Online Affect Tracking with Multimodal Kalman Filters

Krishna Somandepalli
somandep@usc.edu

Rahul Gupta
guptarah@usc.edu

Md Nasir
mdnasir@usc.edu

Brandon M. Booth
bbooth@usc.edu

Sungbok Lee
sungbokl@usc.edu

Shrikanth S. Narayanan
shri@sipi.usc.edu

Signal Analysis and Interpretation Laboratory
University of Southern California, Los Angeles, CA, USA, 90089

ABSTRACT

Arousal and valence have been widely used to represent emotions dimensionally and measure them continuously in time. In this paper, we introduce a computational framework for tracking these affective dimensions from multimodal data as an entry to the Multimodal Affect Recognition Sub-Challenge of the 2016 Audio/Visual Emotion Challenge and Workshop (AVEC2016). We propose a linear dynamical system approach with a late fusion method that accounts for the dynamics of the affective state evolution (i.e., arousal or valence). To this end, single-modality predictions are modeled as observations in a Kalman filter formulation in order to continuously track each affective dimension. Leveraging the inter-correlations between arousal and valence, we use the predicted arousal as an additional feature to improve valence predictions. Furthermore, we propose a conditional framework to select Kalman filters of different modalities while tracking. This framework employs voicing probability and facial posture cues to detect the absence or presence of each input modality. Our multimodal fusion results on the development and the test set provide a statistically significant improvement over the baseline system from AVEC2016. The proposed approach can be potentially extended to other multimodal tasks with inter-correlated behavioral dimensions.

Keywords

Multimodal affective computing, arousal, valence, linear dynamical systems, Kalman filters

1. INTRODUCTION

Developing computational models for automatic emotion recognition and affect sensing has been an active field of research over the past few years. Researchers are increasingly using the arousal-valence (A-V) scale for a continuous and dimensional representation of emotional quality [1, 2]. This is especially true when one is interested in its dynamic evolution. Continuous tracking of a person's affective state on the A-V scale during the course of dyadic interactions and natural conversations has been investigated with some success. The Audio/Visual Emotion Challenge and Workshop (AVEC, [3]) has offered such opportunites and previous AVEC workshops have led to several investigations and novel outcomes in the field of affective computing (e.g., [4, 5]).

The AVEC 2016 challenge uses the REmote COLlaborative and Affective interactions (RECOLA [6]) dataset for the Multimodal Affect Recognition Sub-Challenge (MASC). The RECOLA dataset consists of synchronized and continuous data from multiple modalities recorded during dyadic interactions in French on a video conference that requires completion of a collaborative task. In this paper, we present a continuous affective state tracking system based on linear dynamical models with Kalman filters [7]. Our approach is inspired from the fact that arousal and valence are latent affective dimensions, thus making online state tracking systems such as Kalman filters a suitable choice. We modify the Kalman filtering scheme to incorporate the multimodal nature of this problem by accounting for the presence or absence of input modalities.

Contributions to affective computing research may be classified into two categories: feature extraction and model development. Within the scope of AVEC challenges, several studies have made contributions to both these fields. For instance, the work in [8] proposed new audio-visual features to capture speech spectrum characteristics and facial landmarks to track affect. Novel modeling schemes using template based methods and ensemble canonical correlation analysis were proposed in [9] and [10] respectively. Other novel methods proposed in previous AVEC challenges include the use of multi-scale temporal modeling [11], log-gabor filters [12] and recurrent neural networks [13]. In most of the aforementioned studies, performance of arousal prediction is better than that of valence. There is also strong theoretical and experimental evidence reporting that affect dimensions are inter-correlated (e.g., [14]). Leveraging this aspect, a fusion framework was proposed in [14] by modeling the state correlations and covariances. Following this lead, we use the predicted arousal as an additional noisy observation to improve the performance of tracking valence. Previous studies (e.g., [4]) have shown that acoustic features are most predictive of arousal and video features more predictive for valence. Additionally, while performing annotation of natural dyadic conversations, human annotators may not always be able to observe the subjects' faces or voices to make reliable judgements of arousal and valence. Motivated by these factors, we propose a knowledge-based conditional framework to select among different multimodal Kalman filters while performing online tracking.

In summary, the contributions of this paper are two-fold:

WOODSTOCK '97 El Paso, Texas USA

1) We employ linear dynamical systems to model unimodal predictions from trained regressors to initially predict arousal and then use the predicted arousal as an additional observation for valence prediction in an online fashion using Kalman filters.
2) We propose a conditional framework to select among different multimodal Kalman filters while performing online tracking based on cues that incorporate observability of audio-visual data. The proposed approach could be adopted for general behavioral coding where different behavioral codes are inter-correlated.

