Temporal Bayesian Fusion for Affect Sensing: Combining Video, Audio, and Lexical Modalities

Arman Savran, Houwei Cao, Ani Nenkova, and Ragini Verma

***Abstract*—The affective state of people changes in the course of conversations and these changes are expressed externally in a variety of channels, including facial expressions, voice, and spoken words. Recent advances in automatic sensing of affect, through cues in individual modalities, have been remarkable; yet emotion recognition is far from a solved problem. Recently, researchers have turned their attention to the problem of multimodal affect sensing in the hope that combining different information sources would provide great improvements. However, reported results fall short of the expectations, indicating only modest benefits and occasionally even degradation in performance. We develop temporal Bayesian fusion for continuous real-value estimation of valence, arousal, power, and expectancy dimensions of affect by combining video, audio, and lexical modalities. Our approach provides substantial gains in recognition performance compared to previous work. This is achieved by the use of a powerful temporal prediction model as prior in Bayesian fusion as well as by incorporating uncertainties about the unimodal predictions. The temporal prediction model makes use of time correlations on the affect sequences and employs estimated temporal biases to control the affect estimations at the beginning of conversations. In contrast to other recent methods for combination of modalities our model is simpler, since it does not model relationships between modalities and involves only a few interpretable parameters to be estimated from the training data.**

***Index Terms*—Acoustic, affective computing, arousal, Bayesian fusion, emotion recognition, facial expressions, lexical, multimodal, particle filter, power, speech, temporal fusion, turn-based, valence.**

I. Introduction

PEOPLE rely on disparate and asynchronous cues, including expressions on the face, changes in voice characteristics, and what is being said, to interpret each other's affective states. In stark contrast, automated affect recognizers are incapable of reliably combining information from all modalities. While there have been great advances in single modality affect recognition for face, voice, and words, none has been powerful enough in isolation. So researchers have eagerly turned their attention to multimodal prediction.

Yet so far results reported in the literature have shown only modest benefits from multimodality [1], and a fair number of studies have concluded that multimodal prediction is in fact inferior to that of the best single modality. Moreover, the improvements from modality fusion are much smaller on datasets of natural spontaneous emotions compared to those with acted emotions.

In contrast to these prior findings, we describe a remarkably successful model for combination of modalities. We cast multimodal affect recognition as a temporal Bayesian data fusion problem [2], where each unimodal predictor is modeled as an affect sensor. The proposed approach keeps track of uncertainties about the affect values due to imprecise unimodal affect predictors in time. Therefore, during estimation it uses information from the past as well, filtering noise and partially compensating for asynchronous unimodal predictions.

We develop our method following the dimensional affect theory [3], [4], which has been used extensively in recent work on spontaneous affect. According to this theory, emotion classes are points in a multidimensional space of affect and the location of states changes continuously in time. This representation is particularly suitable for spontaneous natural emotion expression which may be a blend of several discrete emotions or where the possible emotion classes are so numerous that annotation and learning will become impossible.

We present the first application of Bayesian data fusion for affect recognition. It is greatly suited for the task and in contrast to prior work, we are able to report a high percentage improvement over the best unimodal predictor. Moreover, unlike prior systems that posit complex relationships of modalities to be estimated from training data, our model does not involve learning of any relationship between modalities. Our fusion is governed by only a few parameters to model certainty of different channels and to model temporal priors, which can be obtained by appropriate application of simple statistics.

Another contribution of this paper is interpretative analysis. We perform a number of experiments to exactly pinpoint which are the aspects of the problem that we have handled better in the new framework. The key ingredient for success turns out to be the proper modeling of the stochastic affect generating processes as well as the imperfection of

Manuscript received July 2, 2013; revised March 5, 2014, August 16, 2014, and September 29, 2014; accepted September 30, 2014. This work was supported by NIH Clinical Center under Grant R01-MH-073174. This paper was recommended by Associate Editor M. Pantic. This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a PDF file containing some examples to illustrate the method in the paper. This material is 322 KB in size.

A. Savran, H. Cao, and R. Verma are with the Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: arman.savran@uphs.upenn.edu).

A. Nenkova is with the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2362101

unimodal predictions for optimal Bayesian fusion. In addition, we show that strong temporal bias is present in the annotation of certain affect dimensions, which can be easily modeled in our probabilistic framework, greatly improving predictive power. Moreover, we model the imperfection of single modalities in two layers, including how certain each modality is about its prediction and how certain we are about that certainty, which helps filter out unreliable predictions.

II. Prior Work

Automated affect recognition has been studied in different disciplines and sub-fields, each primarily interested in a given subset of modalities [3], [5], [6]. The visual modality is the most studied one, with a particular focus on facial expressions, but also including body movement, posture, and gesture sub-modalities. The second most studied modality is voice. Examples of other modalities, like text, EEG and other bio-signals, eye gaze, and context are relatively rare.

The literature on automatic emotion recognition independently done with facial expressions and voice is abundant. For facial expression analysis, almost all prior work [7] deals with categorical emotion descriptions or the detection of facial action units [8] used as intermediate representation for emotion analysis. There are three widely adopted approaches to extract information related to facial expression: detection and tracking of facial feature points [9], fitting face models to model shape, and/or appearance such as active appearance models (AAMs) [10], and image analysis by basis functions like Gabor wavelets [11] or by texture descriptors like local binary patterns (LBP) [12].

For acoustic analysis of affect, the most widely discussed features are those related to prosody, including pitch, intensity, and duration. Most recent studies have shown that spectral features such as MFCCs as well as voice quality features [13], [14] also effectively capture emotional expressions. Lexical information, represented as bag-of-words (BOW), term-frequency, and n-gram language models, has also been found helpful [15].

In contrast, multimodal affect recognition is limited compared to the richness of single modality studies. The most common modality combination has been face + voice, followed by trimodal combination of face + voice + posture and gesture. Those multimodal systems, as well as other minor multimodal combinations such as voice + text or EEG + biosignals, have been quantitatively analysed recently by D'Mello and Kory [1]. Their comprehensive meta-analysis of 30 multimodal studies has revealed that there have been consistent but modest improvement due to multimodality compared to the accuracies of the best unimodal recognizers. They report that the average relative improvement resulting from a multimodal prediction over the best unimodal recognizer is 8.12%. There is great variation. While some authors have reported impressive multimodal gains between 16% and 27%, others have reported negligible improvements, or even considerable degradations. Overall, much higher improvements have been obtained on acted databases (12.1% average improvement); the improvements on natural databases like the one we work with here have been small (4.39% average improvement) [1].

The most commonly adopted fusion techniques are naive feature-level (in which all features are combined together to learn a classifier) and decision-level fusion (in which a classifier for each modality is trained separately and their predictions are combined by rules) [16]–[20]. A few hybrid fusion strategies combine feature-level and decision-level fusion [21], or decision-level and model-level fusion (in which more than one classifier for each modality is trained and their predictions are combined by rules) [9]. Some methods can be considered as both feature-level and decision-level fusion, for instance as in boosting [22], [23], where each feature is viewed as a weak classifier. For example, fusion by several classification algorithms has been compared and meta-decision trees were found to be superior to Gaussian mixture models, support vector machines, and multilayer perceptrons in [24]. Some approaches have relied on mixture of Gaussian process classification to directly address some of the unique characteristics of the problem such as missing values from some modalities or noisy predictions [25]. Further complexity has been introduced by additionally learning temporal patterns of fusion. For example, string-based fusion, where a string represents a series of certain events in the video and acoustic channels outperforms feature-level fusion [26]. Hidden Markov models (HMMs) have been applied to learn temporal relations between audio and video streams via error weighted semi-coupled HMMs [9], or by concatenated HMMs [27]. Another example of learning temporal patterns is neural networks working with long short-term memory principle [28]. However, direct comparison of system performance is impossible since the employed modalities, extracted features, affect models, and databases vary.

Recently, two multimodal emotion recognition grand challenges took place, AVEC 2011 [29] and AVEC 2012 [12]. They have greatly facilitated progress on multimodal affect prediction as they enable comparison of unimodal and multimodal methods on a fixed affect model and database. Both challenges call for recognition in four affect dimensions, VALENCE, AROUSAL, POWER, and EXPECTANCY. AVEC'11 deals with binary classification task (high/low or negative/positive prediction) on each dimension, AVEC'12 deals with continuous real-value estimation. The same database is used in both challenges. It is composed of recording of subjects during spontaneous dyadic conversations.

Most teams in AVEC'11 developed unimodal systems exploiting either face or voice features alone. Only a handful attempted multimodal prediction and neither reported considerable improvements. Some of the multimodal systems even led to degradation for some affect dimensions. More multimodal methods were developed for AVEC'12. The best performing methods include linear least squares regression [10] where correlations between affect dimensions are incorporated by combining predictors from all the affect dimensions; fuzzy inference [30], which uses various rules derived manually by observations on the dataset including use of the character's emotions; co-HMM fusion [27] that learns temporal patterns for fusion; and particle filter on regression (see [31]) which performs temporal estimation. Of these,

only Nicolle *et al*. [10] and Savran *et al*. [31], the winners of the fully continuous and word-level sub-challenges respectively [32], reported both unimodal and multimodal performances. However, our preliminary work [31] was the only method that demonstrated improvement for all dimensions and with high margin.

Unlike prior temporal fusion studies [9], [26]–[28], our approach does not aim at learning temporal relationships between different modalities. Instead, we model stochastic affect generating processes to be used as strong temporal priors in Bayesian fusion. Avoiding complex relations between modalities also minimizes high risk of over-learning and means easier training with fewer parameters. The Bayesian framework is well-established, with many examples in the data fusion literature on sensor networks, autonomous vehicle control, etc., [2]. Its main advantage is its effectiveness in handling imprecise sensors by temporal estimation. In order to cope with the deficiencies of existing Bayesian methods, such as inability to represent ambiguous and conflicting cases, we also put forward an uncertainty model to capture the uncertain precision of unimodal sensors.

