Modeling emotion understanding over time in complex narratives

Desmond C. Ong, *Member, IEEE Computer Society*, Zhengxuan Wu,
Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki

Abstract—Human emotion unfolds and evolves dynamically over time, but only a minority of affective computing research focuses on time-series understanding—for example, recognizing emotions at every moment of an interaction. Affective dynamics have been difficult to capture in affect computing, in part due to the inherent difficulties in modelling time-series data and of collecting high-quality time-series datasets. Here, we address this gap. We first provide a review of different time-series approaches which have been, or can be, productively applied to affective computing, including deep neural network and generative models, and models that can handle asynchronous event-based input. Second, we introduce the first version of the Stanford Emotional Narratives Dataset (SENDv1), a rich video dataset of self-paced emotional narratives, annotated for emotional valence over time, and designed for time-series modelling of naturalistic behavior. We apply several time-series modelling approaches to our dataset: a Long Short-Term Memory model, a Multimodal Variational Recurrent Neural Network, as well as an event-based Recurrent Marked Temporal Point Process model. These models were applied to predict emotional valence as a function of three modalities: visual (facial expressions), acoustic (pitch and other paralinguistic features) and linguistic (i.e., the semantic content of the story). We discuss comparisons between the different approaches, and end by discussing their potential for time-series applications. We hope that this paper highlights the challenges—as well as several state-of-the-art solutions—for time-series modelling in affective computing.

Index Terms—Multimodal Emotion Recognition, Time-Series, Naturalistic Stimuli, Deep Learning, Probabilistic Modelling, Event-based models

1 Introduction

Emotions are an integral part of our everyday lives. They are also fundamentally dynamic. John wakes up feeling sad that he has to get out of his warm bed; he feels happy when he checks his phone and receives a nice email; this turns to anger when he realizes his roommate had clogged the bathroom. Our emotions evolve dynamically over time, and are situated in the context of the day's events and our history of prior experiences.

We are entering an age where we will soon share our homes, hospitals, and offices with artificial agents. In order for these artificial agents to successful co-exist with people, they will have to seamlessly understand people's thoughts and emotions. In particular, these agents will have to learn models of human emotion based on observations of situated, naturalistic human behavior [1], [2]. The field of affective computing has made exciting progress in this direction, with a large body of research focusing on recognizing emotions from faces [3], paralinguistics (e.g., pitch, prosody) [4], body gestures [5], language [6], as well as actively integrating different modalities into making *multimodal* judgments [7], [8].

- *D. C. Ong and Z-X. Tan are with the A*STAR Artificial Intelligence Initiative, Agency for Science, Technology and Research, Singapore 138632. E-mail: desmond.c.ong@gmail.com*
- *Z. Wu is with the Department of Management Science and Engineering, Stanford University.*
- *M. Reddan is with the Department of Psychology, University of Colorado, Boulder.*
- *I. Kahhale and J. Zaki is with the Department of Psychology, Stanford University.*
- *A. Mattek is with the Department of Psychology, University of Oregon.*

Manuscript received XXXXXX XX, 2019; revised XXXXXX XX, 2019.

However, the vast majority of such work in affective computing is still unable to capture the *dynamics* of emotion as they unfold over time and as they change in response to different events: These models are often unable to handle **time-series emotion recognition**. Specifically, we define time-series modelling as taking in temporally continuous input data and producing temporally continuous output, with an explicit consideration of how information is propagated over time. A social robot at home having a conversation with its user would have to take in a continuous stream of sensor data, process them, and reason about their user's emotions over time, perhaps after every second or after every sentence, as well as across the conversation [9].

Time-series modelling is challenging. For one, we need to model how emotions at one time-point depend on the emotions at previous time-points: Is this a linear or non-linear relationship, and how far back does this temporal dependence stretch? How does this influence decay over time? Psychologically, this refers to how "sticky" emotional states are: How long do they tend to last [10]? How do emotional states transition into other emotional states [11]? Second, we need to deal with the problem of potentially asynchronous data. In a multimodal emotion recognition setting, visual input may be sampled at 30Hz or 60Hz, acoustic signals and psychophysiological signals are sampled at several thousand Hz, and conversational content may only vary on the order of seconds. How should we integrate information across all these different time-scales [12], [13]? Third, real world settings often come with missing data: Perhaps a particular sensor stops measuring data for certain periods of time (or more commonly, a participant's face moves out of the view of the camera). As it turns

out, most models cannot easily handle missing data, which occurs frequently in continuous, time-series data.

What does the field need to move towards time-series models in affective computing? We suggest that the biggest barriers are due to the inherent difficulty of building computational time-series models, and the difficulty of collecting high-quality datasets. To address this first gap, we conduct a review covering different machine-learning-based approaches to time-series modelling (Section 2). We begin by discussing the most commonly used time-series techniques in affective computing: deep neural network models, part of a broader class of *discriminative* models. We also cover two other classes of time-series approaches—*generative* approaches and *event-based* approaches—which are comparatively less common within affective computing, but offer interesting modelling capabilities that hold exciting potential for understanding emotions.

We turn next to discuss the second gap: Researchers need high-quality time-series datasets on which to train models. These are expensive to construct, in terms of both the production of stimuli and the collection of time-series annotations of emotion and affective labeling [14]. There are several existing time-series datasets that have been used by the affective computing community, mostly through the Audiovisual Emotion Challenges, a series of challenges held annually since 2011 [15]. AVEC is a large and collaborative multi-institutional effort that involves collating, curating, and releasing datasets, and has catalyzed much of the research in time-series affective computing. In every AVEC challenge to date, part of the challenge requirements involves producing time-series labels on a common dataset. The first two challenges [15], [16] had researchers predict valence over time on the SEMAINE dataset [17]. The SEMAINE dataset consists of recordings of volunteers interacting with a "Sensitive Artificial Listener", an artificial agent programmed to respond in emotional stereotypes (e.g., happy and outgoing, or angry and confrontational [15]). The third and fourth AVEC challenge [18], [19] asked for predictions of valence and arousal (and dominance in the fourth challenge) on the AViD-Corpus, a series of recordings of volunteers performing several tasks such as reading aloud excerpts from storybooks and describing the story behind a given picture (as in the Thematic Appreciation Test). The fifth and sixth challenge [20], [21] involved predicting valence and arousal on the RECOLA database [22], which included pairs of individuals collaborating on a task via remote conferencing. Finally, the seventh and eighth challenge [23], [24] required predictions of valence, arousal and likability ratings on the Sentiment Analysis in the Wild (SEWA) dataset, which also involved dyads discussing their views on a commercial that both individuals had viewed. The SEWA dataset was also collected "in the wild" using participants' personal webcams rather than in a controlled lab environment, unlike the previously-mentioned datasets. More recent challenges that involve predicting emotions or empathy over time include the 2018 OMG-Emotion [25] and the 2018 Affect-in-the-Wild challenge [26], which were collections of YouTube videos of spontaneous emotion displays, and the 2019 OMG-Empathy challenge, which had videos of a research volunteer listening to a confederate recount scripted emotional stories.

There is a constant demand for high-quality time-series datasets, especially for training models for different affective computing applications. Indeed, the diversity of datasets we mentioned does not cover the range of social interactions that affective computing researchers are interested in[1]. In Section 3, we introduce our contribution towards providing emotionally-rich time-series data: the Stanford Emotional Narratives Dataset, version 1 (SENDv1), an annotated video dataset of unscripted autobiographical emotional narratives, and in Section 4 we report the results of several time-series modelling approaches on this dataset.

From this point onwards, we choose not to use "continuous" to describe the time-series nature of the models or data. This is to avoid confusion with another potential meaning of "continuous", which is to produce graded or dimensional outputs [9]. That is, instead of producing an emotion classification (e.g. *happy* vs. *sad* vs. *neutral*) or a binary judgment (e.g. high or low valence), such models would predict a real-valued judgment on some interval or ordinal scale [27]. We will stress here that the choice of a dimensionally-continuous output is an orthogonal modelling decision from dealing with temporally-continuous data, and hence we will not use "continuous" to avoid ambiguity.

In the rest of this paper, we provide a review of time-series modelling, with a focus on affective computing (Section 2). We then introduce a novel naturalistic multimodal dataset consisting of unscripted emotional life stories (Section 3). In Section 4, we describe implementations of three time-series approaches to modelling this dataset, and discuss the results in light of the modelling assumptions. Finally, we end with a discussion of how to extend the ideas discussed in this paper, such as to building personalized and longitudinal models (i.e., affective computers that interact with an individual over many sessions, potentially over a lifespan).

2 Time-Series Models

In this section, we provide an overview of time-series approaches applicable to affective computing. We focus on three classes of models: (1) discriminative, (2) generative, and (3) event-based models such as point process models. For reasons of space, we do not cover linear models, such as autoregressive or moving average models traditionally used in econometrics and other fields: rather, we focus on machine learning models that are more amenable to complex input data. The goal of this section is to highlight some of the modelling assumptions underlying these popular approaches as well as the relations between the approaches.

Let us begin with some notational definitions. Let X_t^k be the vector of input features for sequence k at time t—this could be a vector of say, facial expression features or vectors of multimodal features—and Y_t^k be the corresponding vector of outputs at time t, such as categorical labels of emotion classes or real-valued scores or probabilities. We use $X_{t_1:t_2}^k$ and $Y_{t_1:t_2}^k$ to denote a series of these inputs and outputs from times t_1 to t_2, inclusive. Given n paired training sequences

1. Ideally, we would need to build models that generalize across different datasets, i.e., transfer learning. But that itself is technically challenging, especially for time-series modelling, and out of the scope of this paper.