2. BACKGROUND

Affect tracking typically comprises two systems: feature extraction, which provides a low-level representation of the audio, visual and/or physiological recordings; and modeling approaches that translate the low-level descriptors into high-level affect-related information. Audio features, typically referred to as acoustic low-level descriptors (LLD), include a wide range of features that cover spectral, cepstral, prosodic and voice quality information. Several studies have shown the utility of LLD for A-V prediction, especially for arousal. In videos, the goal is to extract features that can capture the change and intensity of facial expressions over the duration of a task. These video features can be classified into appearance- and geometry-based. A few popular examples of the appearance-based features include multiscale local-binary patterns (LBP) and histogram of gradients (HOG) modeled using bag of words (BOW). A variant of LBP features, examined in spatio-temporal volumes of the video after convolving with 2D Gabor filter-banks, (LGBP-TOP) [15], has been recently used for automatic facial expression recognition [15] as well as affect tracking from video (e.g., [3, 4]). Video geometric features include identifying landmarks on the face as well as the shoulders (e.g., [14]) or the whole body (as with MoCap, [16]). These landmarks are then tracked to acquire low-level descriptors of the dynamics of facial or body gestures.

Physiological recordings of electro-cardiogram (ECG) and electro-dermal activity (EDA) have been used extensively to measure arousal, valence, categorical affect, among other mental states [17]. At rest, both of these signals are *tonic* in nature and thus *phasic* changes are used to measure more immediate stimuli responses. Slowly evolving changes to the tonic frequency for both ECG and EDA signals have been correlated with higher levels of arousal. Heart rate (HR; tonic) and heart rate variability (HRV; phasic) extracted from the ECG signal are typically used to quantify physiologic changes in the autonomic nervous system. Skin conductance level (SCL; tonic) and measures of skin conductance response (SCR; phasic) provide a complementary view.

Affect tracking is usually performed with human-annotated arousal and valence for 'gold standard' ratings. Modeling approaches here are generally supervised ones and can be broadly classified into three distinct categories: regression-based methods; continuous-time, discrete-state models; and continuous-time, continuous-state models. Support vector machines for regression (SVR, [18]) is perhaps the most widely used regression method for A-V prediction. The work in [19] was among the first to propose SVR for A-V recognition where audio-visual features were used to estimate continuous valued 'emotion primitives' which were then mapped to discrete emotions. Several studies (e.g., [20, 21]) have shown reliable affect tracking from a single modality such as music as well as multimodal data [4, 3]. Although such regression approaches predict affective dimensions on a continuous scale, they do not account for the dynamical evolution of these dimensions. In contrast, discrete- and continuous-state modeling methods are desired in order to account for state dynamics.

Discrete-state methods first quantize the continuous affect dimensions into discrete levels to discriminate between coarse categories such as active-passive [22], negative-positive [23] or multiple levels (e.g., four to seven levels as in [24]). Some commonly used approaches for modeling the relation between the features and the discretized levels are hidden Markov models (HMM, [23]) and conditional random fields [24]. Although many of these approaches post-process the predicted discrete levels to a continuous space, the inherent property of quantization results in information loss. In this context, continuous-state models are perhaps best suited for continuously tracking affect dimensions.

Recently, continuous-state methods coupled with regression approaches have shown promising results in affect tracking. Long short-term memory (LSTM, [25]) recurrent neural networks and Gaussian mixture models [26] were shown to perform superior to SVR methods in [27]. Another study [14], employed a bidirectional LSTM model in conjunction with an output-associative framework to achieve improved performance in affect prediction. Following this trend, a deep bidirectional LSTM was proposed in [5] which was the winner of the 2015 AVEC challenge.

A somewhat less explored continuous-state approach for affect tracking is using Kalman Filters [7]. Good performance for predicting arousal and valence by modeling acoustic features from music using Kalman filters was demonstrated in [28]. Given a small number of observations, only a few parameters need to be estimated for Kalman filters. They have the additional benefit of performing online tracking by propagating the predicted means and covariances of the state in time. In the proposed approach we use a combination of SVRs and Kalman filters to perform affect tracking. Due to the high dimensional nature of the multimodal features, we first employ SVR to acquire arousal and valence separately for each modality similar to the baseline paper [3]. The resulting predictions from individual modalities are treated as noisy observations of the underlying state which is assumed to be known during late fusion. To the best of our knowledge, no previous work has used Kalman filters for late fusion of unimodal predictions to perform online tracking of arousal and valence.