III. Multimodal Affect Database

We use the AVEC 2012 Grand Challenge dataset [12], corresponding to the Solid-SAL part of the SEMAINE database [33], which is composed of interactions between human subjects and emotionally stereotyped characters acted by humans. The characters respond to the emotional state of their conversational partner rather than to what the partner says [33]. There are four characters with unique moods: 1) even-tempered and sensible; 2) happy and outgoing; 3) angry and confrontational; and 4) sad and depressive.

The subjects have conversations with each of the characters in the AVEC challenge, thus there are at least four conversations per subject, exhibiting different combinations of affect states. The total duration of the footage is about 7.5 h with more than 50 000 words. The dataset contains 95 conversations and is divided into training, development, and test subsets with 31, 32, and 32 conversations in each respectively. Every subset contains data from eight subjects. There are no subjects in common in the training and test set, however, some subjects appear in both training and development.

The database is comprised of recordings of video (upper body video), audio, and manual speech transcripts. Video resolution is 780×580 pixels, 8 bits per sample, and 49.979 frames/s. Audio is recorded at 48 kHz with 24 bits per sample. Also, audio and video streams are synchronized with an accuracy of 25 μs.

All frames in the dataset are annotated for the four dimensions of affect: 1) VALENCE; 2) AROUSAL; 3) EXPECTANCY; and 4) POWER. VALENCE indicates if the feeling is positive or negative. AROUSAL quantifies the subject's overall inclination to be active or inactive; EXPECTANCY indicates to what extent the subject is anticipating what is going on or is unaware like in surprise. POWER is the sense of being in a position to direct events versus being at their mercy. Two to eight people annotated each conversation using a tool for real-time annotation as they watch the video recording. The ground-truth label is obtained by taking the average of the individual annotations at each frame. For training and testing of speech-modality-based predictors, the word-level ground-truths are generated by taking the average of the frame labels over each word. The evaluation measure for dimensional affect prediction is the correlation between gold-standard annotations and the predicted values, averaged across all conversations.

IV. Single Modality Representations

We develop affect predictors independently for video, audio, and lexical modalities. All the predictors perform markedly better than the unimodal baselines of the AVEC12 challenge.

For video, we base the predictions on LBP which are extracted in 10×10 uniform blocks over normalized face images (200×200 pixels), as provided in the AVEC 2012 challenge [12]. For normalization, face and eye detection are done by the OpenCV's Viola–Jones detector. The challenge baseline method first removes rightmost and leftmost LBP blocks and then merges the remaining ones in groups of 2×2. For our feature set, we keep all the blocks unmodified, and also use the face and eye coordinates, and then compute temporal statistics over them. Without temporal statistics, the feature vector length is 59 bins $\times$ $(10 \times 10) + 8$ coordinates $= 5908$. After calculating mean and variance statistics for each frame over intervals of five durations (of $\{2^k\}_{k=-1}^{3}$ s) ending at that frame, the total feature length is: $[(5 \times 2)$ temporal $+$ 1 static$] \times 5908 = 64\,988$. Finally, we reduce this high number of features to the most discriminative 200 features (which provides good compromise between performance and training duration) by applying AdaBoost as in [11] after creating binary labels for high and low values (binary: above versus below the mean of the entire training set).

However, training over all frames in the training set is not feasible since there are over total of 500 000 frames. Therefore, we train the video regression only on an average of 5 700 representative frames for each dimension such that affect values are at least half standard deviation away from the mean, where mean and standard deviation are estimated over the entire training set. This is realized by resampling randomly without replacement and with rejection of neighboring frames. We empirically determine, via experimentation on the development dataset, the rate of the resampling as 40/min and rejection neighborhood size as 1 s.

For audio and lexical modalities, we develop predictors operating on speaker turns. To evaluate word-level or frame-level estimation, we simply assign the prediction obtained for the turn as the value for all constituent words and frames. We use the openSMILE toolkit [34] to extract 988 acoustic features corresponding to 19 functional summary statistics of 26 prosodic and spectral low-level descriptors (LLD) and their first-order time derivatives, as shown in Table I. We adopt a BOW representation of each speaker turn for the lexical modality. The feature space is defined by the words that appear at least three times in the training set. There are 1048 such words. Each turn is represented by a vector of length 1048, each component corresponding to one of the words. The value

TABLE I
VOICE FEATURES: LLD AND FUNCTIONALS

LLD (26)	Functionals (19)
intensity, loudness	max, min, mean, standard dev.
F0, F0 envelope	skewness, kurtosis
probability of voicing	value/range/relative position of extremes
zero-crossing rate	offset/slope/linear err./quad. err. of linear regress.
MFCC 1–12, LSF 1–8	quartile 1–3, 3 inter-quartile ranges

of the component is one if the corresponding word occurs in the speaker turn, and zero if the word does not appear in the utterance. Many studies have shown the advantage of exploiting both acoustic and lexical information [35]–[37]. An in-depth comparison for different acoustic and lexical representations specifically for the AVEC12 dataset can be found in [38].

We use support vector machine regression (SVR) for all unimodal predictors. For video, SVRs are trained on the training set and optimized according the average correlation coefficients across conversations in the development set. For audio and lexical predictors, we combine training and development sets for training, since increasing training set size considerably improved the performance for speech modalities. For audio, SVR parameter search is performed via subject independent cross validation cross-validation on the training and development set.

In Table II, we compare our single modality correlation performances on the AVEC challenge test set [12], listed along the official baseline method. For video, we do frame-level fully continuous predictions, while word-level predictions are evaluated for the speech modality. To compare the modalities on speech regions, where all three are available, we also evaluate word-level performances of video by using the mean value of the frame-level predictions over the words.

We see in Table II, that our video correlation performances are substantially higher than the baseline video for every affect dimension. For frame level fully continuous predictions, it is twice as high as the baseline on average, and for word level, the improvement is about 50%; we observe slightly lower score for word level POWER.

The performance of our acoustic and lexical predictors is also higher than that of the baseline acoustic predictor. Our turn-level acoustic features achieve average correlation score of 0.099, which is higher than that of the word-level baseline results of 0.081 reported in [12].[1] These results confirm the superiority of the larger units for continuous affect recognition, which has also been discussed in [39].

Finally, our BOW lexical representation shows remarkably strong prediction power. It achieves average score of 0.178, which is more than double improvement compared with the baseline acoustic features.

According to the word-level evaluation, video is the best among the three modalities for AROUSAL and VALENCE; and lexical is the best for EXPECTANCY and POWER. To test

[1]Reference [12] reports baseline results as average of absolute values of the correlation coefficients, hence obtains higher scores, especially for audio. The advantage of our turn-level acoustic features is more substantial when we compared it with the actual averaged correlation score of 0.027 achieved with the baseline word-level acoustic features.

TABLE II
UNIMODAL PREDICTORS AND SVR FUSION VERSUS THE OFFICIAL BASELINE [12] ON THE TEST SET VIA AVERAGE CORRELATION COEFFICIENTS[1]

Method	Arousal	Expect.	Power	Valence	Avg.
Fully Continuous - Video					
Baseline (LBP) [12]	0.082	0.108	0.049	0.125	0.093
Temporal LBP	0.256	0.177	0.126	0.248	0.202
Word Level - Video					
Baseline (LBP) [12]	0.091	0.114	0.121	0.143	0.117
Temporal LBP	**0.198**	0.139	0.102	**0.256**	**0.174**
Word Level - Speech					
Baseline Acoustic [12]	0.119	0.075	0.056	0.076	0.081
Turn-based Acoustic	0.104	0.112	0.054	0.122	0.099
Turn-based Lexical	0.092	**0.268**	**0.152**	0.201	0.178
Word Level - Fusion					
SVR Fusion	0.153	0.253	0.053	0.273	0.183

simple learning-based fusion, we applied SVR with linear kernel using the unimodal prediction outputs as features. It is trained on the development set and search for capacity parameter is done by fourfold subject-independent cross-validation. Table II shows its word-level performances on the test set. We see that SVR fusion improves VALENCE, from 0.256 (video) to 0.273 correlation score. On the other hand, it degrades the performance for other affect dimensions.

V. BAYESIAN FUSION

We apply probabilistic inference to fuse continuous decisions made by video, audio, and lexical predictors. The inference is based on Bayesian filtering: first, we make a prediction about the current affect state given only past observations, then update the belief state by current observations.

Here, we show the derivation of the Bayesian inference update equation by treating affect as a Markov process of order K. Let x_t be the affect state at discrete time t (at video frame t of the conversation) that we aim to estimate, and the video, audio, and lexical predictions at time t be the elements of the measurement vector $\mathbf{z}_t$. The posterior distribution of x_t given the present and all the past measurements from single modality affect predictions, $\mathbf{Z}_t$, is obtained using Bayes' rule

$$p(x_t|\mathbf{Z}_t) = \frac{p(\mathbf{z}_t|x_t, \mathbf{Z}_{t-1}).p(x_t|\mathbf{Z}_{t-1})}{p(\mathbf{z}_t|\mathbf{Z}_{t-1})} \tag{1}$$

$$= \frac{p(\mathbf{z}_t|x_t).p(x_t|\mathbf{Z}_{t-1})}{p(\mathbf{z}_t|\mathbf{Z}_{t-1})} \quad \text{(cond. indep.)} \tag{2}$$

$$\propto \underbrace{p(\mathbf{z}_t|x_t)}_{\text{fusion likelihood}} . \underbrace{p(x_t|\mathbf{Z}_{t-1})}_{\text{temporal prediction prior}} . \tag{3}$$

We need to derive an expression for the temporal prediction prior shown in (3). It is obtained by marginalizing for x_t the joint probability of states at different time indices using the product rule and Markov property of order K (see the Appendix for the derivation)

$$p(x_t|\mathbf{Z}_{t-1}) = \int \underbrace{p(x_t|x_{t-K:t-1})}_{\text{transition density}} \prod_{k=1}^{K} p(x_{t-k}|\mathbf{Z}_{t-k})\, dx_{t-K:t-1}. \tag{4}$$