$\{(X^k_{1:T_k}, Y^k_{1:T_k}),\ 1 \leq k \leq n\}$ where T_k is the final time point of sequence k, the goal is to train a model that can predict the sequence of outputs $Y^j_{1:T_j}$ given a new input sequence $X^j_{1:T_j}$, for some $j > n$. Without loss of generality, this new predicted sequence could also be an extension of a previously-observed sequence.

2.1 Discriminative Models

Given a set of emotion outputs Y_t and a set of input features X_t, one approach is to directly model how we can predict the output labels from the input features. Such *discriminative* models [28] are widely used in machine learning for both classification (e.g., predicting an emotion category) and regression problems (e.g., predicting a real-valued number). Linear and logistic regression, the Support Vector Machine/Support Vector Regression [29], random forest classifiers [30] and deep neural networks like Convolutional Neural Networks [31], are amongst the most popular discriminative machine-learning models applied within (non-time-series) affective computing [7], [8].

A vanilla feed-forward neural network transforms inputs X into outputs Y via nonlinear transformations through intermediate, hidden layer(s) h. The most straightforward way to extend feed-forward neural networks to model time-series data is to allow the hidden layer at one time point to influence the hidden layer at subsequent time points. Adding such a "recurrency" between hidden states results in an architecture known as the Recurrent Neural Network [32], shown in Fig. 1a. A RNN is a neural network in which hidden states at time h_t depends on the input features at that time X_t and the hidden state at the previous time-point h_{t-1}, via some function f with parameters θ. The hidden states subsequently predict the outputs via g with parameters ϕ:

$$\begin{aligned} h_t &= f_\theta\left(X_t, h_{t-1}\right) \\ Y_t &= g_\phi\left(h_t\right) \end{aligned} \tag{1}$$

In common parameterizations, f_θ and g_ϕ return a linear combination of their arguments filtered through a nonlinear activation function (e.g., the hyperbolic tangent, the sigmoid, the softmax, or the retified linear/ReLU functions):

$$\begin{aligned} h_t &= \tanh(W_X \cdot X_t + W_h \cdot h_{t-1}); \\ Y_t &= \text{softmax}(W_Y \cdot h_t) \end{aligned} \tag{2}$$

In the rest of the paper, we refer to this general function as a Multilayer Perceptron, or MLP (i.e., $h_t = \text{MLP}(X_t, h_{t-1})$). The weight matrices W_X, W_h, and W_Y are shared across all time steps and learnt via stochastic gradient descent on the backpropagation of errors.

One limitation of vanilla RNNs is that they do not readily capture long-range dependencies. Hochreiter and Schmidhuber [33] proposed adding memory units, or cells, within an RNN, which are able to "remember" information over arbitrarily-long intervals. These Long Short-Term Memory (LSTM) networks are rapidly becoming one of the most popular variants of the RNN.

Many researchers have since used RNNs and their LSTM variants to recognize emotion from speech and from video. [34], [35], [36] and [26] all used a Convolutional Neural Network to learn hidden layer features from individual video frames, along with a recurrency between hidden layers at consecutive times—thus, combining the time-independent CNN with a RNN. Many others have used LSTMs to recognize emotions from video data. [37], [38] and [39] were some of the earlier papers that worked on comparing multimodal LSTMs with Support Vector Regressions and other approaches for valence and arousal classification recognition on the SEMAINE dataset. This subsequently led to a surge of interest in applying LSTMs, especially to time-series emotion recognition on the AVEC 2015 [40], [41], AVEC 2017 [42], [43], AVEC 2018 [44], and OMG-Empathy 2019 [45] challenges. Other noteworthy examples are [46], who investigated bidirectional LSTMs (where there is another recurrence that goes "backwards" in time), and [47] who built an LSTM with electroencephalography (EEG) input. These papers have collectively found that RNNs/LSTMs are a powerful model for time-series emotion recognition, whether they rely on extracted low-level features, or combined with features extracted using CNNs.

Discriminative approaches, by and large, are the most popular type of time-series approaches we discuss, because they are a flexible approach that makes little assumptions about the nature of the data. At their heart, these approaches perform excellent pattern recognition, and find the best nonlinear functions that maps the input behavioral features to the output emotion via minimizing how badly the model predicts the output (also called the loss function). For tasks like emotion recognition from faces, deep approaches like Convolutional Neural Networks are by far the best performing state of the art. One drawback, however of making less structural assumptions about the data is that these discriminative approaches, especially deep neural network approaches like LSTMs, tend to require larger amounts of data to learn and perform well.

There is another important modelling decision for such models: how to deal with asynchronous inputs. Multimodal time-series input often come in at different sampling frequencies, and discriminative approaches require some kind of binning to synchronize them [12], [13]. One popular method (and the one that we use in this paper) is feature fusion, also called early fusion, where the input modalities are oversampled or undersampled (or otherwise averaged) to a common sampling rate. This allows the multimodal features to be concatenated into a single feature vector within a given time window to be fed into a model [35], [40], [41]. A second way to achieve such "synchronization" is decision fusion (or late fusion), which involves fitting a separate time-series model to each modality, operating at their own sampling frequencies, but connecting these individual models further into the computation to predict outputs [42], [48].

2.2 Generative Models

A second class of time-series approaches instead focus on modelling the causal structure behind the generation of the data [1], [49]. As we highlighted in the opening example, emotions evolve over time, and emotions may cause behavior like displaying emotional expressions. Thus, taking a modelling approach that is more sensitive to the underlying emotional phenomena, we may be interested in explicitly writing out how, say, the emotions evolve over time

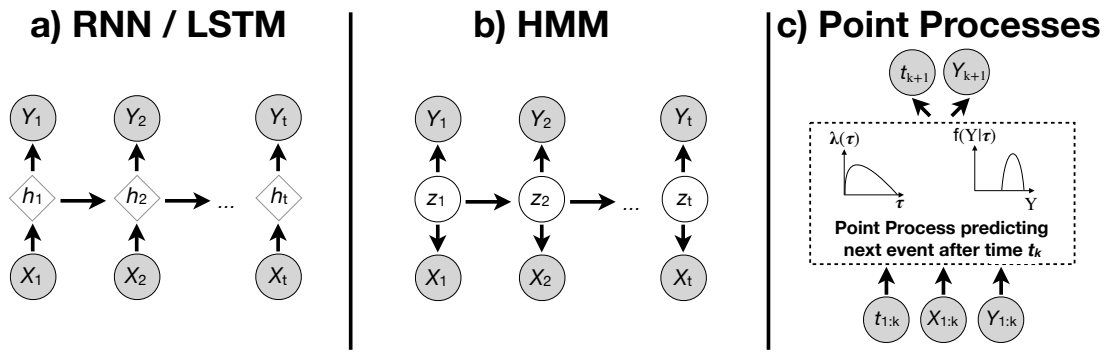


Fig. 1. Diagramatic overview of the different time-series approaches reviewed. We use conventional Bayesian Network notation, where circles represent random variables, shaded shapes represent observable quantities, and unshaded shapes represent latent quantities. We also use diamonds to represent deterministic values computed from random variables. (a) A representation of a Recurrent Neural Network (RNN) model. Note that, unlike in Bayesian Network notation, arrows here represent **information flow** in the model. The inputs X_t (e.g., emotional expressions) are mapped onto hidden states h_t to produce output labels Y_t (emotion labels), and there is a recurrency between consecutive hidden states. See Equation 1. The goal of the discriminative approach is to find a function that best discriminates the outputs given the inputs, modelling $P(Y|X)$. Note that Long Short-Term Memory (LSTM) networks, are variants of RNNs where the hidden layers also includes "memory" units that allow longer-range information dependencies. (b) A representation of a Hidden Markov Model (HMM). Here in (b), arrows represent **causal influence**. In the generative approach, there is some hidden (emotional) state z_t, which "causes" people to display emotional expressions X_t and also "causes" observers to rate these as certain emotional states Y_t. The goal of the generative approach is to model the joint distribution $P(X, Y)$, in the case of the HMM, by invoking and marginalizing out latent variables $P(X, Y) = \sum_z P(X, Y|z)P(z)$. (c) A very general illustration of the marked point-process approach. When we are at time t_k (with associated events X_k, Y_k), the point process is conditioned on the event history up to time t_k and is parameterized by two functions. The first, $\lambda(\tau)$ gives the distribution of event occurrences times, while the second, $f(Y|\tau)$, denotes the effect of an event at τ on the dependent variable Y. These distributions give the time of the next event $t_{k+1} = t_k + \tau$ and the predicted signal Y_{k+1}

($Y_t \rightarrow Y_{t+1}$), and how emotions cause emotional expressions ($Y_t \rightarrow X_t$). Generative models offer this flexibility, but with their own share of modelling assumptions and challenges. More generally, generative models aim to model the joint distribution of the observed data, both the inputs X and the outputs Y, or $P(X, Y)$. Indeed, the parameters in generative models are fit by maximizing the (log-)likelihood of the data under the model. By contrast, the discriminative models described in the previous subsection directly model the outputs given the features $P(Y|X)$, and are often trained by minimizing some loss function, which does not correspond directly to likelihood (see [28] for more discussion).