3. METHODS

We briefly introduce the corpus used for Multimodal Affect Recognition Sub-Challenge (MASC) of AVEC 2016, the baseline features provided as part of the challenge, and the additional features we use in our proposed approach. The unimodal predictions acquired are modeled as noisy observations in a linear dynamical system to track the underlying state of arousal and valence using Kalman filters. We then propose a conditional framework that selects different filters according to available modalities over the duration of the task. Detailed performance evaluation is conducted to assess our proposed approach.

3.1 Unimodal Predictions

3.1.1 Baseline Corpus and features

As previously described, we use the RECOLA database [6] as part of the MASC challenge in AVEC 2016 in all our experiments. The multimodal data in RECOLA include au-

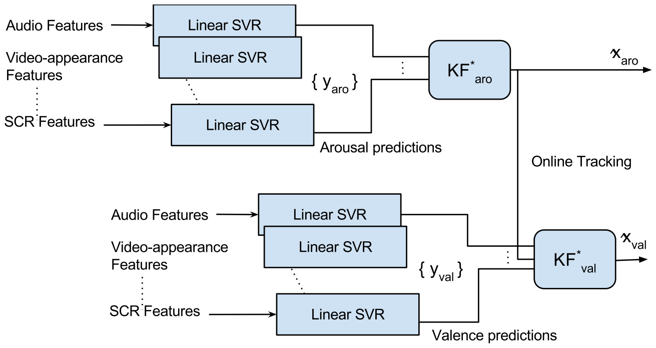

Figure 1: Overview of the proposed system

dio, video, ECG and EDA recordings. Time-continuous ratings were obtained for arousal and valence at 25fps. The inter-rater reliability measured by intraclass correlation coefficient ([3]) is high (ICC>0.8) for both arousal and valence indicating reliable gold standard ratings. Furthermore, the audio-visual recordings were of high quality with little background clutter in the video and clear speech. Data from nine distinct subjects was provided for each of training, development and testing. The subjects were gender balanced within and across groups.

To avoid repetition, we refer to [3] for details on the feature extraction procedures for all features provided as part of the MASC. Briefly, the audio features include a minimalistic set of acoustic LLD as per recommendations in [29]. Video appearance features included LGBP-TOP [15] features reduced by applying principal component analysis. Video geometric features involve aligning forty-nine landmarks to a mean-shape and tracking them for the duration of the video. Physiological features are extracted from EDA signals (SCL and SCR) and ECG signals (HR and HRV).

3.1.2 Additional Features

We supplement the baseline features with other features described below. Among these features, face status and voicing probability are used in the subsequent conditional framework.

Face Status P_v: As demonstrated in [30], LBP and other pixel-based appearance features suffer when the face pose is not frontal due to alignment errors. To quantify this aspect of observability of the subject's face, we use the face detection status from the *dlib* library [31] as a binary feature. This error measure can robustly quantify the observability since the face detection is conservative and fails when the face is approximately non-frontal (e.g., bending down, profile faces, etc).

Voicing Probability P_a: In this work, we use voicing probability as a cue to determine whether audio features are reliable. Several studies have used voicing probability for emotion recognition to extract speech features from voiced regions [32]. It has also been used directly for several emotion classification or clustering problems (e.g., [33, 34]). To extract voicing probability, we use the algorithm introduced in [35] implemented in the *Kaldi* toolkit [36]. It relies on a pitch tracking method that computes the probability of the frame being voiced. This method involves post-processing of the normalized cross-correlation function (NCCF) of the speech signal. Voicing probability along with face status are used as cues to select different Kalman filters in the conditional framework proposed in section 2.2.3.

Sparse dictionary representation of EDA (SD-EDA): Prior literature has shown the effectiveness of using EDA measures to predict valence and arousal (e.g., [37]). The baseline EDA feature set provides statistical moments of the noise-reduced signal every 40ms which does not directly capture SCRs. In our approach, artifacts are removed by fitting the original signal into a predetermined EDA-specific shape through sparse representation techniques [38]. SCR detection is performed with the *Ledalab* software [39]. The final EDA measures include the mean SCL, number of SCRs and mean SCR amplitude over non-overlapping 5sec windows. These measures of SCR are commonly used in analyzing EDA data and have been related to high arousal [40]. We use a 0.02μS minimum SCR amplitude threshold to remove noisy SCR shape matches. The data is inflated using linear interpolation to produce a time series at the same temporal resolution as the baseline data.