The posterior distribution of the affect state x_t, $p(x_t|\mathbf{Z}_t)$, is recursively obtained according to (1) and (4), that is,

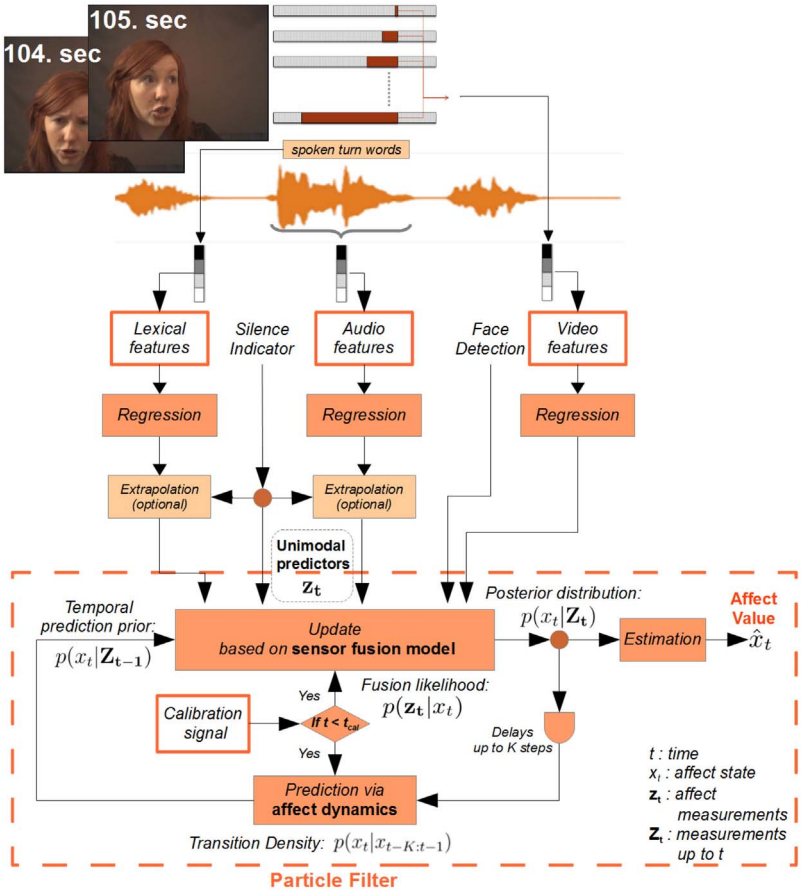


Fig. 1. Flowchart of the temporal Bayesian fusion.

by propagating it in time through a transition density, $p(x_t|x_{t-K:t-1})$, and then by updating with new measurements $\mathbf{z}_t$ from different modalities according to the fusion likelihood $p(\mathbf{z}_t|x_t)$. However, this is only a conceptual solution for Bayesian inference because the integral in (4) is analytically intractable. Nevertheless, an approximate solution can be obtained by means of sequential importance sampling principles. Specifically, we implement this as a particle filter (Section V-E). The overall procedure for multimodal prediction is illustrated as a flowchart in Fig. 1.

In the rest of this paper, we propose specific models for computation of the fusion likelihood and the transition density. For the fusion likelihood, we first model each unimodal predictor as a sensor of affect with fixed precision parameter in Section V-A. In Section V-B, we propose a more sophisticated view of the problem by considering uncertainties on the precision parameters, thus we model sensor precisions not as a single number but as a probability distribution.

In order to determine the appropriate temporal memory K and models for the transition density, we qualitatively identify the stochastic affect generating processes based on linear temporal correlations in Section V-C. There is one more temporal characteristic that we model. Annotations can be biased to some default at the beginning of the conversation. Given the temporal nature of our model, this initial bias can impact the precision of prediction in later stages as well. So in Section V-D, we propose a method to discover and incorporate the initial annotation bias as a trend model.

A. Sensor Fusion With Known Precision

Now, we treat the video, audio, and lexical affect predictors as sensors measuring the affect state x through different modalities. The goal is to combine (fuse) the three predictions in a new, more accurate measurement. In this section and in Section V-B, we focus exclusively on the fusion likelihood part of our Bayesian model and discuss how to compute the posterior in (1), when we assume a noninformative prior for x_t, $p(x_t|\mathbf{Z}_{t-1}) \propto 1$. In Section V-C, we will provide details about the computation of the temporal prediction prior.

First, we assume that each sensor provides noisy measurements with known constant precision λ. That is, we assume that the prediction from a single modality is not accurate but gives some indication where the real-value falls. The distribution of the real value is modeled by a Gaussian with mean equal to the prediction of the single modality and variance inversely proportional to the precision of the sensor. The lower the precision of a modality, the more probable it is that the real affect state is further away from value it predicted.

We now give a formal description of the intuitions above. For sensor m, the measurement noise corresponds to a random variable with Gaussian distribution $N(0, \sigma_m^2)$. The variance σ_m^2 is inversely proportional to the precision of the sensor,

$\sigma_m^2 = 1/\lambda_m$, so the Gaussian is parametrized as $N(0, \lambda_m^{-1})$. We assume the sensors are conditionally independent given the affect state x, hence the likelihood of the measurement vector $\mathbf{z} = [z^{\text{vid}}, z^{\text{aud}}, z^{\text{lex}}]^T$ is

$$p(\mathbf{z}|x) = \prod_m p(z^m|x) = \prod_m N\left(z^m|x, \lambda_m^{-1}\right). \tag{5}$$

Then the posterior—corresponding to the predicted value after fusion of individual modalities—becomes proportional to the product of Gaussians since we assume $p(x_t|\mathbf{Z}_{t-1}) \propto 1$

$$p(x|\mathbf{z}) \propto \prod_m N\left(x|z^m, \lambda_m^{-1}\right). \tag{6}$$

The product of Gaussian densities is unnormalized Gaussian with mean m_F and precision λ_F. Hence, after division with the evidence, the posterior equals to the Gaussian

$$p(x|\mathbf{z}) = N\left(x|m_F, \lambda_F^{-1}\right) \tag{7}$$

$$\lambda_F = \sum_m \lambda_m \tag{8}$$

$$m_F = \frac{\sum_m \lambda_m . z^m}{\lambda_F}. \tag{9}$$

Note that the expected value of the affect state after fusion, $E[x|\{z^m, \lambda_m\}] = m_F$, is a convex summation of the single modality predictions. Simply put, the fusion output is a weighted average of the predictions of individual unimodal affect predictors where weights are proportional to the precisions. On the other hand, the precision of the posterior distribution, λ_F, specifies the certainty that the fusion output for the current frame is correct. Smaller precision corresponds to high certainty and larger precision indicates that even the predicted value after fusion is likely to be incorrect. In the full Bayesian filtering model, where we compute the temporal prediction prior (1) instead of using an uninformative prior, λ_F has important role. It will determine the relative weights of the fusion posterior and the temporal prior.

We now describe the final missing detail about the fusion model, namely how we estimate the precision of each single-modality sensor. Precision parameters are estimated on the development set via a three-step procedure. First, both the predictions from single modalities and the ground-truth affect values are standardized by subtracting the mean and dividing by the standard deviation for all values of that affect dimension on the training set. Standardization of the ground-truth values is performed for each modality separately because video is available all the time but audio/words are not available during regions of silence. The mean and standard deviation from each modality is recorded and is later used for standardization of the predictions in testing.

Next, we calculate mean squared error mse_m between normalized predictions and ground-truth values over the entire development set. Finally, the precision of each modality is defined based on mean squared error as

$$\widehat{\lambda}_m = 1/\text{mse}_m. \tag{10}$$

The corresponding fusion precision is $\widehat{\lambda}_F = \sum_m \widehat{\lambda}_m$.

Relating single-modality precision with mean square error is intuitive, given that the task is to predict real-value affect dimensions. However, the official evaluation measure for the AVEC challenge is the average, over all conversations, correlation coefficient between the estimated affect value and the ground-truth annotations rather than MSE. Therefore, we also experiment with an alternative definition of precision, which combines MSE and correlation performance as

$$\widehat{\lambda} c_m = c.\mu_{\rho_m}, \quad c = \frac{\widehat{\lambda}_F}{\sum_m \mu_{\rho_m}} \tag{11}$$

where μ_{ρ_m} is the average Pearson's correlation between the predictions for sensor m and ground-truth values over the development conversations.[2]

Notice that the precision adjustment made by (11) does not change the overall fusion precision, since $\sum_m \widehat{\lambda} c_m = \widehat{\lambda}_F$. Thus, the certainty in the fusion prediction will still be influenced only by MSE of the single modality sensors.

B. Sensor Fusion With Prior on Precision

Single modality predictors are characterized not only by their precision but also by their consistency of prediction. The inconsistency of a predictor can be high, for instance due to idiosyncratic expressions of affect. To model this additional aspect, we assume the value of the sensor precision is unknown but that we know its probability distribution. The parameters of this distribution will be informed not only by the MSE/correlation for a given modality but also by their variation across different conversations in the development set. In this case, the uncertainty of single modality prediction is defined as

$$p(z^m|x) = \int p(z^m|x, \lambda_m) p(\lambda_m|x) d\lambda_m. \tag{12}$$

Equation (12) gives the marginal likelihood for a unimodal prediction output since we integrate out the unknown precision. The second term inside the integral is the prior density on the precision, λ_m. Notice that since it does not depend on the state x, condition on x is ignored. The first term inside the integral is still modeled as a Gaussian with variance related to the sensor precision.