Let us illustrate this with a classic time-series generative model, the Hidden Markov Model (Fig. 1b). In this model, we posit that there is a latent (unobservable) variable z_t. This z_t is a discrete, categorical variable, and it could be something that the modeller may want to label (e.g., a discrete emotion category like *happy* or *sad*), or it could also be some unknown "state of the world" that the modeller may be agnostic about labeling. First, the latent variable at the current time step z_t "causes" both the input features X_t and the output labels Y_t via an emission function or emission model $z_t \rightarrow (Y_t; X_t)$. The model's emission probabilities encode how observations are "emitted" from the hidden states. Second, the latent variable at the current time step z_t evolves to the next time step z_{t+1} via a transition function $z_t \rightarrow z_{t+1}$ with transition probabilities governing how one hidden state may transition to another. The X_t's and the Y_t's are only connected via the z_t's, and each z_t is only influenced by the z at the preceding time-point.

The HMM allows one to set priors on both the transition and emission models. For example, one might have a theory that emotions tend to be "sticky" over the time-scale of the time steps, so the emotional state z_t would likely be similar to the preceding state z_{t-1}. Alternatively, emotion A may be more likely to precede emotion B than emotion C: These could all be set in the transition model via weights in a multinomial distribution. These priors are updated after observing the data. More generally, we can define parameterized distributions, and find the parameters θ that maximize the probability of the data under the model:

$$\begin{aligned} z_t &\sim P_\theta(z_t|z_{t-1}) \\ X_t &\sim P_\theta(X_t|z_t) \\ Y_t &\sim P_\theta(Y_t|z_t) \\ \theta^* &= \arg\max_\theta P_\theta(X_1, \ldots, X_T, Y_1, \ldots, Y_T) \end{aligned} \tag{3}$$

Hidden Markov Models have been used for many years to recognize time-series emotions, especially from continuous speech. [50], [51] and [52] all explored using HMMs to classify speech into discrete emotion categories. [53] did a more systematic investigation of how various parameters of HMMs (e.g. number of states or mixtures per state, input lengths) impact their performance at recognizing emotions in speech. The latent variable in a HMM can also capture different types of variability: for example, emotion dynamics within an utterance, versus emotion dynamics within a conversation across multiple utterances. [54] modelled exactly these two levels of emotion dynamics using a HMM with two hierarchical layers of latent variables. [55] also applied a multilevel HMM to recognize emotions from sequences of facial expressions.

Researchers have also tried other similar generative models to emotion recognition. For example, a Kalman filter is similar to a HMM with the main difference being that the hidden states are continuous and, with some linearity assumptions, allows modelling of graded variables like valence: [56] applied Multimodal Kalman Filters to recognize valence and arousal over time on the AVEC 2016 challenge. Working on the same dataset, [57] applied a Gaussian

Process Regression model, which is similar to a Bayesian Regression in that they assume a generative process over the parameters of a regression model (in this case, assuming that the covariance structure is the result of a Gaussian Process). [48] also used a Gaussian Process Regression, as well as a Gaussian Mixture Regression (which assumes that the model parameters are a result of a "mixture" or combination of multiple Gaussians) on the AVEC2017 dataset.

2.2.1 Integrating discriminative and generative approaches

Compared to discriminative approaches, generative approaches make more assumptions about the underlying structure of the data, such as which variables "cause" which other variables and how. These modelling assumptions provide an inductive bias [58], [59] that helps models to learn faster with less data. Generative models also allow the model to learn different sources of variability. For example, by using hierarchical latent variable models [54], we could potentially learn general emotion-cue mappings (e.g., people tend to smile like so when happy) as well as person-specific mappings (Bob tends to smile like *that* when happy).

Generative models also tend to be more computationally expensive to train, compared to discriminative approaches. Moreover, inference in these models is often NP-hard in all but the most simple models, and so many algorithms rely on approximate-inference algorithms. Fortunately, this situation is improving. In recent years, researchers have worked on models that merge the benefits of the discriminative and generative approaches, for example, by using techniques from deep learning to produce more efficient approximate-inference algorithms. In non-time-series domains, the Variational Autoencoder [60] has become a popular and flexible deep generative model—a generative model parameterized by neural networks, and where inference in the model can be approximated by maximizing a variational lower bound on the log-likelihood of the model. We recently proposed [49] that such deep generative approaches allow affective computing researchers to leverage the advantages of both generative and discriminative approaches.

Within time-series modelling, there are also a handful of promising examples of such integration. For example, in a Deep Markov Model [61], [62] or Deep Kalman Filter [63], one can parameterize the generative edges in the model (e.g., the emission and transition functions) using neural networks as in Eqn. 2. In another example, [64] and [65] both introduced a latent variable into a RNN to help it model different sources of variability (e.g. inter-subject variability) in the data. Within affective computing, [66] combined an LSTM and a Dynamic Baysian Network to extract word-level linguistic features for predicting emotional valence and arousal.

We hope that in due course, these contemporary hybrid techniques will improve by leveraging strengths of both approaches, and subsequently be adopted within the affective computing community. In Section 4, we also present an implementation of a modified Variational Recurrent Neural Network [64] that tries to combine these approaches.

2.3 Event-based models

The discriminative and generative models discussed thus far make a common assumption about the nature of the time-series data: The data is segmented into equally-sized time windows, and in every time window, there is an underlying output variable (emotion) that can be inferred using the behavioral cues present in that and previous time windows. Thus, the "basis" upon which the models operate is fixed time intervals. If we return to the example in the opening paragraph and hark back to many decades of emotion theory, many theories agree that emotions arise as a response to *events* in the world. Appraisal theories of emotion [67], [68], in particular, hold that emotions arise as a result of an agent's subjective evaluation ("appraisal") of an event in the world. Even before one gets to modelling the subjective appraisal [2], [49], this theoretical position suggests an emphasis on *events* as the underlying basis to predict emotions, rather than (or in addition to [1]) behavioral cues within a fixed time window.

This alternative approach to formulating time-series modelling is exemplified in event-based models, of which we shall focus on point process models [69]. We note up front that being "event-based" is orthogonal to the discriminative vs. generative distinction in the earlier sections: notably, point process models can also be discriminative [70], [71] or generative [72], [73]. The distinction we make is that the point process model tries to predict *events*. Events are usually defined as having two parts: their occurrence time, as well as the value of the dependent variable signal at the event[2]. Predicting events entails first specifying a distribution of event occurrences over time, commonly using Poisson Processes, Cox Processes, Hawkes Processes, Gaussian Processes, or some other learnable distribution. Let us denote the distribution of event occurrence times using a Conditional Intensity Function $\lambda(\tau)$, such that $\lambda(\tau)d\tau$ gives the expected number of events occurring within the next $d\tau$ time-units, conditioned on the event history up to τ. Second, for each event that occurs at time τ, the model predicts the value of the output Y using the conditional density $f(Y|\tau)$, again conditioned on the event history up to τ. The distribution of Y at time t is then given by:

$$f(Y,t) = \lambda(t)f(Y|t) \tag{4}$$

In Section 4.5, we implement a variant of this point process approach.

While event-based approaches are not as common within affective computing, one notable and very recent example is the work of Wataraka and colleagues [74], who proposed an event-filter model to model continuous valence and arousal prediction over time from speech events. In their model, a vocal event j produces an emotional response $h_j(t)$ and occurs at times captured by the function $\varphi_j(t)$. For example, if event j denotes laughter, then $h_j(t)$ represents the change in emotional valence signalled by a single laughter episode (e.g., a sharp increase, followed by some decay back to baseline; assumed to be the same across all episodes), and $\varphi_j(t)$ captures all the occurrences of laughter in the signal. Then, the emotional signal $Y(t)$ is then proportional to the sum of the convolution $h_j(t) \bigoplus \varphi_j(t)$ across all events j. They tested their event-filter model on the AVEC 2018

2. The effect on the dependent value signal is referred to as the "mark" of a "marked" point process

dataset, and it performed better than the audio-channel-only baselines for the AVEC 2017 and 2018 challenges [74].

In contrast to non-event-based approaches where one has to decide how to synchronize data from different modalities—e.g., using feature fusion, decision fusion, or some combination—event-based approaches avoid this decision and allows the model to handle asynchronously-arriving data. This may be useful in setting where some signals may come irregularly. For example, if we are interested in the emotions of two people are playing a competitive game, certain game events ("Player 1 wins") would be strong signals to their emotions, but difficult to capture in a non-event-based approach. One drawback—which is also shared by decision fusion methods—is that considering the inputs in each modality as separate events does not let the model learn or take into account the correlations (or interactions) between different modalities.

Though event-based approaches are still new within affective computing, we believe that such approaches may provide ways of implementing some desirable psychological assumptions. For example, these approaches, by conceptualizing emotion as an output of discrete events, may provide the most theoretically-satisfying implementation of the dynamics of emotional appraisals in time-series emotion recognition.

2.3.1 Interim discussion

Thus far we have covered three different classes of time-series modelling, each with their own set of assumptions. Though the "textbook" examples of these approaches that we discussed are quite different, the distinctions between these approaches are getting more blurred. This is especially true in recent years with hybrid approaches combining discriminative and generative approaches (e.g., [64], [66]), and discriminative and event-based approaches (e.g. [70]). We believe that in due course, we will see more research applying these contemporary modelling ideas to affective computing, incorporating more psychological theory to solve more difficult problems. In the next section, we introduce a new dataset, and in Section 4, we apply several of these modern modelling approaches to our data.