Teager energy-based MFCC (TEMFCC): The previous study in [41] demonstrated TEMFCC features to be more robust than MFCC for classification of discrete emotion categories, particularly in a noisy environment. TEMFCC features are computed by applying the non-linear Teager energy operator Ψ ([42]; equation 1) to the magnitude of the discrete Fourier transform in the process of computing MFCC.

$$\Psi(s[n]) = s[n]^2 - s[n-1]s[n+1] \tag{1}$$

3.1.3 Unimodal predictions

Separate arousal and valence predictions are obtained from individual modalities as described in the AVEC2016 paper [3] and the scripts provided as part of the MASC challenge. The regression task is performed using linear SVR provided with the liblinear library available in WEKA [43]. Data from the nine subjects in the development set is used to test the performance as well tune for different parameters after fitting the SVR models on the training set (see [3] for details). Unimodal predictions are first obtained from the eight feature-sets provided as part of the MASC baseline (audio, video-appearance, video-geometric, EDA, SCL, SCR, ECG, HRHRV). We conducted additional experiments by adding TEMFCC features to the baseline audio (audio+TEMFCC) and adding SD-EDA features as described earlier. The best delay parameter for augmented audio features is the same as that for baseline features (2.8s). However, the best delay parameter for the SD-EDA features is zero sec. The resulting predictions are post-processed to correct for bias and scaling issues as described in [3]. These unimodal predictions are used as observations in the proposed late fusion approach using linear dynamical systems.

3.2 Linear Dynamical System with Late Fusion

As described earlier, Kalman filters are ideally suited for continuous state tracking. We first present linear dynamical system formulation for tracking arousal and valence using the unimodal predictions. The state-space system parameters we use here are estimated using maximum likelihood and expectation maximization (EM) algorithms.

Linear dynamical systems can be described by a state equation and an observation equation as follows:

$$x_{t+1} = Ax_t + w_t \tag{2}$$

$$y_t = Cx_t + v_t \tag{3}$$

Here, we assume the gold standard ratings to be the continuous state, $\{\mathbf{x}\}_1^T = (x_1, x_2, ..., x_T)$. The predictions from different input modalities are treated as *noisy* observations or measurements, $\{\mathbf{y}\}_1^T = (y_1, y_2, ..., y_T)$ of the underlying state with discrete time indexed by t with total number of time points, T. The state noise w_t, and observation noise v_t, are zero-mean and normally distributed random variables with *output noise covariance* Q and *state noise covariance* R respectively. A is the *state dynamics matrix* which controls the state-evolution in time and C, the *observation matrix* which relates the observations to the state. The evolution of the state and its relation to the observations are assumed to be linear time-invariant.

3.2.1 Estimation of state equation parameters

We assume the gold standard ratings to be the latent state which is completely known in the training set. Additionally, since the state dynamics noise is assumed to be a zero-mean Gaussian random variable, we use maximum likelihood (ML) estimation to estimate the state dynamics matrix, A and the state noise covariance R of the autoregressive process in equation 2. The likelihood function of the ML estimate for A can be written as

$$L(\{\mathbf{x}\}_2^T; A, R) = \\ K \exp(\frac{1}{2}\sum_{t=1}^{T-1}(x_{t+1} - Ax_t)^\top R^{-1}(x_{t+1} - Ax_t)) \tag{4}$$

where $K = \frac{1}{(2\pi)^{(T-1)/2}|R|^{(T-1)}}$

For a given $|R| > 0$, maximizing L is equivalent to minimizing the squared error, $\mathcal{E}$ which is also the least-squares estimate of A;

$$\mathcal{E} = \sum_{t=1}^{T-1}((x_{t+1} - Ax_t)^\top(x_{t+1} - Ax_t)) \tag{5}$$

Following this, the state noise covariance is estimated with the covariance of the residuals from the ML estimate, $\hat{A}^{ML}$.

$$R \approx \frac{1}{T-1}\sum_{t=1}^{T-1}((x_{t+1} - \hat{A}^{ML}x_t)^\top(x_{t+1} - \hat{A}^{ML}x_t)) \tag{6}$$

In all our experiments, the state parameters are estimated for arousal and valence separately on the training set.

3.2.2 Estimation of observation equation parameters

The observations or measurements in our context are the unimodal predictions acquired as described in section 2.1.3. While the state to be tracked is a one-dimensional quantity, the observations are multidimensional. Lacking a physical model to describe the observation equation, a simple method to estimate the observation matrix C and observation noise covariance Q is to use ML estimation. Without additional constraints, this is equivalent to the least-squares estimation or the Moore-Penrose pesudoinverse which in this context would be an ill-posed problem since $dim(y) > dim(x)$. Previous work, [28] has employed a convex optimization approach in order to achieve stable estimates for the system parameters.