We use Gamma density with shape parameter α and the rate parameter β as the prior distribution on precision, since it is a maximum entropy distribution of positive continuous random variables. Thus, dropping the sensor index m, (12) becomes

$$p(z|x) = \int N(z|x, \lambda^{-1}) Ga(\lambda|\alpha, \beta) d\lambda \tag{13}$$

$$= \int \left[\frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp(-\frac{1}{2}\lambda(z-x)^2\right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\lambda\beta)\right] d\lambda \tag{14}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int \lambda^{1/2} \lambda^{\alpha-1} \exp\left(-\frac{1}{2}\lambda(z-x)^2 - \lambda\beta\right) d\lambda \tag{15}$$

$$\propto \int \lambda^{a-1/2} \exp\left(-\lambda\left[\frac{(z-x)^2}{2} + \beta\right]\right) d\lambda = I \tag{16}$$

[2]When a given modality m has a negative average correlation, it is ignored in fusion. However, the average correlations for all three were positive.

where $\Gamma(x)$ is the Gamma function. Notice that I is the integral of an unnormalized Gamma density with shape parameter $\alpha' = \alpha + 1/2$ and rate parameter $\beta' = \beta + (z-x)^2/2$. Therefore, I equals to the normalizing constant of the Gamma density, i.e., $\Gamma(\alpha')\beta'^{-\alpha'}$. Hence, substituting I in (15), we obtain

$$p(z|x) = \frac{\beta^{\alpha}\Gamma\left(\alpha+\frac{1}{2}\right)}{\sqrt{2\pi}\Gamma(\alpha)}\left[\frac{1}{\beta+\frac{(z-x)^2}{2}}\right]^{\alpha+\frac{1}{2}}. \quad (17)$$

Then the fusion posterior (when $p(x_t|\mathbf{Z}_{t-1}) \propto 1$), becomes

$$p\left(x|\mathbf{z}\right) \propto \prod_m \left[\frac{1}{\beta_m+\frac{(z^m-x)^2}{2}}\right]^{\alpha_m+\frac{1}{2}}. \quad (18)$$

In Section V-A, we modeled the uncertainty about the sensor measurements via the precision parameter λ_m, i.e., assuming Gaussian density for the measurement noise. Now, we model our uncertainty about the sensors themselves, as well by (18) using the parameters α_m and β_m. First, we make the expected value of the Gamma distributed precision variable equal to $\widehat{\lambda}c_m$ (11). The expected value of the precision for Gamma distribution $Ga(\lambda|\alpha_m, \beta_m)$ is $\alpha_m/\beta_m = \widehat{\lambda}c_m$, $\alpha_m = \beta_m.\widehat{\lambda}c_m$. Parameter β_m controls the inverse scale of the distribution and thus quantifies the degree of uncertainty on the precision, which in turn is determined based on the correlations between predicted values and ground-truth annotations on the development set. Variations of the correlation on different conversations for sensor m, ρ_m, is indicative of the uncertainty. Therefore, we treat ρ_m as a random variable, and relate the value of the beta parameter to its variance, $\sigma^2_{\rho_m}$, after scaling according to the overall precision by c (11), as $\beta_m = 1/(c^2.\sigma^2_{\rho_m})$.

C. Temporal Model of Affect Dynamics

In our complete framework for multimodal affect prediction, we apply sensor fusion in a Bayesian filtering setting where the prior distribution is given by the temporal prediction density (3), instead of using an uninformative prior. Temporal predictions about the affect states follow a temporal prediction density, obtained according to the transition density, $p(x_t|x_{t-K:t-1})$ as expressed in (4).

To determine the appropriate model which best approximates the dynamics of affect, we perform statistical tests and qualitative analysis on the development set. We consider a family of time-series models, called autoregressive integrated moving average (ARIMA) models. An ARIMA(P,D,Q) model is composed of three parts: 1) autoregression part (AR) of degree P; 2) integration part (I) of degree D; and 3) moving average part (MA) of degree Q. The guiding assumption is that linear correlations between the affect values at nearby points in time are sufficient to determine the order of the Markov dependencies (how many past states will influence the current prediction). The estimated model coefficients determine the strength of the dependencies with past states.

The model identification procedure requires estimation of autocorrelation and partial correlation coefficients. We estimate them over all the conversation sequences in the development dataset. However, we ignore the first 30 s of each sequence since those regions are treated separately as explained in Section V-D. We estimate autocorrelations on the ground-truth annotations in the development set. All development conversations are concatenated in a single series of affect annotations for the analysis. The concatenation sites, corresponding to the end of one conversation and the beginning of another, could influence negatively the estimation of model parameters because there no temporal continuity whatsoever is expected. Therefore, we skip the samples (continuous affect annotation corresponding to a video frame) at the concatenation sites so that samples from the consecutive conversation sequences are never used in the same iteration while estimating the autocorrelations. The partial correlation coefficients are estimated from the autocorrelations by fitting autoregressive models using Yule–Walker equations. An autoregressive model of order P is in the form

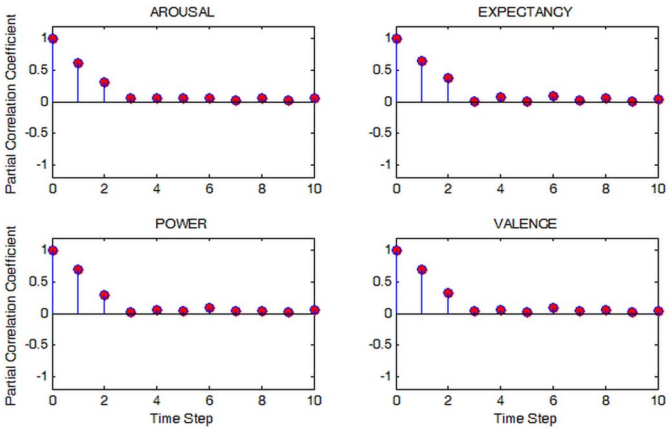


Fig. 2. Partial correlation coefficients after differencing operation on the development set, ignoring the first 30 s of each conversation.

$$y_t = a_0 + \sum_{k=1}^{P} a_k y_{t-k} + v_t \quad (19)$$

where a_k are the autoregressive coefficients and $v_t \sim N(0, \sigma^2)$ is i.i.d. process noise sequence.

The model identification procedure starts with determination of the degree of the integration part, D, of the ARIMA model. For this purpose, we apply unit root test, more specifically augmented Dickey–Fuller [40] test for unit roots, which is one of the commonly used tests in time-series analysis. The test accepted the null-hypothesis of being nonstationary for $D = 0$ and rejected for $D \geq 1$, for each affect dimension. This means that affect generating processes are stationary after first order differencing, i.e., $D = 1$. Therefore, in the rest of the identification procedure all the estimations are done after preprocessing the series with first order differencing in which the time series for analysis is obtained by $y_t = x_t - x_{t-1}$.

In the second step of the identification, we examine the autocorrelation and partial correlation coefficients. In contrast to autocorrelations, partial correlations quantify the correlation between the affect variables with their values at previous time steps after the correlations of all the other time steps are removed. The estimated partial correlation coefficients are shown in Fig. 2. For each affect dimension the temporal

dependency is short, only between a current state and a couple of preceding states. There is gradual decrease of the autocorrelations with more distant states. This analysis clearly indicates that the affect generating process of each dimension does not involve MA component, hence $Q = 0$. On the other hand, partial correlation cut-offs at time step two suggests there are AR components with $P = 2$. Consequently, the affect generating process can be modeled as ARIMA(2,1,0).[3] The process equation is obtained by substituting $y_t = x_t - x_{t-1}$ in (19)

$$x_t = b_0 + \sum_{k=1}^{K} b_k x_{t-k} + v_t \tag{20}$$

where $K = P + D = 2 + 1 = 3$ and

$$\begin{aligned} &b_0 = a_0, \quad b_1 = 1 + a_1, \quad b_K = -a_{K-1} \\ &b_k = a_k - a_{k-1}, \quad \text{for} \quad k = 2, \ldots, K-1. \end{aligned} \tag{21}$$

Hence, we assume the order of Markov dependency is $K = 3$ and the resulting transition density is

$$p\left(x_t | x_{t-K:t-1}\right) = N\left(x_t;\ b_0 + \sum_{k=1}^{K} b_k x_{t-k},\ \sigma^2\right). \tag{22}$$

D. Modeling Trends in the Calibration Phase

Model-driven analysis (explained in detail later in this section) revealed that time-dependent biases exist in the annotations at the beginning of each conversation. Prior work in continuous affect estimation also provide evidence that there is such bias, demonstrating considerably high correlation scores when time features are used for regression, either as scalar time index feature [27] or as step functions [30].

The underlying reasons that can cause those biases are not clear. A sensible explanation might be that during continuous annotation of the affect dimensions, annotators may need some time to accumulate observations before forming a clear impression of the affective states. Annotations are performed in real time, so the annotations of the beginning of each conversation may not fully reflect that true annotator judgment of affect state but be instead biased toward common generally expected states for an affect dimension. We call this initial period at the beginning of each conversation a calibration phase and propose a method to estimate these biases and compensate for them during prediction. Then, we integrate the calibration phase model into our Bayesian estimation scheme by modifying both the temporal prediction and sensor fusion components.

We model the trend in the calibration phase with a logistic sigmoid function parametrized by s_a, s_b, s_c, s_d as

$$s(t) = s_a + \frac{s_b}{1 + \exp^{-(t - s_c)/s_d}}. \tag{23}$$

Equation (23) provides a model of the trend as a smooth transition with four degrees of freedom. s_a and s_c determine the location of the sigmoid on the affect value and time axis, respectively; s_b determines the affect value departure from the start of the conversation due to the trend; and s_d gives the duration of this trend. Therefore, the proposed time model is flexible, with smooth evolution.

[3]We also performed Bayesian fusion experiments for alternative P. For $P = 1$, prediction is improved but not as much as for $P = 2$. For $P = 3$, the prediction improvements is the same as for $P = 2$ for the average performance across the four dimensions.