3 The Stanford Emotional Narratives Dataset (SEND)

In order to build affective computers that can understand human emotions in real life, we need high-quality time-series datasets with naturalistic emotion expressions. Here, we introduce the first[3] version of the Stanford Emotional Narratives Dataset (SENDv1). The SENDv1 consists of unscripted narratives of people recounting important and emotional life stories: It captures spontaneous naturalistic expressions as well as complex semantic content. These stories also have many different emotional trajectories, and thus provide a rich set of data for time-series modelling.

We refined the experimental protocol for the collection of the SENDv1 following our previous work [75], [76], and we report finer details of the current stimulus dataset collection in [77]. All experiments were approved by the Stanford University Institutional Review Board. Participants ("targets") were brought into the lab and told to think about the three most positive and three most negative events that they would feel comfortable sharing in front of a video camera. Recording was self-paced: targets were left alone in the room and talked for as long as they wanted about each event. After they finished recording the videos, targets were shown each video again, and were asked to give consent for us to use the videos in future experiments. We included several levels of consent, namely, no consent (upon which the video were deleted in front of the target), (i) whether they consented for the research team to view the videos, (ii) whether they consented for participants in future experiments (e.g., annotators in a crowd-sourced experiment) to view the videos, (iii) whether they would allow members of the public to view this video. This procedure of giving individual consent after re-watching each video was designed to protect targets' privacy and comfort. We note that all the videos in the SENDv1 were consented for participants in future experiments (i.e., annotators) to watch, although not all of them are consented for the public to view.

We selected a subset of 193 clips containing 49 unique targets. This set was chosen such that: (i) the target's face was always in the camera, (ii) the clips did not contain sensitive (e.g. mental health, suicide) content, and (iii) the clips had some narrative flow (rather than stream of consciousness rambling). These clips were also cropped for length, such that the final clips on average lasted 2 minutes 15 seconds (for a total of 7 hrs 15 mins). Targets in these clips talked about positive events such as winning a prize in school or going on vacation, to negative events like having a loved one pass away or experiencing a romantic breakup. These narratives were unscripted and spontaneous, and capture natural variation in emotion expression as the target is speaking.

We divided the current subset into Training (60% of the dataset, 117 videos, 4 hrs 26 mins long), Validation (20%, 38 videos, 1 hr 23 mins long) and Test (20%, 38 videos, 1 hr 26 mins long) sets. Importantly, we included 5 targets that appeared only in the Validation set, and 6 that appeared only in the Test set, to test the generalizability of our models to novel targets.

3.1 Independent Observer Ratings

We recruited a separate group of participants ("observers") on Amazon Mechnical Turk to watch the selected videos clips and provide ratings of how the target in the video felt along the valence dimension. Observers saw each video along with a continuous sliding scale underneath, and were asked to rate using their mouse how they thought the target was feeling as they were speaking in the video (and not how the target may have been feeling during the event they were describing). Observers were reminded to move the scale as the target is speaking to continually reflect the target's emotions. They used a visual analog scale that we divided into a hundred points, ranging from "Very Negative" to "Very Positive", measuring emotional valence, and the ratings on the scale were sampled every 0.5s. Many previous studies have used similar continuous rating dials, scales,

3. We plan to keep adding to the dataset, especially targets from more diverse demographic, socio-economic, and cultural backgrounds.

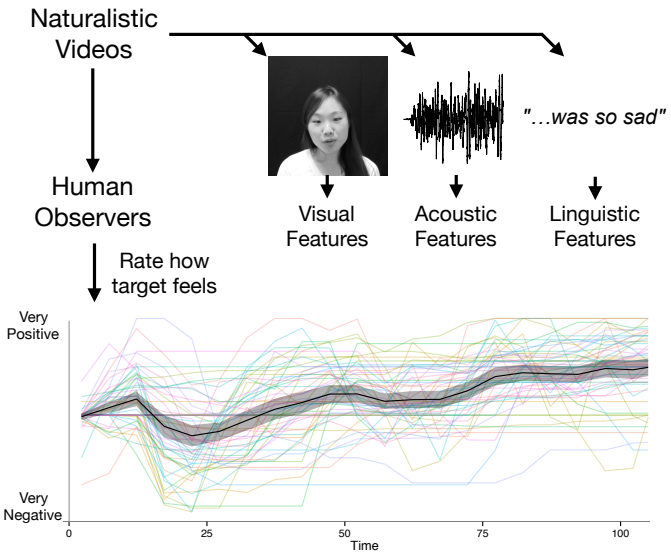


Fig. 2. Brief illustration of the dataset. Top: The SENDv1 consists of multimodal naturalistic videos, with rich, minutes-long information in three modalities: video, acoustic (i.e., paralinguistic) and linguistic (i.e., semantic content of the story). Bottom: The videos are annotated by independent observers for the target's emotional valence over time, using a continuous slider. We show example data from one video, where each colored line represents an individual observer's rating, and the black line and black region represents the mean rating with standard error. We observe that there is a lot of heterogenity in the way observers respond to the cues in the video (as well as heterogeneiety in the way they use the scale). In this paper, we report results of models trained on the mean observer rating.

or joysticks [26], [78], [79], [80], [81] to provide continuous valence ratings of videos.

Due to the complex nature of the stimuli, we aimed to get a large number of ratings (>20) per video for more reliability. Hence, we recruited 700 observers, who each watched 8 videos. To ensure that observers were paying attention, we included two comprehension checks per video, which were True/False questions pertaining to the content of the video. Overall, observers got both attention check questions correct on 82% of trials, one question correct on 15% of trials, and zero of two correct only on 2% of trials. We excluded trials with zero or one questions correct, which resulted in a total of 4607 rating vectors, giving an average of 23.9 rating vectors per video. We calculated the mean of the observer ratings and used them as the "gold-standard" emotional valence labels to be predicted.

3.2 Model Evaluation

We will use the Concordance Correlation Coefficient (CCC [82]) as the metric to compare our models' predictions for a time-series video with the gold-standard ratings. The CCC has been used in previous affective computing studies and challenges [21], [23]. Intuitively, the CCC captures the expected discrepancy between the two vectors, compared to the expected discrepancy if the two vectors were uncorrelated. The CCC for two time-series vectors X and Y is:

$$\begin{aligned}\text{CCC}_{XY} &\equiv 1 - \frac{E\left[(X-Y)^2\right]}{E\left[(X-Y)^2\right]|_{\text{setting } \rho_{XY}=0}} \\ &= 1 - \frac{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2 - 2\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2} \\ &= \frac{2\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2} \end{aligned} \quad (5)$$

where $\rho_{XY} \equiv \text{cov}(X,Y)/(\sigma_X\sigma_Y)$ is the (Pearson) correlation coefficient, and μ and σ denotes the mean and standard deviation respectively. Like ρ, the CCC measures agreement, where $+1$ means that the two time-series are in perfect agreement and 0 means that they are uncorrelated. The CCC also penalizes bias in the model's predictions via the $(\mu_X - \mu_Y)^2$ term in the denominator.

4 Modelling

In this section, we present several time-series approaches to model valence ratings on the SENDv1. We implement three different models, approximately spanning the breadth of models we reviewed earlier: a discriminative Long Short-Term Memory (LSTM) model, a deep generative Multimodal Variational Recurrent Neural Network model, and a Recurrent Marked Temporal Point Process model, each of which we adapted from previously published work. The goal of this section is to compare the different approaches, especially in the types of assumptions they make about the data.

As is conventional practice, we train our models only on the Training Set, and use the models' performance on the Validation set to choose hyperparameters. We then use these optimized settings to report results on the Test set. In addition to reporting mean results (e.g. on the Validation set), we also report standard deviations: This is to show the variability in model performance across the different videos in a particular partition of the dataset. Reporting SDs or other statistics is not common in Machine Learning, but we feel that this should be an area for improvement. Finally, the code for our models, written in PyTorch in Python, can be found at: https://github.com/desmond-ong/TAC-EA-model.

4.1 Human Benchmark

First, we wanted to establish how human observers perform on this task. This serves two purposes: First, it gives readers an intuition as to how difficult this task is. Second, it provides a quantitative benchmark with which to compare our modelling results in the next few sections.

The emotional valence labels that our models predict are the average of all the observer ratings. We wanted to calculate how well each individual observer j predicts this averaged rating—but because the averaged rating contains observer j's rating, we calculated the CCC of j's rating with the average, **subtracting out j's rating**. If we use $\mathcal{K}$ to denote the set of observers for video k (of length T_k), $R^j_{1:T_k}$ for observer j's ratings and $R^{\mathcal{K}\setminus j}_{1:T_k}$ as the mean of all the other observers less j, then the mean human CCC on video k is:

$$\overline{\text{CCC}}_k = \frac{1}{|\mathcal{K}|}\sum_{j\in\mathcal{K}} \text{CCC}\left(R^j_{1:T_k}, R^{\mathcal{K}\setminus j}_{1:T_k}\right) \quad (6)$$

where $|\mathcal{K}|$ is the number of observers for video k.