In this work we use the EM algorithm for linear dynamical system presented by Shumway and Stoffer [44] to estimate C and Q. We use a simple modification to the approach in [44] by assuming C to be unknown as implemented in Python[1]. A summary of the expectation (E-step) and maximization (M-step) steps are given below where EM aims to iteratively find $\theta = \{(C, Q) | \max_\theta P(\{\mathbf{y}\}_1^T; \Theta)\}$.

Table 1: Summary of Kalman filter equations

State prediction	$\hat{x}_{t+1\|t} = A\hat{x}_{t\|t}$
Covariance prediction	$P_{t+1\|t} = AP_{t\|t}A^\top + R$
Innovation (measurement error)	$\hat{y}_{t+1} = y_{t+1} - C\hat{x}_{t+1\|t}$
Innovation covariance	$S_{t+1} = CP_{t+1\|t}C^\top + Q$
Kalman gain	$K_{t+1} = P_{t+1\|t}C^\top S_{t+1}^{-1}$
State update	$\hat{x}_{t+1\|t+1} = \hat{x}_{t+1\|t} + K_{t+1}\hat{y_{t+1}}$
Covariance update	$P_{t+1\|t+1} = (\mathbf{I} - K_{t+1}C)P_{t+1\|t}$

We define the log likelihood function L as:

$$L(\{\mathbf{x}\}_1^T, \theta) = \log P(\{\mathbf{y}\}_1^T, \{\mathbf{x}\}_1^T; \theta) \tag{7}$$

and compute the expected log likelihood at iteration i:
E-step:

$$\Theta_i = \mathbb{E}_{\{\mathbf{x}\}_1^T}[L(\{\mathbf{x}\}_1^T, \theta) | \{\mathbf{y}\}_1^T, \theta_i] \tag{8}$$

M-step:

$$\theta_{i+1} = \arg\max_\theta(\Theta_i) \tag{9}$$

The parameters, C and Q are re-estimated by solving the partial derivative of Θ_i with respect to C and Q^{-1} respectively, after setting to zero. Detailed derivation of the EM steps can be found in [45]. Skipping intermediate steps, the final update equations are below:

$$C^{new} = \Big(\sum_{t=1}^{T} y_t\mathbb{E}[x_t|\{\mathbf{y}\}_1^T]^\top\Big)\Big(\sum_{t=1}^{T}\mathbb{E}[x_t x_t^\top|\{\mathbf{y}\}_1^T]\Big)^{-1} \tag{10}$$

$$Q^{new} = \frac{1}{T}\sum_{t=1}^{T}(y_t y_t^\top - C^{new}\mathbb{E}[x_t|\{\mathbf{y}\}_1^T]y_t^\top) \tag{11}$$

In all our experiments the parameters C and Q are estimated over the entire training partition of the development set.

3.2.3 Kalman filters for conditional online tracking

Before describing the conditional aspect of our framework, we first present the Kalman filter equations we use in our approach. Note that because $x_{t|\cdot}$ is a Gaussian random variable, it is sufficient to only keep track of the conditional means and covariances denoted as follows:

$$\hat{x}_{t|t'} = \mathbb{E}[x_t|y_{0:\,t'}] \tag{12}$$

$$P_{t|t'} = \mathbb{E}[(x_t - \hat{x}_{t|t'})(x_t - \hat{x}_{t|t'})^\top|y_{0:\,t'}] \tag{13}$$

The conditional mean and covariance are initialized to zero and one respectively. The measurement update and time update equations involved in the forward Kalman filter recursions [7, 45] are summarized in **Table 1**.

From the unimodal predictions acquired from the training set, we perform experiments on the development set in a leave-one-subject-out fashion. So, for each subject in the development set, we use the remaining eight subjects' data to estimate the state parameters A and R with ML estimation and the observation parameters C and Q using EM. We then perform online filtering on the left-out subject's unimodal predictions. Since our linear dynamical system model uses zero-mean Gaussian random variables, it is important to compute the bias terms of the state and observation equations. For each subject tested, we compute the measurement bias $\bar{y}$ as the average of each measurement vector from the

[1] https://pykalman.github.io/

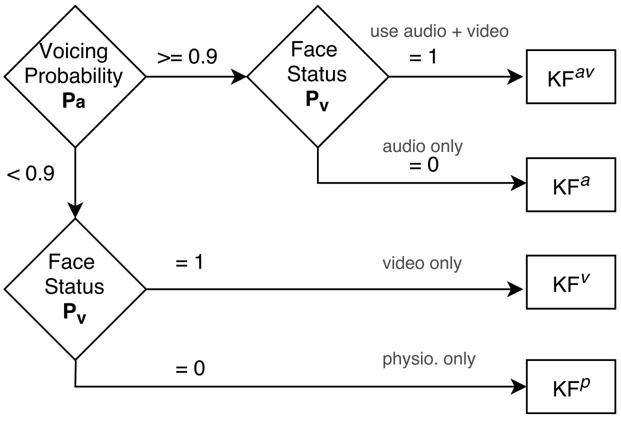

Figure 2: Conditional framework for Choosing Kalman Filters (KF)

training partition. Similarly, the state bias $\bar{x}$ is the average of the known state. Given a test subject's data, we first remove the pre-computed measurement bias. Subsequent to Kalman filtering, the state bias is added to the predicted state.