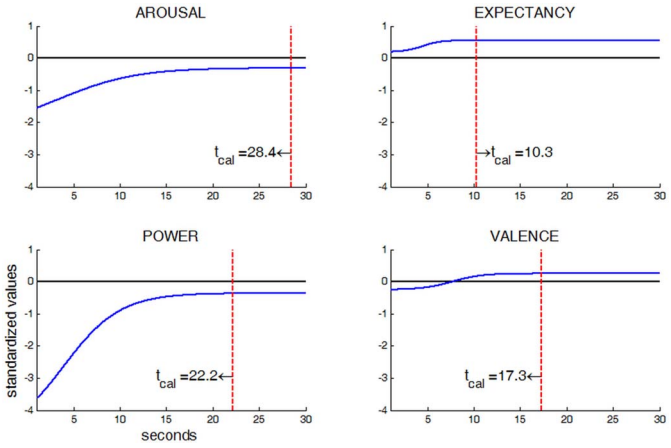


Fig. 3. Fitted logistic functions (on the training set) that model the trend in the first 30 s. t_{cal} is the estimated calibration phase duration [24].

The parameters of the logistic function are estimated by fitting the time values (in units of frames) to all of the standardized affect labels over the first 30 s of conversations from the training set, using nonlinear minimization. We fixed this 30 s duration empirically after trials on the training data over longer durations. The end of calibration is estimated as the point where the sigmoid is assumed to converge (in practice we assume $e^{-6} \approx 0$ for convergence) to the saturation point

$$t_{\text{cal}} = \begin{cases} s_c + 6.s_d, & \text{if } s_b \geq 0 \\ s_c - 6.s_d, & \text{if } s_b < 0 \end{cases}. \tag{24}$$

In Fig. 3, we see the fitted logistic function in the standardized scale of the corresponding affect dimension, i.e., the center is the mean and the unit is in one standard deviation of the data. The estimated calibration durations are marked by the intersection with the vertical dotted red lines. We observe that the progression of the calibration phase is rather different for each affect dimension. For POWER and AROUSAL, the bias is high, several standard deviations of the data. It also appears that their calibration durations are long, 28.4 s for AROUSAL and 22.2 for POWER. POWER has the biggest slope. On the other hand, biases for EXPECTANCY and VALENCE have smaller range and slopes with shorter durations, 10.3 and 17.3 s, respectively.

For temporal prediction in the calibration phase, we again identified the temporal process as ARIMA(2,1,0) process as described in Section V-C. However, the estimation is done separately from the rest of the model, only on the beginning segments up to the sigmoid saturation time points t_{cal} (24). In addition, the trend, i.e., the estimated logistic functions, are subtracted from the reference affect annotation in these segments. The model is in the form

$$x_t = s_t + c_0 + \sum_{k=1}^{K} c_k . \left[x_{t-k} - s_{t-k}\right] + v s_t \tag{25}$$

where c_k are the model coefficients in the calibration phase and $vs_t \sim N(0, \sigma_s^2)$ is i.i.d. process noise sequence.

We also modify the update likelihood in the calibration phase by making it less informative so that time-dependent predictions have more influence on the estimation. Another use of reducing the influence of likelihood would be to mitigate any misleading effect due to unreliable temporal features at the very beginning. Some of the multiscale video features we use are extracted from frames in a sequence of 8 s and these will be undefined at the beginning of interaction.

To implement this modification, we again resort to the use of a sigmoid function. We start with high sensor measurement noise variance (low precision) for all single modality predictors, and a logistic function is used to gradually decrease the variance until we reach the actual estimated values. Very high variance at the beginning yields a broad Gaussian which is similar to uniform uninformative prior in the region of interest. This is realized according to variance scaling function $r(t)$

$$r(t) = A + \frac{1 - A}{1 + \exp^{-(t - t_0)/s_r}}. \tag{26}$$

Here, we set $t_0 = t_{\text{cal}}/2$ and $s_r = t_{\text{cal}}/12$ (assuming $e^{-6} \approx 0$) so that the logistic sigmoid is centered at the estimated calibration phase center and converges to the saturation value, 1.0, when the calibration phase ends, i.e., at time t_{cal}. This scaling function is applied on λ_F (8) to obtain time-dependent precision in the calibration phase

$$\lambda_{Fs}(t) = \frac{1}{r(t)} \lambda_F. \tag{27}$$

We set $A = 100$ to initially have a sufficiently broad Gaussian density as measurement noise, i.e., to generate imprecise sensor models.

E. Particle Filter-Based Estimation

We apply particle filter to realize the conceptual solution of temporal Bayesian inference depicted in (1) and (4). The particle filter implements sequential Monte Carlo simulation using a chosen importance density [41]. Thus, the posterior distribution, $p(x_t|\mathbf{Z}_t)$, is approximated via importance sampling principle by N weighted particles, $\{x_t^i, w_t^i\}_{i=1}^N$ as

$$p\left(x_t|\mathbf{Z}_t\right) \approx \sum_{i=1}^{N} w_t^i \delta\left(x_t - x_t^i\right) \tag{28}$$

where $\delta(x)$ is Dirac delta function and weights w_t^i are the importance sampling weights. These weights are updated at each time step by the fusion likelihood, $w_t^i = p(\mathbf{z}_t|x_t^i, \lambda_{Fs}(t))$, either using (7) or (18).

We use sampling importance resampling (SIR) filter [41], also known as Bayesian Bootstrap filter. The SIR filter uses the transition density, in (4), as the importance density from which samples are drawn. Thus, given the posterior from the previous time steps, the filter first performs N predictions according to the transition density

$$x_t^i \sim p\left(x_t|x_{t-1}^i, x_{t-2}^i, \ldots, x_{t-K}^i\right) \tag{29}$$

which necessitates keeping history of the past K particle sets.

Recursive filtering runs on standardized observations and state variables. Proper initialization of this recursive filtering is important because it can considerably influence the subsequent estimations. Time-dependent biases estimated as described in Section V-D for the initial phase of the affect sequences provide good starting points. Hence, we initialize the particles deterministically according to (23), i.e., $x_{-k}^i = s(-k)$. However, when we test our model without a calibration phase as we do in Section VI-C, initialization is done via a standard normal distribution on the standardized state variables, $x_{-k}^i \sim N(0, 1)$, so that posterior converges quickly after a few steps.

An issue for continuously filtering the single modality outputs is that features for each modality may not be extracted at every point in time. For instance, audio measurements only exist when the subject is speaking; video measurements are not available if the face is not detected. At times, it is even possible that none of the measurements are available. We consider two solutions for this problem. The first is that for each point in time, we use only the available measurements to compute the likelihood of the state. Thus, for each combination of measurements we have different likelihood models. In case there is no observation at all, all the particles have the same uninformative uniform likelihood, $p(\mathbf{z_t}|x_t^i) = 1$, that is, estimations are based only on the predictions. We represent the available unimodal predictor outputs (measurements) at time t with the nominal variable c_t which indicates a mapping of $\mathbf{z_t}$ to a new observation vector $\mathbf{z_t^{c_t}} \in R^{n_{c_t}}$ where n_{c_t} is the number of available measurements.

As another solution, we fill missing measurements for speech modalities with values from preceding speech-turns before applying filtering. This speech modality extrapolation over silence regions assumes that affect predictors of speech can also be good predictors for the following silence regions. However, durations of speech-turns and silence regions vary and can be quite long. Therefore, this may not be an effective solution. In Section VI-C, we also compare the estimations with and without speech extrapolation.

For inference, we apply minimum mean square error estimation (MMSE), which is equivalent to computing expected values over the posteriors. We use 50 particles, as we found experimentally that there is no significant change on the estimations after 50 particles.

VI. Fusion Experiments and Discussion

We extensively evaluate our approach, teasing apart the aspects of the fusion model that contribute most to performance. First, in Section VI-A, we compare Gaussian and marginalized likelihood models in Bayesian fusion together with regression and averaging-based fusion. Then, we show the advantage of extrapolating speech-modality predictions over silence intervals and of calibration predictions in Section VI-C. We investigate the effect of temporal prediction in fusion in Section VI-D. Section VI-E is dedicated to comparison with previous work on the same database. Demonstration of how the proposed marginalized likelihood method can be useful in preventing counter-intuitive

TABLE III
PERFORMANCE COMPARISON OF BAYESIAN FUSION WITH FUSION BY AVERAGING AND SVR ON THE TEST SET. CORRELATION PERFORMANCES ARE LISTED TOGETHER WITH PERCENTAGE OF RELATIVE IMPROVEMENT WITH RESPECT TO BEST SINGLE MODALITY (BEST M.) PERFORMANCE. (E: EQUAL WEIGHTED, P: PRECISION WEIGHTED, C: CORRELATION WEIGHTED, G: GAUSSIAN LIKELIHOOD, AND M: MARGINALIZED LIKELIHOOD)

