Schmidt, Gerardo H Valadez, Luca Bogoni, Linda Moy, and Jennifer G Dy, "Modeling annotator expertise: Learning when everybody knows a bit of something," in *International conference on artificial intelligence and statistics*, 2010, pp. 932–939.

[20] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie, "The multidimensional wisdom of crowds," in *Advances in neural information processing systems*, 2010, pp. 2424–2432.

[21] Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Yan Yan, *"l_{1}-norm low-rank matrix factorization by variational bayesian method," Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 4, pp. 825–839, 2015.

[22] Anders Eriksson and Anton Van Den Hengel, *"Efficient computation of robust low-rank matrix approximations in the presence of missing data using the l 1 norm,"* 2010.

[23] Rongfeng Zhang and Tadeusz J Ulrych, *"Physical wavelet frame denoising," Geophysics*, vol. 68, no. 1, pp. 225–231, 2003.

[24] Zhouchen Lin, Minming Chen, and Yi Ma, *"The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," arXiv preprint arXiv:1009.5055*, 2010.

[25] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran, *"Representation of facial expression categories in continuous arousal-valence space: Feature and correlation," Image and Vision Computing*, 2014.

[26] Mohammad Soleymani, Anna Aljanaki, Yi-Hsuan Yang, Michael N Caro, Florian Eyben, Konstantin Markov, Björnn Schuller, Remco Veltkamp, and Frans W Felix Weninger, "Emotional analysis of music: A comparison of methods," in *Proceedings of ACM International Conference on Multimedia-MM 2014, November 3-7, Orlando, Florida, USA*, 2014.

[27] Naveen Kumar, Rahul Gupta, Tanaya Guha, Colin Vaz, Maarten Van Segbroeck, Jangwon Kim, and Shrikanth Narayanan, "Affective feature design and predicting continuous affective dimensions from music," in *MediaEval 2014 Multimedia Benchmark Workshop, Barcelona*, 2014.

[28] Dogan Can, Panayiotis Georgiou, David Atkins, and Shrikanth S. Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features," in *Proceedings of InterSpeech*, Sept. 2012.

[29] Matthew Black, Athanasios Katsamanis, Chi-Chun Lee, Adam Lammert, Brian Baucom, Andrew Christensen, Panayiotis Georgiou, and Shrikanth S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *In Proceedings of InterSpeech*, Makuhari, Japan, Sept. 2010.

[30] Jackson Liscombe, Giuseppe Riccardi, and Dilek Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[31] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic, *"Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1299–1311, 2014.

[32] Christopher M Bishop et al., *Pattern recognition and machine learning*, vol. 1, springer New York, 2006.

[33] John Neter, William Wasserman, and Michael H Kutner, *"Applied linear regression models,"* 1989.

[34] Mário AT Figueiredo, "Lecture notes on bayesian estimation and classification," 2004.

[35] Makoto Yamada and Masashi Sugiyama, "Dependence minimizing regression with model selection for non-linear causal inference under non-gaussian noise.," in *AAAI*, 2010.

[36] Chong Gu, *"Adaptive spline smoothing in non-gaussian regression models," Journal of the American Statistical Association*, vol. 85, no. 411, pp. 801–807, 1990.

[37] Opel Y. Ousley, Rosa I. Arriaga, Michael J. Morrier, Jennifer B. Mathys, Monica D. Allen, and Gregory D. Abowd, *"Beyond parental report: Findings from the rapid-abc, a new 4-minute interactive autism," Technical report series: report number 100 (http://www.cbi.gatech.edu/techreports), Center for Behavior Imaging, Georgia Institute of Technology*, 2013.

[38] Rahul Gupta, Chi-Chun Lee, Lee Sungbok, and Shrikanth Narayanan, "Assessment of a child's engagement using sequence model based features," in *Workshop on Affective Social Speech Signals, Grenoble*, 2013.

[39] Rahul Gupta, Daniel Bone, Sungbok Lee, and Shrikanth Narayanan, *"Analysis of engagement behavior in children during dyadic interactions using prosodic cues," Computer Speech & Language*, vol. 37, pp. 47–66, 2016.

[40] Daniel Messinger and Alan Fogel, *"The interactive development of social smiling," Advances in child development and behaviour*, vol. 35, pp. 328–366, 2007.

[41] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan, *"Toward practical smile detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, 2009.

[42] Yu-Hao Huang and Chiou-Shann Fuh, "Face detection and smile detection," in *Proceedings of IPPR Conference on Computer Vision, Graphics and Image Porcessing, Shitou, Taiwan, A5-6*, 2009, p. 108.

[43] Thibaud Sénéchal, Jay Turcot, and Rana El Kaliouby, "Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[44] M Cox, J Nuevo-Chiquero, JM Saragih, and S Lucey, *"Csiro face analysis sdk," Brisbane, Australia*, 2013.

[45] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, *"Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[46] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[47] Arthur Albert, *Regression and the Moore-Penrose pseudoinverse*, Elsevier, 1972.