Qualitative observations on the AVEC 2016 data show relative decrements in arousal whenever the primary speaker (subject) remains silent. Similar decrements in valence are observed when the subject's face is not clearly observable. Additionally, it is reasonable to assume that video features and therefore the unimodal predictions are unreliable when the subject's face is not entirely frontal. As described previously, we use voicing probability to quantify the primary speaker's voice activation and a binary feature, face status to quantify the observability of the subject's face.

As with the unimodal predictions, arousal and valence are best predicted with audio and video features respectively. Hence, for late fusion we select either audio, video, audio and video, or physiological predictions using voicing probability and face-status as cues. We design four different Kalman filters: audio only ($KF^a; dim(y^a) = 1$), video only ($KF^v; dim(y^v) = 2$), audio and video ($KF^{av}; dim(y^{av}) = 3$) and physio only ($KF^p; dim(y^p) = 5$). Parameters for each of the filters are estimated in a leave-one-subject-out fashion as described before. Different filters are designed for arousal and valence, denoted as, KF^*_{aro} and KF^*_{val} where * refers to the variable modality. Since we use Kalman filters to perform online tracking, it ensures that the predicted state means and covariances are propagated in time thereby preserving the dynamic evolution of the state being tracked.

We use a conditional logic framework to select from the four filters while performing frame-wise online tracking. At each frame, if the face is observable (face status = 1) and the subject is speaking (voicing probability > 0.9), we use KF^{av}, else if only the face is observable or subject is speaking, we use KF^v or KF^a accordingly, else use KF^p. For valence tracking, as mentioned earlier, we use the predicted arousal as an additional observation. The overall algorithm for the proposed approach is given in **Algorithm 1**. The threshold for voicing probability is determined to be 0.9 using grid search on the development set by varying the probability values from 0.6 to 1.0 in steps of 0.05. The appropriate bias term $\bar{x}$ is added back to the arousal and valence estimates acquired by tracking. Additional experiments are conducted without the conditional framework to evaluate the performance of the system.

```
Algorithm 1: Conditional Online Tracking
Input: Unimodal arousal/valence predictions of time
       duration, $T$; $\{\mathbf{y}^*_{\mathbf{aro}}\}^{\mathbf{T}}_{\mathbf{1}}$ and $\{\mathbf{y}^*_{\mathbf{val}}\}^{\mathbf{T}}_{\mathbf{1}}$ with
       corresponding $KF^*_{aro}$ or $KF^*_{val}$
Output: Tracked arousal and valence; $\{\hat{\mathbf{x}}_{\mathbf{aro}}\}^{\mathbf{T}}_{\mathbf{1}}$ and
        $\{\hat{\mathbf{x}}_{\mathbf{val}}\}^{\mathbf{T}}_{\mathbf{1}}$
Parameters: Vocing probability, $P^a$ and face
            status, $P^v$ and frame, $t$
$\hat{x}_{1,aro} = 0; \hat{x}_{1,val} = 0; t = 2$
while ($t \leq T$) do
    Choose $KF^*_{aro}$ and $KF^*_{val}$ according to Figure 2
    $\hat{x}_{t,aro} \leftarrow KF^*_{aro}(\hat{x}_{t-1,aro}, y^*_{aro})$
    $y'^*_{t,val} \leftarrow \{y^*_{t,val}, \hat{x}_{t,aro}\}$
    $\hat{x}_{t,val} \leftarrow KF^*_{val}(\hat{x}_{t-1,val}, y'^*_{t,val})$
    $t = t+1$
end
```

3.3 Experiments and Performance Evaluation

We first perform unimodal predictions on arousal and valence separately using the methods described in sections 2.1.1–2.1.3. In our preliminary experiments, unimodal performance with audio+TEMFCC and SD-EDA is slightly better (not statistically significant) than the baseline audio and SCL/ SCR/ EDA features respectively. Taking this into account, we only retain predictions from audio+TEMFCC, video-appearance, video-geometric, ECG, HRHRV and SD-EDA. These predictions are grouped into the three audio, video and physiological modalities in our subsequent fusion framework.