Fusion	Arousal	Expectancy	Power	Valence	Avg.
Frame Level - Over Speech Turns After 30 sec.					
Best M.	0.192 : vid	0.264 : lex	0.115 : lex/vid	0.272 : vid	0.211
SVR	0.180 : -6%	0.234 : -11%	0.076 : -34%	0.303 : 11%	0.198 : -10%
Aver.-E	0.162 : -16%	0.193 : -27%	0.104 : -10%	0.266 : -2%	0.181 : -14%
Aver.-P	0.157 : -18%	0.195 : -26%	0.113 : -2%	0.259 : -5%	0.181 : -13%
Aver.-C	0.178 : -7%	0.200 : -24%	0.106 : -8%	0.293 : 8%	0.194 : -8%
Bay.-G	0.318 : 66%	0.291 : 10%	0.121 : 5%	0.352 : 29%	0.270 : **28%**
Bay.-M	0.281 : 46%	0.298 : 13%	0.160 : 39%	0.362 : 33%	0.275 : **33%**
Frame Level - After 30 sec.					
Best M.	0.208 : vid	0.229 : lex	0.093 : vid/lex	0.288 : vid	0.205
SVR	0.176 : -15%	0.123 : -46%	0.088 : -5%	0.295 : 2%	0.170 : -16%
Aver.-E	0.177 : -15%	0.146 : -36%	0.065 : -30%	0.267 : -7%	0.164 : -22%
Aver.-P	0.173 : -17%	0.147 : -36%	0.072 : -23%	0.263 : -9%	0.164 : -21%
Aver.-C	0.186 : -11%	0.145 : -37%	0.078 : -16%	0.287 : 0%	0.174 : -16%
Bay.-G	0.282 : 36%	0.273 : 19%	0.115 : 24%	0.320 : 11%	0.248 : **22%**
Bay.-M	0.260 : 25%	0.285 : 24%	0.134 : 44%	0.324 : 13%	0.251 : **27%**
Frame Level - Whole Sequence					
Best M.	0.256 : vid	0.272 : lex	0.173 : lex	0.248 : vid	0.237
SVR	0.260 : 2%	0.183 : -33%	0.042 : -76%	0.247 : 0%	0.183 : -27%
Aver.-E	0.258 : 1%	0.192 : -29%	0.156 : -10%	0.251 : 1%	0.214 : -9%
Aver.-P	0.261 : 2%	0.193 : -29%	0.161 : -7%	0.248 : 0%	0.216 : -9%
Aver.-C	0.275 : 7%	0.191 : -30%	0.155 : -10%	0.263 : 6%	0.221 : -7%
Bay.-G	0.435 : 70%	0.297 : 9%	0.246 : 42%	0.313 : 26%	0.323 : **37%**
Bay.-M	0.402 : 57%	0.295 : 8%	0.259 : 50%	0.306 : 23%	0.315 : **35%**

fusion due to conflicting predictions is given in the Appendix.

A. Evaluation of Bayesian Multimodality Fusion

We evaluate our Bayesian multimodality fusion not only by comparing the SVR fusion described in Section IV but also by comparing with several averaging methods. Recall that weighted averaging is equivalent to performing Bayesian fusion under Gaussian likelihood model and ignoring the temporal prediction prior (9). We consider three types of linear combinations: 1) equal weights; 2) weights proportional to estimated single modality precisions (10); and 3) weights proportional to the mean of the per conversation correlation coefficients (11), as described in Section V-A.

These fusion methods are trained on the development set, however, skipping the first 30 s of each sequence to ensure that time-biases do not affect the fusion (see Section V-D). For the averaging and SVR fusion methods, video regression outputs are directly used over the silence regions rather than fusion, and in case of missing video features the last seen features are employed for prediction. On the other hand, Bayesian fusion inherently handles the missing speech modalities by altering the likelihood functions and by means of temporal predictions (we also evaluate fusion with speech extrapolation in Section VI-C.)

In Table III, we compare the correlation coefficient performances of the fusion methods on the test set. The table also shows relative improvement over the single best modality, equal to the percent of original correlation performance added thanks to the use of modality fusion techniques. Evaluations are done in three settings: 1) over whole conversations; 2) ignoring the first 30 s; and 3) over speech regions after dropping the first 30 s. We omit these beginning durations in order not to confound the evaluations with the annotation time-biases as discussed in Section V-D. Also, since we can have all the modalities only over the speech-turns, evaluations over speech turns provides more accurate comparisons. Video is the best performing modality for AROUSAL and VALENCE, and lexical is the best for EXPECTANCY in all evaluation settings. The lexical modality is the best for POWER in whole sequence evaluations, but it is on par with the video modality in the other two evaluation settings (difference in correlations is less than 0.05).

Fig. 4. Several example frames (cropped faces) from the 25. test clip.

When the first 30 s of the conversation are ignored in testing, most fusion methods for all dimensions perform markedly worse. This remarkable drop may be due to the training of single modality regressions on the whole sequences, which may partly establish mappings onto values dominant over the initial calibration stage.

We see in Table III that temporal Bayesian fusion performs the best for all dimensions for all the evaluation settings, with big performance difference. For frame-level evaluation over speech turns after 30 s, the average gain of temporal Bayesian fusion with respect to the best single modality results is 33%, whereas we observe degradation with other fusion methods: -10% for SVR and -8% for the best performing averaging method which uses the weights proportional to correlation coefficients. Recall that in the Bayesian method, input sensor precisions (or their expected values in case of marginalized likelihood) are also proportional to correlations (11). The poor performance of SVR, may be because of the sometimes very noisy outputs of the single modality regressors (especially audio predictions of POWER), which may make learning of mapping patterns for fusion impossible. On the other hand, averaging and our Bayesian likelihood models are not trained for complex mapping functions at all. The relative differences of single modality and fusion methods in performance with skipping the first 30 s but including silence regions are all similar. In this setting, the gain from temporal Bayesian fusion is 27%. When we evaluate on the whole sequence domain, SVR performances decrease whereas averaging and temporal Bayesian performances increase; the gain from the Bayesian fusion becomes 37%.

Finally, we observe that temporal Bayesian fusion with unknown precision model achieves slightly higher performance on average (omitting the first 30 s) compared to the known precision model. The improvement is much higher with the POWER dimension, but lower on AROUSAL. This higher improvement with POWER can be related to the greater uncertainty on that dimension. The unknown precision model has the advantage of handling conflicting predictions, and in the Appendix, we demonstrate how it can be helpful to reduce big errors due to those conflicting predictions from single modalities.

TABLE IV
FRAME-LEVEL EVALUATIONS OF SPEECH MODALITY EXTRAPOLATION OVER SILENCE REGIONS (EXT.) AND CALIBRATION PREDICTION (CAL.), VIA CORRELATION PERFORMANCES AND PERCENTAGE OF RELATIVE IMPROVEMENT WITH RESPECT TO BEST SINGLE MODALITY (BEST M.)

	Arousal	Expectancy	Power	Valence	Avg.
Effect of Speech Modality Extrapolation over Silence Regions (Ext.) - After 30 sec.					
On Devel Set					
Best M.	0.231 : aud	0.253 : lex	0.077 : vid	0.364 : vid	0.231
No Ext.	0.291 : 26%	0.235 : -7%	0.121 : 57%	0.400 : 10%	0.262 : 21%
Ext.	0.349 : 51%	0.328 : 30%	0.124 : 61%	0.422 : 16%	0.306 : 39%
On Test Set					
Best M.	0.208 : vid	0.229 : lex	0.093 : vid/lex	0.288 : vid	0.205
No Ext.	0.260 : 25%	0.285 : 24%	0.134 : 44%	0.324 : 13%	0.251 : 27%
Ext.	0.263 : 26%	0.295 : 29%	0.150 : 61%	0.362 : 26%	**0.268 : 36%**
Effect of Calibration Prediction (Cal.) with Speech Modality Extrap. - Whole Sequence					
On Devel Set					
Best M.	0.250 : vid	0.252 : lex	0.143 : vid	0.360 : vid	0.251
No Cal.	0.429 : 72%	0.309 : 23%	0.251 : 76%	0.448 : 24%	0.359 : 49%
Cal.	0.517 : 107%	0.316 : 25%	0.585 : 309%	0.452 : 26%	0.467 : 117%
On Test Set					
Best M.	0.256 : vid	0.272 : vid	0.173 : lex	0.248 : vid	0.237
No Cal.	0.410 : 60%	0.320 : 18%	0.289 : 67%	0.356 : 44%	0.344 : 47%
Cal.	0.478 : 87%	0.343 : 26%	0.570 : 229%	0.350 : 41%	**0.435 : 96%**

B. Comparison on Example Sequence

To better understand how fusion performs using the different methods in Table III, we show both single modality and fusion estimation on the same video clip in Fig. 5. All the graphs show the standardized values of the affect states, with red lines for estimations and blue dotted lines for ground-truth. This example is clip 25 from the test set of the AVEC database (clip 125 in the SEMAINE database). It has a duration about 2 min (6253 frames), and several frames are shown in Fig. 4. In Fig. 5,[4] on the left-side, we see video, acoustic, and lexical predictions of AROUSAL which obtains almost zero correlation coefficient, both for frame-level and word-level evaluations. Notice that due to the varying lengths of speech-turns, speech modality predictions are different length segments of constant values. However, on the right-side, we see that all the fusion methods improve the correlation coefficient performance: 0.219 with SVR, 0.221 with averaging, and 0.470 with Bayesian fusion. It seems that for this example video and speech modality predictors are truly complementary. For instance, at 70 s we observe a jump which is predicted by the acoustic sensor to some extent but not by the video predictor. As seen from Fig. 4, while the subject is speaking during that time, we did not observe facial expressions. The example demonstrates the clear superiority of the Bayesian fusion; on it, SVR is generating temporal inconsistencies and averaging is missing the important changes in affect state as the one in 70 s.

C. Incorporating Silence Intervals and Calibration Prediction

Here, we evaluate Bayesian fusion performance by applying speech modality extrapolation over the silence regions and calibration prediction. The top section of Table IV compares with and without speech extrapolation performances both on the development and test sets. In the top section of the table, all evaluations are performed on the parts of the conversations after the first 30 s, in order to remove confounds due to annotation bias at the very beginning of the conversation. We see that speech extrapolation increases the average performance on the development and test sets. The percentage of improvement of prediction due to fusion compared to the best unimodal predictor on the test set rises from 27% to 36%, and on the development set from 21% to 39%. These results show that it is beneficial to extrapolate speech modality predictions over the following silence regions. When extrapolation is done, multimodal fusion likelihood can still be evaluated, otherwise the update stage in Bayesian estimation is realized only by video predictions.