Using Eqn. 6, the mean and standard deviation of observer CCC on the **training set** was $.45 \pm .14$, the mean (and SD) observer CCC on the **validation set** was $.47 \pm .120$, and finally, the mean (and SD) observer CCC on the **test set** was $.46 \pm .14$.

4.2 Feature Extraction

To facilitate comparison across the different model types, we chose to extract features from all the modalities and combine them into a multimodal input feature vector (also called feature fusion or early fusion).

Audio Features. We used openSMILE v2.3.0 [83] with the accompanying `emobase` configuration file to extract 988 low-level acoustic features for every 1-second window.

Text Features. We collected professional annotations for all the videos, and used forced alignment[4] to assign timestamps to individual words. We used 300-dimensional GloVe word embeddings [84] as a representation for each word.

Visual Features. We used the Emotient software by iMotions[5] to extract 20 Action Units [85] for each frame.

4.3 Using Long Short-Term Memory Networks

As we noted in our review, the Long Short-Term Memory (LSTM) deep neural network is one of the most popular and successful discriminative approaches to time-series emotion recognition. They provide a flexible framework that can learn general nonlinear functions from multimodal input features (X_t) to an emotion output (Y_t, in our case, valence). We implemented three variants of a general LSTM architecture, which aims to model $P(Y_t|X_{1:t})$. In addition to a "vanilla" **LSTM**, we also built an autoregressive-LSTM (**AR-LSTM**), inspired by linear autoregressive models: Autoregression is when the model explicitly uses its predicted label at the previous time-steps to predict the label at the current time-step. We were interested to see if explicitly adding an autoregressive layer would help the performance of the model. Our third variant is the Encoder-Decoder LSTM (**ED-LSTM**), which have been successfully applied to predict sequences in other domains (e.g., [86]).

All of our three LSTM variants shared the first few steps (Fig. 3): First, the LSTM layer computes hidden states $h_1, \ldots, h_t$ from the input history. Next, we compute a local attention layer [87], [88] using a Multilayer Perceptron with a attention window of length l. This means that, at time t, we compute a set of l attention weights which are then used to weight the hidden states at the previous few timesteps, to give a context vector c_t:

$$h_t = \text{LSTM}(X_{1:t}) \tag{7}$$

$$\{a_{t-l+1}, \ldots, a_t\} = \text{MLP}(X_t) \quad \text{(attention weights)} \tag{8}$$

$$c_t = \sum_{j=0}^{l-1} a_{t-j} h_{t-j} \quad \text{(context vector)} \tag{9}$$

After this point, our three LSTM variants differed in their computation flow. In our vanilla **LSTM** model, the context vector is then fed into another MLP to decode the predicted output $\hat{Y}_t$. In our **AR-LSTM**, the context vector as well as the output at the previous time step $\hat{Y}_{t-1}$ is fed into the MLP to predict $\hat{Y}_t$. Finally, in our **ED-LSTM**, we added a second LSTM to predict the output (note that this LSTM will also implicitly use $\hat{Y}_{t-1}$, as in Fig. 3), and trained it using 50% teacher-forcing [32], meaning that with 50% probability on the training cases, the ED-LSTM was fed the actual value at the previous time step Y_{t-1}, while on the remainder, the ED-LSTM used its predictions on the previous time-step $\hat{Y}_{t-1}$.

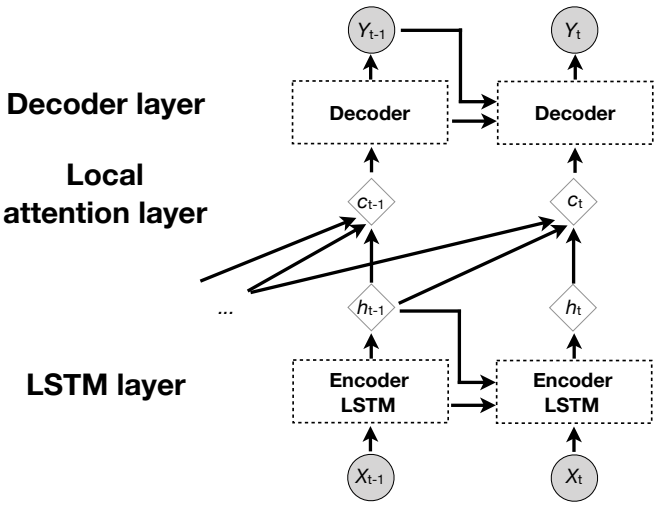


Fig. 3. Illustration of the LSTM models we applied. X_t is a multimodal vector of embedded features from all three modalities, and Y_t is the output of the model, a real-valued valence rating. The first layer in all our models, puts X_t through an LSTM layer to a hidden layer representation h_t. The local attention layer of length l computes a set of l attention weights: $a_{t-l+1}, \ldots, a_{t-1}, a_t = f(X_t)$, and computes the context variable c_t as a linear combination of the hidden units: $c_t = \sum_{j=0}^{l-1} a_{t-j} h_{t-j}$. The context vector c_t is then fed into a decoder layer to provide the final output Y_t. In our vanilla LSTM model, the decoder layer is a simple multilayer perceptron with a final regression output, $Y_t = f(c_t)$. In our autoregressive-LSTM (AR-LSTM) variant, the decoder layer is a MLP that also weights the prediction at the previous time-point, Y_{t-1}, so $Y_t = f(c_t, Y_{t-1})$. Finally, in our Encoder-Decoder-LSTM (ED-LSTM) model, the decoder layer is a second LSTM layer.

$$\text{LSTM:} \quad \hat{Y}_t = \text{MLP}(c_t) \tag{10}$$

$$\text{AR-LSTM:} \quad \hat{Y}_t = \text{MLP}(c_t, \hat{Y}_{t-1}) \tag{11}$$

$$\text{ED-LSTM:} \quad \hat{Y}_t = \text{LSTM}(c_t, \hat{Y}_{t-1}) \tag{12}$$

We used the Mean Squared Error (i.e., $MSE(\hat{Y}^k_{1:T_k}, Y^k_{1:T_k}) = \sum_{t=1}^{T_k} (\hat{Y}_t - Y_t)^2$ as the loss function to be minimized. We trained all our models with an initial dropout layer (on the input embeddings) of 0.1, which helps to regularize the learnt weights and help prevent overfitting [89]. Our best-fitting attention window, optimized on the validation set, was different for each LSTM variant: we used $l = 10$ for LSTM, $l = 1$ (no attention) for AR-LSTM and $l = 3$ for ED-LSTM.

4.3.1 LSTM Results

We summarize all our model results in Table 1. Our LSTM model performed the best when using only the visual features, achieving a mean CCC (with standard deviation) of $.19 \pm .31$ on the Validation set, and this went on to achieve a similar performance of $.23 \pm .31$ on the Test set. Using

4. https://github.com/ucbvislab/p2fa-vislab
5. https://imotions.com/emotient/

bimodal and multimodal features did not help the LSTM model perform as well as using only visual features.

Adding an autoregressive layer (AR) does not seem to improve the performance of the model. Indeed, upon qualitative inspection of the predictions, we found that the AR-LSTM model's predictions seemed to "drift" (e.g., gradually increasing or decreasing over the course of the video), and was less able to predict inflection points and other notable changes in the output labels, possibly because it learnt a. Overall the AR-LSTM seemed to perform the worst of the three LSTM variants.

Finally, the ED-LSTM was the best performing of the three variants, but its top performance, using both text-and-visual features, achieved only a CCC of .22 $\pm$ .27 on the Validation set, and did slightly worse at .15 $\pm$ .34 on the Test set. Overall for the ED-LSTM, adding more modalities seem to boost performance compared to the unimodal models, but the performances are still admittedly not high. Strangely, all our LSTM models do not seem to learn with only the Text features (LSTM: .00 $\pm$ .01, AR-LSTM: .08 $\pm$ .14, ED-LSTM: .04 $\pm$ .27).

One possibility why these models are not doing so well compared to previous research using LSTMs is that one of the most powerful features of neural networks—which we do not leverage here—is the ability to extract features directly from the raw data. For example, many previous models use a CNN on the raw images to extract visual features (e.g., [34], [35]), rather than calculating visual features separately as we did here. The weights of such a CNN will be modified during training, which "optimizes" the feature extraction process for this particular task. We chose not to do that here, and to have the same input data across all models to facilitate comparison, although we think that learning the feature extraction will likely improve the performance of the LSTM models.

Another possibility for the low performance is that the SENDv1 dataset is especially varied in terms of its narrative content, with complex speaker-specific and story-specific relationships between the input modalities and the valence ratings. Since LSTMs are not designed to model these implicit sources of variation, they may end up performing well on a certain subset of examples, but poorly on others. Indeed, this is what we saw when examined the LSTM predictions—for some videos, there was a very close match between human ratings and model predictions, but for other videos, there was substantial divergence.

4.4 Using a Multimodal Variational Recurrent Neural Network

Given the above limitation of LSTMs, it makes sense to consider models which can account for sources of variation such as narrative style or speaker-dependent attributes. One way to do this is to build a generative model of the inputs X_t and the outputs Y_t, modelling them as generated from some lower-dimensional latent state z_t. By training the model to accurately predict both X_t and Y_t, it could then automatically learn a good latent representation z_t that captures the aforementioned sources of variation. If the model learns to map particular dimensions of z_t onto these sources of variation, it could then go on to learn that some of them are irrelevant for prediction emotion, thereby allowing the model to generalize well across videos.