[48] Govind S Mudholkar, *"Fisher's z-transformation," Encyclopedia of Statistical Sciences*, 1983.

[49] Marnie E Rice and Grant T Harris, *"Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r.," Law and human behavior*, vol. 29, no. 5, pp. 615, 2005.

[50] Sam Roweis and Zoubin Ghahramani, *"A unifying review of linear gaussian models," Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[51] James M Joyce, "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science*, pp. 720–722. Springer, 2011.

[52] Alfredo HS Ang and Wilson H Tang, *"Probability concepts in engineering," Planning*, vol. 1, no. 4, pp. 1–3, 2004.

[53] Michael Kearns, Yishay Mansour, and Andrew Y Ng, "An information-theoretic analysis of hard and soft assignment methods for clustering," in *Learning in graphical models*. Springer, 1998.

[54] Fred Jelinek, *"Speech recognition by statistical methods," Proceedings of the IEEE*, vol. 64, pp. 532–556, 1976.

[55] Daphne Koller and Nir Friedman, *Probabilistic graphical models: principles and techniques*, MIT press, 2009.

Rahul Gupta received a B.Tech. degree in Electrical Engineering from Indian Institute of Technology, Kharagpur in 2010 and a Ph.D. degree in Electrical Engineering from University of Southern California (USC), Los Angeles in 2016. His research concerns development of machine learning algorithms with application to human behavioral data. His dissertation work is on the development of computational methods for modeling non-verbal communication in human interaction. He is the recipient of Info-USA exchange scholarship (2009), Provost fellowship (2010-2014) and the Phi Beta Kappa alumni in Southern California scholarship (2015). He was part of the team that won the INTERSPEECH-2013 and INTERSPEECH-2015 Computational Paralinguistics Challenges. He is a member of the IEEE.

Kartik Audhkhasi received a B.Tech. degree in Electrical Engineering and a M.Tech. degree in Information and Communication Technology from Indian Institute of Technology, Delhi in 2008. He received a Ph.D. degree in Electrical Engineering from University of Southern California (USC), Los Angeles in 2014. He is currently a Research Staff Member in the Watson Multimodal Group at IBM Watson. His research focuses on automatic speech recognition, natural language processing, and machine learning. He was the recipient of the Annenberg and IBM Ph.D. fellowships. He was part of the team that won the INTERSPEECH-2013 Computational Paralinguistics Challenge. He also received best paper and teaching assistant awards from the Electrical Engineering Department at USC and was a 2012/13 Ming Hsieh Institute Ph.D. Scholar. He is a member of the IEEE.

Zachary Jacokes received his Bachelor's degree in Psychology from Emory University. His research interests include Autism Spectrum Disorders and traumatic brain injuries, specifically regarding their effect on brain volumetrics and neural connectivity. He is also interested in virtual and augmented reality and their potential applications in treating Autism Spectrum Disorders (ASDs). He is currently a data scientist and programmer at the University of Southern California's Lab of Neuro Imaging, where he manages the database of ASD subjects for a project funded by the Autism Centers of Excellence.

Agata Rozga received a BA in Psychology from the University of California, Berkeley and an MA and PhD in Developmental Psychology from the University of California, Los Angeles. She completed a postdoctoral fellowship through the Center for Behavior Neuroscience at Georgia State University. She is currently a Senior Research Scientist in the School of Interactive Computing at Georgia Institute of Technology, where she directs the Child Study Lab. A developmental psychologist and autism researcher by training, she collaborates with computer scientists to develop novel computational tools and methods to objectively measure behaviors relevant to studying typical and atypical development. Using these tools, she aims to shed new light on early disruptions in social-communicative development in autism.

Shrikanth (Shri) Narayanan is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics, Psychology, Neuroscience and Pediatrics and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems and applications with direct societal relevance. [http://sail.usc.edu]

Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is editor in chief for IEEE Journal of Selected Topics in Signal Processing, an editor for the Computer Speech and Language Journal and an associate editor for the IEEE Transactions on Affective Computing, APSIPA Transactions on Signal and Information Processing and the Journal of the Acoustical Society of America. He was also previously an associate editor of the IEEE Transactions of Speech and Audio Processing (2000-2004), IEEE Signal Processing Magazine (2005-2008), IEEE Transactions on Multimedia (2008-2011) and the IEEE Transactions on Signal and Information Processing Over Networks (2014-2015). He is a recipient of a number of honors including Best Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 20102011 and ISCA Distinguished Lecturer for 2015-2016. Papers co-authored with his students have won awards including the 2014 Ten-year Technical Impact award from ACM ICMI, Best Student Paper award at ICASSP-2016, Interspeech 2015 Nativeness Detection Challenge, Interspeech 2014 Cognitive Load Challenge, Interspeech 2013 Social Signal Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2009 Emotion Challenge, and other awards at IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005 and ICSLP 2002. He has published over 700 papers and has been granted seventeen U.S. patents.