In the bottom section of Table IV, we compare the use the of the calibration prediction on both development and test sets via correlation scores computed over the complete conversations, including the first 30 s. We also perform speech modality extrapolation in this experiment because this improves results as we discussed above. For this reason, the correlation scores on complete conversation evaluations in the bottom section of Table IV is higher than the scores in Table III. We see substantial improvements via calibration prediction, increasing the fusion benefits from 49% to 117% on the development set, and from 47% to 96% on the test set. POWER and to a lesser extent AROUSAL prediction greatly improve. EXPECTANCY and VALENCE prediction improves only slightly. These outcomes are in accordance with the plots in Fig. 3, which points out big differences on affect values depending on time for some of the dimensions. Their prediction consequently improves the performance. Moreover, when we compare rows with title *Ext.* and title *No Cal.*, although they correspond to exactly the same method, we see substantially higher correlation performances with the latter for POWER and AROUSAL. This is mainly due to temporal video features which are also trained with the biased frames, since the latter is whole sequence evaluation.

D. Temporal Aspects in Fusion

So far, we firmly established that Bayesian fusion leads to impressive improvements compared to the SVR and averaging methods. Averaging is equivalent to simple Bayesian fusion with uninformed prior, as explained by (6). However, it remains unclear whether this gain can be attributed to use of temporal prior itself rather than to the combination of modalities. To find out the answer, we design an experiment where we compare video regression, temporal Bayesian inference on video regression, temporal Bayesian fusion, and simple Bayesian fusion [averaging via correlation weights in (11) according to (9)]. The improvements due to temporal estimation are quantified by comparing video regression and its temporal Bayesian inference in case of single modality, and by comparison of simple Bayesian fusion and temporal Bayesian in case of multimodality.

This set of experiments also allows us to assess the improvement in affect prediction due to fusion. When we compare video regression and simple Bayesian fusion, we quantify the gain from fusion without using temporal prior. In contrast, if we compare Bayesian inference on video versus temporal Bayesian fusion, we get a sense of the fusion gain under the same temporal prediction model.

[4]See the supplementary document for predictions in all the dimensions.

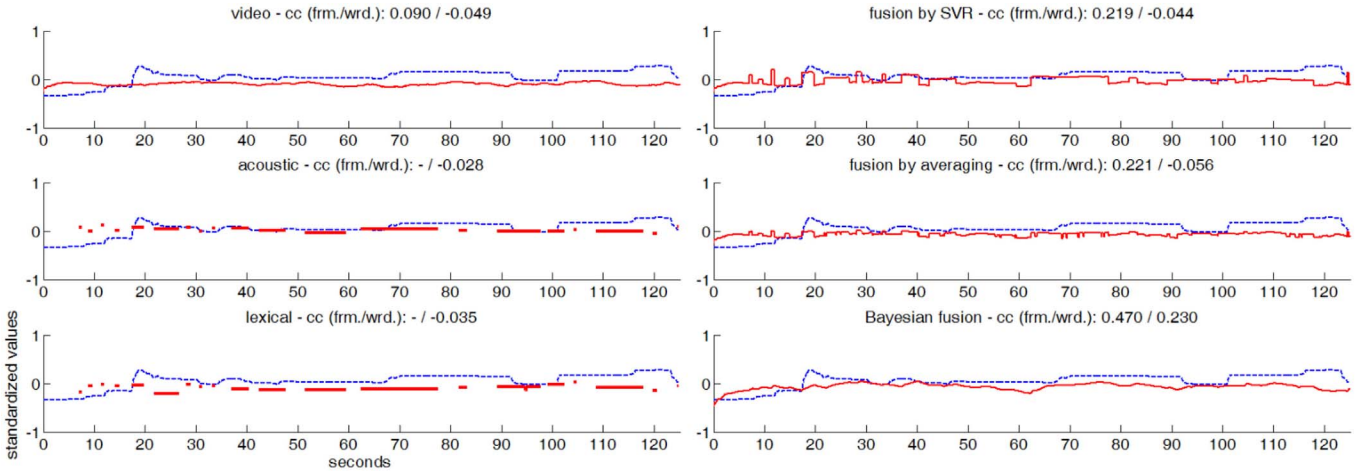


Fig. 5. Comparison of single modality and fusion estimations (truth: dotted-blue, estimation: straight-red) for AROUSAL on the 25. clip in AVEC database. The methods on the left-side from top to bottom are video, acoustic, and lexical modalities, and on the right-side are fusion by SVR, averaging (correlation weighted), and temporal Bayesian fusion (known precision), as given in Table III (*y*-axis: standardized values).

TABLE V
BEST PERFORMING METHODS IN THE AVEC'12 CHALLENGE [12] ARE COMPARED ON THE TEST SET WITH AVERAGE CORRELATION COEFFICIENTS. (.)*: REFLECTS GAIN IN PERFORMANCE DUE TO BOTH MULTIMODAL EFFECT AND MODELING OF TEMPORAL BIAS IN THE CALIBRATION PHASE

	Ours	**Baseline [12]**	**Ozkan et. al. [27]**	**Soladie et. al. [30]**	**Nicolle et. al. [10]**
Performance Results (Fully Continuous / Word Level / Multimodal Effect)					
Arousal	0.48 / 0.36 / 26% (87%)*	0.15 / 0.10 / -13%	0.33 / 0.14 / -	0.42 / - / -	0.64 / - / 20%
Expectancy	0.34 / 0.32 / 29% (26%)*	0.11 / 0.11 / -8%	0.31 / 0.29 / -	0.33 / - / -	0.34 / - / -7%
Power	0.57 / 0.37 / 61% (229%)*	0.14 / 0.07 / -45%	0.45 / 0.29 / -	0.57 / - / -	0.51 / - / 19%
Valence	0.35 / 0.36 / 26% (41%)*	0.15 / 0.11 / -22%	0.18 / 0.17 / -	0.42 / - / -	0.35 / - / -1%
Avg.	**0.44 / 0.35 / 36% (96%)***	**0.14 / 0.10 / -22%**	**0.31 / 0.20 / -**	**0.43 / - / -**	**0.46 / - / 7%**

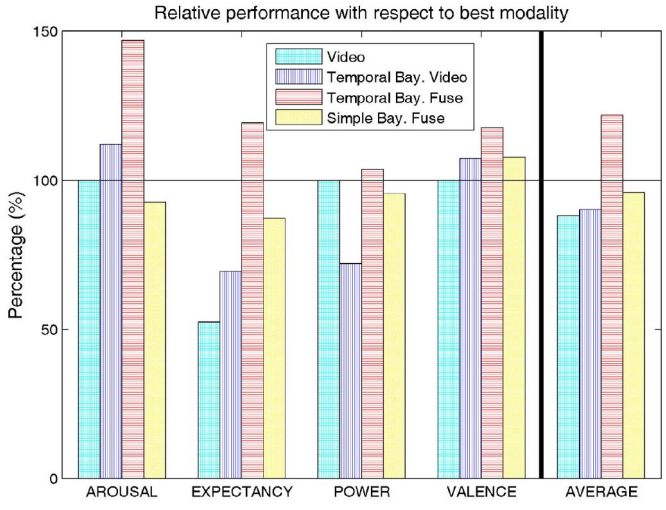


Fig. 6. Comparison of video regression, temporal Bayesian inference on video regression, temporal Bayesian fusion, and simple Bayesian fusion [11]. Evaluation is only over the speech-turns after dropping the first 30 s of the sequences. Bars show the relative percentage with respect to the best of video, audio, and lexical modalities, for each affect dimension.

Fig. 6 shows relative performance percentages with respect to the best among the video, audio, and lexical modalities. Evaluation is done only over the speech-turns, skipping the first 30 s of the conversations on the test set. Here, video is the best modality except for EXPECTANCY and the video bar lengths are at 100% level. The first prominent conclusion supported by this plot is that temporal estimation provides improvements. For the video-only case, Bayesian inference leads to only slight improvement on average, even degrades POWER. In contrast, temporal Bayesian fusion obtains substantial improvements over simple Bayesian fusion (averaging).

Next, we investigate the gains attributable to our fusion in two circumstances. First, we compare the differences in performance without applying temporal estimation, i.e., we look at gains between video regression and simple Bayesian fusion. Second, we compare these with the improvement obtained when temporal estimation is present, i.e., we look at gains between temporal Bayesian inference on video regression and temporal Bayesian fusion. In the latter case fusion improvements are substantial, in contrast to the smaller improvements or even some degradations we witness in the former case. Thus, neither temporal inference nor sensor fusion are responsible for the big improvements alone. It is their combined effect that brings the impressive improvements in our fusion approach.

E. Comparison With Previous Continuous Fusion Methods

Here, we briefly outline the main aspects of competitive approaches that were also evaluated on the same dataset used in this paper.

The best performing methods in the AVEC'12 challenge dataset [32] are shown in Table V. The baseline [12] method obtains 0.136 average absolute correlation over all the four dimensions by using SVR feature fusion. Our fusion method and features perform better than the baseline with very high

TABLE VI
MEAN SQUARE ERROR OF FUSION FOR POWER DIMENSION ON ALL THE TEST SET FRAMES, ON ONLY DETECTED RELIABLE FRAMES, AND ON ALL FRAMES VIA RELIABLE PREDICTION. THE RANGE OF POWER IS [0,1]

Fusion likelihood	All Frames	Only Reliable Frames	Reliable Prediction (All Frames)
Marginalized	0.346	0.328	0.282
Gaussian	0.382	0.305	0.267

margin, with 0.44 correlation score. It is also considerably higher than [27] (0.31) which uses co-HMMs. They employ a few informative video features (smile, gaze, head tilt) and acoustic features as well as a scalar-time index feature to exploit the time-bias.

Soladié *et al.* [30] obtain 0.43 average correlation performance by fuzzy inference. Unlike all other methods, they apply rules rather than training on a dataset. They also make use of the agent's mood, which may greatly contribute to the performance; for instance, when subjects interact with the cheerful agent, they have a tendency to have high valence and arousal, as shown in [30]. As for the other features, they use smile detection and head pose for video, word count, and rate for speech modality. Moreover, they generate response and discourse time features to model the time-biases.