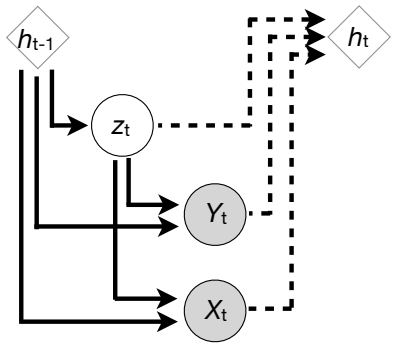


Fig. 4. Graphical structure of the Multimodal Variational RNN (MVRNN), adapted from [64], [90]. The hidden state at the preceding time step h_{t-1} parameterizes all of the distributions at the current time step t. First, we estimate the posterior distribution $Q(z_t|X_t, Y_t; h_{t-1})$ given the true inputs and outputs X_t and Y_t, and sample z_t from the posterior. Then we sample $\hat{X}_t$ and $\hat{Y}_t$ from the generating distribution $P(X_t, Y_t|z_t)$ to compute a reconstruction loss. Finally, we compute the recurrence to h_t using the sampled z_t and the observed X_t, Y_t (replacing X_t with $\hat{X}_t$ only if X_t is missing, and similarly for $\hat{Y}_t$). We use a solid line to indicate "causal flow" (as in a graphical model), and dashed lines to indicate a deterministic computation.

With this rationale, we propose and implement a Multimodal Variational Recurrent Neural Network (MVRNN). We adapted the VRNN, proposed by [64], to handle multiple modalities, by using a method from the (non-time-series) Multimodal Variational Autoencoder [90]. See Figure 4. In our model, at each time step, we sample the latent variable z_t from the approximate posterior $Q(z_t|X_t, Y_t)$, which is parameterized by the hidden state at the previous time step h_{t-1}. We follow [90] and assume a Gaussian prior $P(z_t)$ on the latent space, as well as Gaussian posteriors $Q(z_t|X_{t,m})$ for each input modality $X_{t,m}$ ($1 \leq m \leq M$ where M is the number of modalities); the full posterior $Q(z_t|X_t, Y_t)$ is then a product of Gaussians (itself a Gaussian).

$$
\begin{aligned}
z_t &\sim Q(z_t|X_t, Y_t) \\
&= P(z_t)\, Q(z_t|Y_t) \textstyle\prod_{m=1}^{M} Q(z_t|X_{t,m}) \qquad (13)\\
\text{where } P(z_t) &= \mathcal{N}\left(\mu_{z_t}, \sigma_{z_t}\right), \\
Q(z_t|Y_t) &= \mathcal{N}\left(\mu_{z_t|Y_t}, \sigma_{z_t|Y_t}\right), \\
Q(z_t|X_{t,m}) &= \mathcal{N}\left(\mu_{z_t|X_{t,m}}, \sigma_{z_t|X_{t,m}}\right), \\
\text{and } \mu_{z_t}, \sigma_{z_t} &= \text{MLP}(h_{t-1}), \\
\mu_{z_t|Y_t}, \sigma_{z_t|Y_t} &= \text{MLP}(Y_t, h_{t-1}), \\
\mu_{z_t|X_{t,m}}, \sigma_{z_t|X_{t,m}} &= \text{MLP}(X_{t,m}, h_{t-1})
\end{aligned}
$$

Next, we reconstruct the multimodal inputs $\hat{X}_t$ and outputs $\hat{Y}_t$ from the sampled z_t; these likelihood distributions are also parameterized by h_{z-1}. Finally, the recurrence occurs by computing the next hidden state h_t via a deterministic computation from z_t, X_t and Y_t, parameterized by a Multilayer Perceptron. In the event that there is a missing input modality m at time t, we use the reconstruction $\hat{X}_{t,m}$ in place of the unobserved inputs $X_{t,m}$ to compute h_t.

Similarly, we replace Y_t with $\hat{Y}_t$ if the former is missing.

$$\hat{X}_t \sim P(X_t|z_t) = \mathcal{N}(\mu_{X_t}, \sigma_{X_t}) \quad (14)$$

$$\hat{Y}_t \sim P(Y_t|z_t) = \mathcal{N}(\mu_{Y_t}, \sigma_{Y_t}) \quad (15)$$

$$h_t = \text{MLP}(z_t, X_t, Y_t) \quad (16)$$

$$\text{where } \mu_{X_t}, \sigma_{X_t} = \text{MLP}(z_t, h_{t-1})$$

$$\mu_{Y_t}, \sigma_{Y_t} = \text{MLP}(z_t, h_{t-1})$$

To train the MVRNN, we maximize the Evidence Lower Bound (ELBO) used in variational inference, summed across all timesteps t:

$$\sum_{t=1}^{T} \Big[\mathbb{E}_{Q(z_t|X_t,Y_t)} [\alpha \log P(Y_t|z_t)] + \mathbb{E}_{Q(z_t|X_t,Y_t)} \left[\sum_{m=1}^{M} \lambda_m \log P(X_{t,m}|z_t) \right] - \beta \, \text{KL}[Q(z_t|X_t, Y_t)||P(z_t)] \Big] \quad (17)$$

Here, α, β, and λ_m are weights balancing the importance of each ELBO term, and KL$[Q||P]$ is the Kullback-Leiber divergence between distributions Q and P. By maximizing the ELBO, the network simultaneously learns better generating distributions $P(Y_t|z_t)$ and $P(X_t|z_t)$, while performing regularization by ensuring that the approximate posterior $Q(z_t|X_t, Y_t)$ does not diverge too far from the prior $P(z_t)$.

During training, we gradually increase the weights α and β from zero as we increase the number of epochs. This allows the network to first learn how to reconstruct the inputs X_t by improving $P(X_t|z_t)$, before eventually placing more emphasis on both reconstructing the outputs Y_t and regularizing the network. We also scale each λ_m inversely with the dimensions of each input modality m, ensuring that reconstruction of that modality is not favored simply because it has more feature dimensions.

4.4.1 MVRNN Results

Overall, the MVRNN performed better than the LSTM models (Table 1). Our best-performing MVRNN model—and the best across all the models that we tried—achieved a CCC of .43 $\pm$ .32 on the Validation set, which is very close to the human benchmark of .47 $\pm$ .12. This model's performance on the Test set drops to .32 $\pm$ .35, but it remains the best-performing model on the Test set. We note that this model only used the Text features. Indeed, this is the opposite pattern that we see with the rest of the models: The LSTM models (and the RMTPP model in the next section) all cannot learn well from the Text features, but the MVRNN seems to perform best with these features.

Compared to the LSTM models, the MVRNN theoretically models different sources of variability using the latent variable z_t. Indeed, we predicted from our own qualitative impressions of the SENDv1 dataset that being able to account for different sources of variability would be critical to performance, and we were heartened to see that borne out in the MVRNN model results.

4.5 Recurrent Marked Temporal Point Process Model

Finally, we were interested in taking an event-based approach to modelling our data. We adapted the Recurrent Marked Temporal Point Process model from [71]. In point-process models, the prediction of the next event depends on the history of all the preceding events: [71] used a RNN to embed information from all the previous events into a hidden state vector that summarizes the event history. In our adaptation, we replaced their RNN with an LSTM as a hidden-layer embedding of the previous event history.

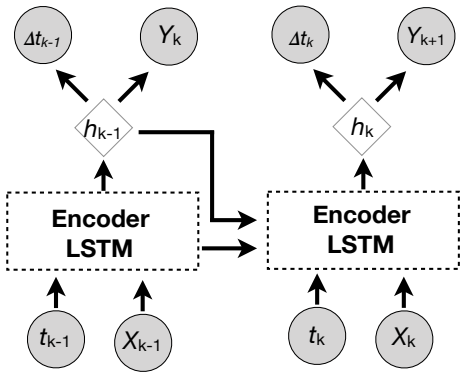


Fig. 5. Adaptation of the Recurrent Marked Temporal Point Process model [71]. Input features of event k with their timestamps (t_k, X_k) are fed into an LSTM which encodes the past history of events into a hidden state h_k. We then train two Multilayer Perceptrons to predict the time to the next event Δt_{k+1} as well as its impact on the output signal Y_{k+1}.

In our model, an event is a tuple that consists of a timestamp, and either some input features (t_k, X_k) or a valence output (t_j, Y_j), and we shall refer to these as input events and output events. Note that unlike the previous non-event-based models, the variables X, Y here are indexed by event, rather than time-window. Input events (t_k, X_k) are fed into an LSTM to provide a hidden layer embedding h_k, which also embeds information from preceding events. From h_k, we learn two more MLPs that predict the time to the next *output event*, Δt_k, and the predicted effect on the emotional signal that is associated with that event, Y_{k+1}. That is, these networks output a prediction of the next output event $k+1$, occuring at time $t_{k+1} = t_k + \Delta t_k$, and with associated emotional intensity, Y_{k+1}.

$$h_k = \text{LSTM}((t_k, X_k)) \quad (18)$$

$$\Delta t_k = \text{MLP}(h_k) \quad (19)$$

$$Y_{k+1} = \text{MLP}(h_k) \quad (20)$$

Note that unlike [71] who learnt a conditional density function, we predicted Δt_k directly with an MLP [70]. We also chose to directly learn Y_{k+1} directly rather than an impulse response function[6].