The main contribution of this paper is late fusion with variable multimodal online Kalman filters based on a conditional framework. To evaluate this system we perform the following experiments:

- System 1: Track arousal and valence separately using all modalities with no conditional framework.
- System 2: Track arousal first and use the predicted arousal to track valence using all modalities with no conditional framework.
- System 3: Track arousal first and use the predicted arousal to track valence with the conditional framework proposed in **Algorithm 1**

Systems 1-3 are tested on the development set in a leave-one-subject-out fashion. In addition to estimating the state bias term from the training partition, we postprocess the outputs as described in [3] to correct for scale and bias. For predicting arousal and valence on the test set, the state-space system parameters were estimated only on the development set to prevent overfitting.

The system performance is quantified using the concordance correlation coefficient (CCC, [46]) as proposed for the MASC competition in AVEC 2016. We also computed Pearson's correlation coefficient (CC) which is the upper limit of the CCC. A significantly lower CCC with respect to CC would indicate that the bias and scale in the predicted state are not similar to the gold standard ratings. In order to examine if the CCC from different experiments are significantly different, we convert them into a z-score with Fisher's r-to-z transformation. The z-scores are then compared using formula 2.8.5 (pg. 54) from [47] returning a p-value from a two-tailed t-test. We use the number of frames per subject

Table 2: Unimodal performance using CCC on the development and test sets

Modality	Arousal	Valence
Audio+TEMFCC	**0.800**	**0.448**
Video-appearance	0.481	0.474
Video-geometric	0.297	0.612
SD-EDA	**0.080**	**0.178**
ECG	0.272	0.159
HRHRV	0.383	0.298

(7501) as the sample size in the test[2]. This test is typically used for comparing correlation coefficients. Since CCC is proportional to CC and has the same range (-1 to 1), we can use this method to compare two independent CCCs without loss of generality.

3.4 Results and Discussion

The CCC values for unimodal performance in predicting arousal and valence separately are shown in **Table 2**. Results show comparable unimodal performance for audio+ TEMFCC features with the baseline[3]. Although increased performance with TEMFCC features compared to MFCC was shown in noisy environments as described in [41], no significant improvement in performance is observed here due to the audio recordings being mostly clean speech. SD-EDA features perform somewhat better than that of the baseline EDA, SCL and SCR features (approximately 1% improvement for arousal and 7% for valence). This shows the utility of shape-based methods for extracting SCL and SCR features as proposed in [38] rather than statistical moments. Detailed analyses and further processing of the SD-EDA features for A-V recognition tasks would be a part of our future work.

The CCC for multimodal performance of the proposed Systems 1-3 is shown in **Table 3**. The baseline performance on the development set as reported in [3] and that obtained using leave-one-subject-out method is also shown for comparison. Overall, we have achieved a better performance for predicting arousal than valence consistent with existing linear modeling frameworks. The relatively lower performance of predicting valence is likely due to the non-linearities in the relationship between the low-level features and valence ratings. Additionally, we assume the evolution of valence dimension to be linear and time-invariant (i.e., represented by the state dynamics matrix, A in equation 2). Our future work would involve additional experiments which can verify these assumptions.

We observe a positive correlation between the gold standard ratings of arousal and valence in the MASC data (CC = 0.42 for training, CC = 0.56 for development set). Consistent with this, a significant improvement is achieved by using the predicted arousal as an additional observation in predicting valence (System 1 < System 2; $p < 0.01$) [3]. This is consistent with previous studies that observe inter-correlations between arousal and valence [14].

The system performance is further improved upon using the conditional framework to select different Kalman filters during online tracking (System 3; see Algorithm 1). Both arousal and valence predictions are significantly higher ($p < 0.01$) compared to Systems 1-2. Furthermore, the performance of System 3 is better than that of the baseline system obtained through leave-one-subject out method on the development set. Both arousal and valence predictions were significantly higher ($p < 0.01$ for both). The CCC for our best performance on the test set is shown in **Table 3** alongside baseline test results for comparison. The proposed system outperforms the baseline system (statistically significant; $p \approx 0.01$). Although both the linear regression approach for late fusion used in the baseline system and the proposed approach are linear models, Kalman filtering has the added benefit of tracking an affect dimension online and continuously by propagating the predicted state in time. Since the dimension of the state being tracked is fixed, choosing different Kalman filters in our conditional framework is possible.