We processed the multimodal input data as separate event streams, according to their sampling frequency. Thus, the audio features were fed in as events every second, and the visual features were fed in every frame. We found that using individual word-level timestamps did not work as well as we expected, even after removing commonly-used stop words (e.g. "the"); thus, we used averaged word-vectors in five-second windows (as in the LSTM and MVRNN models). For each event, missing modalities

6. Alternatively, we assume an instantaneous Dirac-delta impulse response function.

Model	Modalities						
	A	T	V	AT	TV	AV	ATV
Validation CCC (Std. Dev.)							
LSTM	.13 (.23)	.00 (.01)	*.19 (.31)*	.07 (.24)	.08 (.19)	.06 (.23)	.05 (.18)
AR-LSTM	.07 (.18)	.08 (.14)	*.11 (.22)*	.12 (.23)	.08 (.19)	.08 (.17)	.04 (.27)
ED-LSTM	.16 (.29)	.04 (.27)	.14 (.27)	.09 (.28)	*.22 (.27)*	.14 (.28)	.16 (.28)
VRNN	.14 (.27)	**.43 (.32)**	.17 (.26)	.15 (.24)	.37 (.29)	.16 (.24)	.23 (.27)
RMTPP	.10 (.22)	.01 (.02)	*.19 (.25)*	.11 (.22)	.07 (.14)	.10 (.22)	.11 (.23)
Human	–	–	–	–	–	–	.47 (.12)
Test CCC (Std. Dev.)							
LSTM	.05 (.23)	.00 (.01)	*.23 (.31)*	.05 (.18)	.09 (.27)	.10 (.26)	.04 (.18)
AR-LSTM	.05 (.22)	.06 (.14)	*.08 (.19)*	.13 (.24)	.10 (.24)	.01 (.15)	.04 (.19)
ED-LSTM	.15 (.27)	.09 (.29)	.13 (.33)	.04 (.19)	*.15 (.34)*	.11 (.27)	.10 (.27)
VRNN	.08 (.25)	**.32 (.35)**	.16 (.26)	.12 (.21)	.29 (.29)	.14 (.27)	.18 (.25)
RMTPP	.08 (.23)	.01 (.03)	*.10 (.23)*	.09 (.25)	.04 (.09)	.13 (.21)	.06 (.21)
Human	–	–	–	–	–	–	.46 (.14)

TABLE 1
Summary of model results. Modalities—A: Audio, T: Text, V: Visual. Human: mean CCC between an individual human rater and the average of all other human ratings (described in Section 4.1). For each model, we italicize the best performing modality combination on the Validation Set, as well as the corresponding performance of that model-modality combination on the Test set. We also bold the overall best-performing model on the Validation set, which is also the best-performing model on the Test set.

were zero-masked, that is, for a word event that had no accompanying visual and acoustic features, we filled those modalities with zeros in the X_k embedding. Thus, we did not have to restrict the input features to be on the same time-scale, and this model is able to asynchronously handle data streams from different modalities. The model is trained to predict only output events, i.e., that affect the output variable Y.

We made an additional design choice regarding the predicted output event time-stamps. If, when predicting the next output event $k+1$, the predicted arrival Δt_k is too large, such that there is another input event m where $t_m - t_k < \Delta t_k$ (i.e., input event m happens after input event k, but before the "predicted" output event Δt_k time-units later), then we "ignore" the predicted output event in updating Y, and continue with event m as the next input event. This is based on the assumption that if input event m occurs before the predicted change in Y, then we should instead use the information from event m as it is the more recent event. This is a strong assumption that could be relaxed in future improvements of this model.

4.5.1 RMTPP Results

On average, the RMTPP model performed worse than the MVRNN and the ED-LSTM, and performs only slightly better than the LSTM and AR-LSTM. The best-performing RMTPP model used only the visual features, and achieved a CCC of .19 $\pm$ .25 on the Validation set, and a much poorer performance of .10 $\pm$ .23 on the Test set.

We made several assumptions in processing the data into events. We did not identify specific events in the acoustic (e.g. unlike [74]'s vocal bursts) or other modalities. Within the text modality, for example, we could have used NLP techniques to identify salient "events" from the words. Instead, we simply fed in each data stream as a sequence of time-stamped events according to their sampling frequency. Thus, one potential avenue for improvement would be in properly identifying or filtering "events" from the data-streams—and what those events might be will be an interesting theoretical and empirical question.

5 Discussion

We live in a constantly changing environment, and our emotions help us to skillfully navigate the world around us. Fear may motivate us to avoid harm, anger may prepare us to fight, and happiness may promote the forming of cooperative social bonds. Our environments, however, are always changing, and the fluctuations of external events results in corresponding fluctuations in our emotional responding. In order to build artificial intelligence that understands human emotions, one major challenge that researchers have to overcome is the modelling of such emotion dynamics. In this paper, we address just one piece of that puzzle—time-series emotion recognition—and offer a comprehensive review of contemporary time-series modelling approaches that are used or can be used productively in affective computing. In addition, we present a rich naturalistic dataset, the first version of the Stanford Emotional Narratives Dataset (SENDv1), designed precisely for multimodal, time-series emotion recognition. We report the results of three classes of models, with our best-performing model so far being a Multimodal Variational Recurrent Neural Network, which combines some of the advantages of discriminative and generative approaches. That said, we are sure that that future work could optimize all of these models. For example, all our best-performing models used only one or two modalities, and in future work we should optimize these models to better integrate multimodal information to improve performance.

The manner in which the majority of affective computing conceptualizes emotion understanding is primarily via emotion recognition. That is, an affective computer "understands" what a user is feeling if the affective computer perceives and processes behavioural cues like the user's facial expressions, and produces an output of what the user is feeling. This is a difficult task, due to the large complexity of how emotions are expressed in face, voice, and other modalities, and as we mentioned, the field has made much progress on this front [7], [8]. This assumption is also encapsulated in the discriminative time-series approaches we reviewed, which is to find the best (statistical) mapping from the behavioral cue data to an emotion label or rating.

From a psychological perspective, however, emotion recognition is just one of the many ways that people can understand someone's emotions [1], [2]. People understand how others' emotions arises as responses to events in the environment—including via subjectively evaluating the significance of the event, as in Appraisal Theories of emotion [2], [67], [68]—or how emotions evolve in interpersonal interactions [91]. More generally, a theoretically-driven approach would suggest building a causal model of how emotions arise, how they evolve over time, and how they result in behavior [49], and use these causal models as a basis for emotion understanding. This is the assumption behind the generative approach and the event-based approaches, which posit a causal data-generating process. These time-series

approaches are relatively new within affective computing, and we are excited about their potential to capture and model affective dynamics. There is still much work to be done: We note that the generative and event-based models we presented here still do not capture events and appraisals, as an emotion theorist would define them. Our models still define events in terms of behavioral cues, even though we used linguistic cues, and this still does not fully identify events (e.g. "I missed my train") that cause emotions. We think a fruitful set of future directions would be integrating existing models of what constitutes an emotionally-relevant event or cause of emotion (e.g. computational appraisal theories and architectures [68], [92]) into machine-learning models, and most likely in a generative or event-based approach.

More generally, a causal model-based approach may also be more applicable beyond multimodal time-series emotion recognition to "longitudinal" emotion understanding. For example, a medical robot that sees a patient once every few months might need to maintain a longitudinal record of what were the events that happened to the patient (e.g., diagnosis and continual treatment records, progression of the medical condition), in order to decide how best to affectively respond to the patient. Empathic doctors naturally do this, especially if the medical condition is sensitive (e.g., terminal or incurable), and even if there are long gaps between patient visits. Such longitudinal emotion understanding is, in a sense, a generalized version of the time-series problems we discussed in this paper: the observation (patient-robot interactions) may be irregularly spaced, but also driven by other "events" such as test results and other medical information.

In conclusion, time-series emotion recognition is a crucial component of affective computing. In this paper, we have outlined several challenges of—as well as several state-of-the-art solutions to—capturing dynamics in emotion recognition. We hope that this discussion will inspire more ambitious, theoretically-driven modelling using diverse combinations of approaches.

Acknowledgments

The authors would like to thank Emma Master, Kira Alqueza, Michael Smith, and Erika Weisz for assistance with the project, and Noah Goodman, Son Nguyen, and Arushi Goel for discussions about modeling. This work was supported in part by the A*STAR Human-Centric Artificial Intelligence Programme (SERC SSF Project No. A1718g0048), a Stanford IRiSS Computational Social Science Fellowship to DCO, and NIH Grant 1R01MH112560-01 to JZ.

References

[1] D. C. Ong, J. Zaki, and N. D. Goodman, "Affective cognition: Exploring lay theories of emotion," *Cognition*, vol. 143, pp. 141–162, 2015.

[2] ——, "Computational models of emotion inference in theory of mind: A review and roadmap," *Topics in Cognitive Science*, 2018.

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[4] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[5] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 71–82.

[6] R. A. Calvo and S. M. Kim, "Emotions in text: dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2013.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[8] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[9] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[10] P. Kuppens, Z. Oravecz, and F. Tuerlinckx, "Feelings change: Accounting for individual differences in the temporal dynamics of affect." *Journal of Personality and Social Psychology*, vol. 99, no. 6, p. 1042, 2010.

[11] M. Sudhof, A. Goméz Emilsson, A. L. Maas, and C. Potts, "Sentiment expression conditioned by affective transitions and social forces," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1136–1145.

[12] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 4. IEEE, 2005, pp. 3437–3443.

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

[14] B. Schuller, "Multimodal affect databases: Collection, challenges, and chances," *Handbook of Affective Computing*, pp. 323–333, 2014.

[15] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 415–424.

[16] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[17] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[18] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. 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.

[19] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[20] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 3–8.

[21] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[22] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[23] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[24] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 3–13.

[25] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," *arXiv preprint arXiv:1803.05434*, 2018.

[26] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *arXiv preprint arXiv:1804.10938*, 2018.

[27] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, 2018.

[28] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems*, 2002, pp. 841–848.

[29] B. Sun, S. Cao, L. Li, J. He, and L. Yu, "Exploring multimodal visual features for continuous affect recognition," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 83–88.

[30] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 9–16.

[31] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.

[32] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

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

[34] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 467–474.

[35] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.

[36] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 619–623.

[37] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 597–600.

[38] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, 2010.

[39] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[40] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[41] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 49–56.

[42] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 11–18.

[43] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

[44] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 65–72.

[45] Z. X. Tan, A. Goel, T.-S. Nguyen, and D. C. Ong, "A multimodal lstm for predicting listener empathic responses over time," in *OMG-Empathy Challenge workshop at the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2019.*, 2019.

[46] E. Pei, L. Yang, D. Jiang, and H. Sahli, "Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 208–214.

[47] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using eeg signals and facial expressions," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[48] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps, "Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 27–35.

[49] D. C. Ong, H. Soh, J. Zaki, and N. D. Goodman, "Applying probabilistic programming to affective computing," *IEEE Transactions on Affective Computing*, under review.

[50] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–1.

[51] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[52] D.-N. Jiang and L.-H. Cai, "Speech emotion classification with the combination of statistic features and temporal features." in *IEEE International Conference on Multimedia and Expo*. IEEE, 2004, pp. 1967–1970.

[53] J. Wagner, T. Vogt, and E. André, "A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 114–125.

[54] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2401–2404.

[55] I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition from facial expressions using multilevel hmm," in *Neural Information Processing Systems*, vol. 2. Citeseer, 2000.

[56] K. Somandepalli, R. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. S. Narayanan, "Online affect tracking with multimodal kalman filters," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 59–66.

[57] M. Atcheson, V. Sethu, and J. Epps, "Gaussian process regression for continuous emotion recognition with global temporal invariance," in *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, 2017, pp. 34–44.

[58] E. Schulz, J. B. Tenenbaum, D. Duvenaud, M. Speekenbrink, and S. J. Gershman, "Compositional inductive biases in function learning," *Cognitive psychology*, vol. 99, pp. 44–79, 2017.

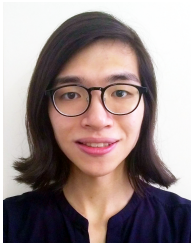
[59] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2017.

[60] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2013.

[61] R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models." in *AAAI*, 2017, pp. 2101–2109.
[62] E. Archer, I. M. Park, L. Buesing, J. Cunningham, and L. Paninski, "Black box variational inference for state space models," in *International Conference on Learning Representations (ICLR) Workshops*, 2016.
[63] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep kalman filters," in *Advances in Approximate Bayesian Inference & Black Box Inference Workshops at NIPS 2015*, 2015.
[64] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.
[65] J. Bayer and C. Osendorfer, "Learning stochastic recurrent networks," in *NIPS 2014 Workshop on Advances in Variational Inference*, 2014.
[66] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.
[67] P. C. Ellsworth and K. R. Scherer, "Appraisal processes in emotion," in *Handbook of Affective Sciences*, . K. R. S. R. J. Davidson, H. Goldsmith, Ed., 2003, vol. 572, p. V595.
[68] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. New York: Cambridge University Press, 1988.
[69] J. G. Rasmussen, "Temporal point processes: the conditional intensity function," 2011, lecture Notes, Jan.
[70] S. Xiao, J. Yan, X. Yang, H. Zha, and S. M. Chu, "Modeling the intensity function of point process via recurrent neural networks." in *AAAI*, 2017.
[71] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song, "Recurrent marked temporal point processes: Embedding event history to vector," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1555–1564.
[72] S. W. Linderman and R. P. Adams, "Discovering latent network structure in point process data," in *International Conference on Machine Learning*, 2014, pp. 1413–1421.
[73] Z. Qin and C. R. Shelton, "Event detection in continuous video: An inference in point process approach," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5680–5691, 2017.
[74] K. Wataraka Gamage, T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "Speech-based continuous emotion prediction by learning perception responses related to salient events: A study based on vocal affect bursts and cross-cultural affect in avec 2018," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 47–55.
[75] J. Zaki, N. Bolger, and K. Ochsner, "It takes two: The interpersonal nature of empathic accuracy," *Psychological Science*, vol. 19, no. 4, pp. 399–404, 2008.
[76] H. C. Devlin, J. Zaki, D. C. Ong, and J. Gruber, "Tracking the emotional highs but missing the lows: Hypomania risk is associated with positively biased empathic inference," *Cognitive Therapy and Research*, vol. 40, no. 1, pp. 72–79, 2016.
[77] D. C. Ong, "Computational affective cognition: Modeling reasoning about emotions," Ph.D. dissertation, Stanford University, 2017.
[78] R. W. Levenson and J. M. Gottman, "Marital interaction: physiological linkage and affective exchange." *Journal of personality and social psychology*, vol. 45, no. 3, p. 587, 1983.
[79] A. M. Ruef and R. W. Levenson, "Continuous measurement of emotion," *Handbook of Emotion Elicitation and Assessment*, pp. 286–297, 2007.
[80] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 488–500.
[81] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: an overview," *International Journal of Synthetic Emotions (IJSE)*, vol. 3, no. 1, pp. 1–17, 2012.
[82] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.
[83] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013, pp. 835–838.
[84] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162
[85] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978.
[86] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.
[87] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2015.
[88] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.
[89] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[90] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018.
[91] B. Mesquita and M. Boiger, "Emotions in context: A sociodynamic model of emotions," *Emotion Review*, vol. 6, no. 4, pp. 298–302, 2014.
[92] S. C. Marsella and J. Gratch, "Ema: A process model of appraisal dynamics," *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.

Desmond C. Ong received his Ph.D. in Psychology and M.Sc. in Computer Science in 2017 from Stanford University. He graduated with a B.A. in Economics (*summa cum laude*) and Physics (*magna cum laude*), with minors in Cognitive Studies and Information Science from Cornell University in 2011. He has been a Research Scientist with the A*STAR Artificial Intelligence Initiative since 2017. His research interests include building computational models of emotion and mental state understanding, using a mix of human behavioral experiments and modeling approaches like probabilistic modeling and machine learning. He is a member of the IEEE Computer Society.

Zhengxuan Wu is completing his M.Sc. in Management Science and Engineering with a concentration in Computational Social Science at Stanford University. He graduated with a B.S. in Aerospace Engineering (*magna cum laude*) and Mechanical Engineering (*cum laude*) from Case Western Reserve University. He also received a M.Sc. in Computer Science from the University of Pennsylvania. His research interests include studying the interplay between emotion and cognition with the applications of machine learning algorithms and computational modelings.



Zhi-Xuan Tan received a B.S. in Electrical Engineering and Computer Science (*magna cum laude*) from Yale University in 2018, and is currently a research engineer with the A*STAR Artificial Intelligence Initiative. Xuan's research interests include computational modelling of human moral psychology, as well as using cognitively-inspired approaches to build AI systems that can better understand and conform to people's intentions, goals, norms, and values

Marianne Reddan is completing her Ph.D. in the laboratory of Tor Wager at the University of Colorado Boulder in a combined degree program that intersects Cognitive Science and Psychology and Neuroscience. Her research interests include modeling the neural and physiological processes underlying emotion expression and modification. She uses machine learning to develop signatures of emotion expression which can then be targeted through behavioral interventions to improve quality of life.

Isabella Kahhale received her B.S. in Cognitive & Brain Sciences (*summa cum laude*), and minors in English and Ethics, Law, & Society from Tufts University in 2017. She is now a full-time research assistant with Professor Jamil Zaki in the Stanford Social Neuroscience Lab. Her research interests include empathy gaps with respect to underserved communities and the impact of emotionally biasing information on legal decision-making.

Alison Mattek obtained her Ph.D. in Psychological and Brain Sciences from Dartmouth College in 2017, and worked as a postdoctoral researcher in Psychology at Stanford University. She is currently a postdoctoral researcher in the Department of Psychology at the University of Oregon. Her research interests include emotion and motivation.

Jamil Zaki received his Ph.D. in Psychology from Columbia University in 2010 and did his postdoctoral work at Harvard University. Since 2012, he has been an Assistant Professor of Psychology at Stanford University. He has won numerous awards, such as the 2017 Sage Young Scholar Award, a 2016 Early Career Award from the Society for Social Neuroscience, a 2015 NSF CAREER Award, and a 2015 Janet T. Spence Award for Transformative Early Career Contribution and a 2013 Rising Star award, both from the Association for Psychological Science. His research interests include empathy and emotion understanding.