Table 3: Leave-one-subject-out (LOSO) multimodal performance using CCC

Method	Arousal	Valence
Development set		
Baseline [3]	0.820	0.702
Baseline with LOSO	**0.793**	**0.659**
System 1 with LOSO	0.783	0.624
System 2 with LOSO	0.783	0.702
System 3 with LOSO	**0.824**	**0.718**
Test set		
Baseline [3]	0.682	0.638
System 3	**0.703**	**0.681**

On examining the subject-specific performance on the development set, we note that the CCC is the least for those subjects that have significantly lower variance of the gold standard ratings than that of the training partition. Since the state equation parameters are estimated over the training partition, this could lead to using a state dynamics matrix that models fast dynamics instead of a slow dynamics. Finally, less than 1% difference is observed between CCC and CC of the final outputs which indicates accurate estimation of the bias terms. Additional experiments conducted where the System 2 is modified to use Kalman smoothing instead of filtering do not yield significantly different results. This indicates that arousal and valence evolve causally, perhaps evidence of the way in which the annotations were performed to acquire gold standard ratings.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a linear dynamical systems perspective to perform late fusion of arousal and valence predictions from multiple modalities. Unimodal predictions from linear SVR are modeled as noisy observations and used to perform a continuous online tracking of the affect state. Leveraging the inter-correlations between arousal and valence; and with a conditional framework to select among different multimodal Kalman filters during online tracking, we are able to outperform the baseline for predicting arousal and valence respectively. The proposed approach could be extended for other multimodal fusion tasks to track continuous behavioral codes.

In light of evidence of superior performance of LSTM compared to SVR, we intend to use recurrent LSTM neural networks as a part of our future work to improve unimodal performance. Arousal and valence ratings typically have slow dynamics; this can be used to adaptively update the state dynamics matrix based on the past segment of the predicted state instead of assuming these dynamics to be time-invariant.

[2] http://www.quantpsy.org/corrtest/corrtest.htm

[3] all p-values are reported on a two-tailed t-test

5. ACKNOWLEDGMENTS

This work is supported by grants from National Science Foundation and National Institutes of Health.

6. REFERENCES

[1] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition and Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[3] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 - depression, mood, and emotion recognition workshop and challenge," *CoRR*, vol. abs/1605.01600, 2016.

[4] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Avec 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*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 3–8.

[5] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 73–80.

[6] 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*, April 2013, pp. 1–8.

[7] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[8] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

[9] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 41–48.

[10] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3729–3733.

[11] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Multi-scale temporal modeling for dimensional emotion recognition in video," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 11–18.

[12] Y. Gu, E. Postma, and H.-X. Lin, "Vocal emotion recognition with log-gabor filters," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 25–31.

[13] 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.

[14] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[15] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, ser. ACII '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 356–361.

[16] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. S. Narayanan, "The USC creativeit database: A multimodal database of theatrical improvisation," in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, Valletta, Malta, May 2010.

[17] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press, 2015, pp. 688–699.

[18] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Advances in Neural Information Processing Systems 9*. MIT Press, 1996, pp. 281–287.

[19] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 787–800, Oct. 2007.

[20] B. Han, S. Rho, R. Dannenberg, and E. Hwang, *SMERS: Music emotion recognition using support vector regression*, 12 2009, pp. 651–656.

[21] H. Xianyu, X. Li, W. Chen, F. Meng, J. Tian, M. Xu, and L. Cai, "SVR based double-scale regression for dynamic emotion prediction in music," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 549–553.

[22] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, ser. ICMI '06. New York, NY, USA: ACM, 2006, pp. 146–154.

[23] M. A. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous

affective data in likelihood space," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, Aug 2010, pp. 3695–3699.

[24] D. Kulic and E. A. Croft, "Affective state estimation for human x2013;robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, Oct 2007.

[25] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, April 2012.

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

[27] 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 *Proceedings Interspeech*, 2008.

[28] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using Kalman filtering," Dec 2010, pp. 655–660.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April 2016.

[30] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 605–611.

[31] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[32] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition." in *INTERSPEECH*, 2006.

[33] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 341–344.

[34] F. Eyben, S. Buchholz, and N. Braunschweiler, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4009–4012.

[35] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[37] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "Affectaura: An intelligent system for emotional memory," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 849–858.

[38] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak, and S. S. Narayanan, "Sparse representation of electrodermal activity with knowledge-driven dictionaries," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 960–971, 2015.

[39] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[40] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, pp. 1017–1034, 2013.

[41] A. Georgogiannis and V. Digalakis, "Speech emotion recognition using non-linear teager energy based features in noisy environments," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 2045–2049.

[42] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, Apr 1990, pp. 381–384 vol.1.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[44] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[45] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," University of Toronto, Tech. Rep., 1996.

[46] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[47] J. Cohen and P. Cohen, *Applied multiple regressionl correlation analysis for the behavioral sciences.* Lawrence Erlbaum Associates., 1983.