Nicolle *et al*. [10] achieve correlation score of 0.46. Their method also involves use of video, audio, and time modeling. For video, they apply AAM and extract multiscale temporal features as well. They also benefit from delay estimation for more robust feature selection. An important difference of their method is normalization by subject via feature normalization which was shown to provide about 10% improvement [10]. They combine outputs of kernel regression-based unimodal predictors using linear least squares. However, to fuse for each affect variable, they combine unimodal predictions of all the affect dimensions.

In terms of multimodal effect, the baseline method shows a big performance drop by deterioration of -22%. Nicolle *et al.* [10] obtain 7% of average improvement by fusion, which is much lower than our improvement of 36% or of 96% by means of calibration prediction (Table IV). Moreover, our fusion obtains improvement and with high margin for every dimension, unlike [10] where some degradation is observed for EXPECTANCY and VALENCE. We cannot compare the multimodal effect with other methods on the challenge due to the absence of unimodal prediction scores.

VII. CONCLUSION

We developed Bayesian fusion for continuous affect estimation, which combines information coming from video, audio, and lexical modalities. To extract affect information from different modalities, we first design effective features and then train regression-based predictors. Dynamics that provide temporal predictions are also learned from training data. We evaluate our single modality and modality fusion methods for the four dimensions of affect—AROUSAL, EXPECTANCY, POWER and VALENCE—on spontaneous dyadic conversation streams where both arbitrary number and duration of speech-turns and silence intervals occur.

Each modality turns out to be important for affect prediction. Video predictions are most accurate for VALENCE and AROUSAL, lexical is the best for EXPECTANCY and POWER.

We consider each single modality predictor as an affect sensor in a temporal Bayesian fusion framework, and propose two types of sensor fusion models. The first one is based on a conventional precision model which assumes sensors work with known precision, i.e. have Gaussian likelihoods, and the resulting effect is weighted averaging with weights proportional to sensor precisions. The uncertainty after the fusion is inversely proportional to summed sensor precisions. Gaussian likelihoods cause overconfidence on the unimodal predictors and therefore, can result in counter-intuitive fusion if the predictions are conflicting. To deal with this issue, we introduce second level of uncertainty by putting prior on precision following Bayesian approach and have marginalized likelihoods. We showed that it slightly improves average performance, but is especially beneficial for the POWER dimension where the uncertainty on the predictors are considerably higher with conflicting predictions.

Temporal Bayesian fusion is realized by particle filter which recursively updates the posterior using fusion likelihood. We assume higher order Markov dependencies and posteriors from several previous time steps are involved in determining the prediction prior in Bayesian fusion. This prediction is made by modeling the affect generating processes with ARIMA model. We show the first time in the literature that doing fusion via temporal prior is the key factor to achieve improved fusion, which is different from prior work where temporal relationships between modalities are modeled [9], [26], [28].

We showed that turn-based speech predictors are useful in predicting the following silence intervals and improves fusion over the silence regions. Moreover, we develop a method to estimate time-biases in the beginning of affect sequences and integrated it into our Bayesian fusion framework. We observed that, especially, POWER and AROUSAL exhibit considerable bias, which is handled successfully by our trend model.

Comparative assessments demonstrate the benefit of our temporal Bayesian fusion, leading to consistent, large improvements for all affective dimensions and dramatically outperforming competitive alternative fusion approaches. We think that these improvements are mostly because emotional cues from different channels often happen asynchronously or information (features) from some modality may be missing altogether at times. Temporal Bayesian fusion helps to compensate these differences by means of temporal prior. Therefore, an interesting future work would be to apply our approach with other types of unimodal predictors, even including temporal predictors, to see the generalization capability of our Bayesian fusion framework.

APPENDIX

A. Temporal Prediction Density

Temporal prediction density (4) is derived starting with Chapman–Kolmogorov equation on $p(\mathbf{X}_t|\mathbf{Z}_{t-1})$ for all the

previous states, $\mathbf{X}_{t-1}$. In other words, the joint probability of states at different time indices is marginalized for x_t

$$p\left(x_t|\mathbf{Z}_{t-1}\right) = \int_{x_0} \cdots \int_{x_{t-1}} p\left(\mathbf{X}_t|\mathbf{Z}_{t-1}\right) d_{x_{0:t-1}}$$
$$= \int_{\mathbf{X}_{t-1}} p\left(x_t|\mathbf{X}_{t-1}, \mathbf{Z}_{t-1}\right) \prod_{n=0}^{t-1} p\left(x_n|\mathbf{X}_{n-1}, \mathbf{Z}_{t-1}\right) d_{x_{0:t-1}}$$

where chain rule is applied. Due to K-order Markov property, x_t only depends on the previous K states. Thus

$$p\left(x_t|\mathbf{x}_{0:t-1}, \mathbf{Z}_{t-1}\right) = p\left(x_t|\mathbf{x}_{t-K:t-1}\right).$$

Also, since x_n does not depend on the future measurements and is conditionally independent given $\mathbf{Z}_n$

$$p\left(x_n|\mathbf{X}_{n-1}, \mathbf{Z}_{t-1}\right) = p\left(x_n|\mathbf{Z}_n\right) \ , \text{ for } n \leq t-1$$

Hence

$$p\left(x_t|\mathbf{Z}_{t-1}\right) = \int_{\mathbf{x}_{0:t-1}} p\left(x_t|\mathbf{x}_{t-K:t-1}\right) \prod_{n=0}^{t-1} p\left(x_n|\mathbf{Z}_n\right) d_{x_{0:t-1}}$$
$$= \int_{\mathbf{x}_{t-K:t-1}} p\left(x_t|\mathbf{x}_{t-K:t-1}\right) \prod_{n=t-K}^{t-1} p\left(x_n|\mathbf{Z}_n\right)$$
$$\left[\prod_{n=0}^{t-K-1} \int_{\mathbf{x}_n} p\left(x_n|\mathbf{Z}_n\right) d_{x_n}\right] d_{x_{t-K:t-1}}$$
$$= \int p\left(x_t|x_{t-K:t-1}\right) \prod_{k=1}^{K} p\left(x_{t-k}|\mathbf{Z}_{t-k}\right) dx_{t-K:t-1}$$

since the integrals for states up to $t-K$ can be factorized and the bracket is canceled out since probability distributions integrate to one.

B. Preventing Counter-Intuitive Fusion

We have introduced a measure of the confidence in each prediction via the marginalized likelihood (Section V-B). It can be immensely helpful in dealing with conflicting predictions from single modalities. For instance, there might be a case where a predictor outputs low negative value and another high positive value. Taking average due to the Gaussian fusion likelihood would produce a neutral value, which would be quite counter-intuitive since it complies with neither of the unimodal predictors. In those circumstances, it would be better to reject the predictions since they are apparently unreliable. Demonstrations of differences in Gaussian and marginalized likelihoods are available in the supplementary document.

Conflicting predictions can simply be detected by finding the modes of the fusion posterior in (18). Since the modes of the posterior can not be outside the 1-D convex hull of the unimodal predictor outputs, we perform search in the range bounded by minimum and maximum of predictor outputs. If more than one peak is found, predictions are conflicting and we reject. We performed an experiment to demonstrate the benefit of detecting counter-intuitive fusion. We first do fusion using marginalized likelihood and accept the biggest mode on the posterior, thus evaluate the performance over all the frames. Second, we reject if there is more than one mode, and evaluate the performance only over the accepted frames. If unreliable conflicting unimodal predictions would cause high errors in fusion, performance over reliable frames must be better. Furthermore, to see if we can obtain better reliable predictions at conflicting frames, we use the last accepted frame prediction in place of those unreliable ones. We observed that detected conflicting cases were insignificant for AROUSAL, EXPECTANCY, and VALENCE dimensions, but considerable for POWER. Table VI shows all these performance results on the test set for POWER. It also shows the Gaussian likelihood fusion results with the same rejected frames. For this experiment, we use MSE over all the sequence frames as the performance metric since it is a direct measure of error and thus enables direct observation of differences due to the counter-intuitive fusion.

We see in Table VI that MSE is lower for reliable frames than for all frames, for both marginalized and Gaussian likelihoods. Moreover, when we apply reliable prediction, we obtain further reduction on the error. These results show us the benefit of detecting conflicting cases as well as applying reliable predictions from the previous frames. Another interesting results is the difference between Gaussian and marginalized likelihoods. When we compare the likelihoods without reliable frame detection, marginalized likelihood achieve less error. This may be because we use the modes in the marginalized case while averaging takes place for the Gaussian case. More explicitly, using the biggest mode over multimodal posterior provides some robustness compared to taking average of unreliable predictions due to the unimodal Gaussian posterior.

REFERENCES

[1] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proc. ACM Int. Conf. Multimodal Interact.*, New York, NY, USA, 2012, pp. 31–38.

[2] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, Jan. 2013.

[3] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Syn. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[4] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth, "The world of emotion is not two-dimensional," *Psychol. Sci.*, vol. 18, pp. 1050–1057, Dec. 2007.

[5] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[7] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, 2003.

[8] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Salt Lake City, UT, USA: A Human Face, 2002.

[9] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, Feb. 2012.

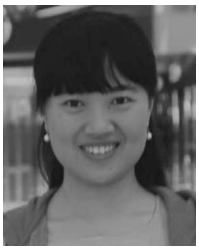
[10] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, New York, NY, USA, 2012, pp. 501–508.

[11] A. Savran, B. Sankur, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image Vis. Comput.*, vol. 30, no. 10, pp. 774–784, Oct. 2012.

[12] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: The continuous audio/visual emotion challenge," in *Proc. ACM Int. Conf. Multimodal Interact.*, New York, NY, USA, 2012, pp. 361–362.
[13] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, pp. 1162–1181, Sep. 2006.
[14] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. Autom. Speech Recognit. Understanding (ASRU)*, Merano, Italy, 2009, pp. 552–557.
[15] B. Schuller, R. Müller, M. K. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 805–808.
[16] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *J. Multimodal User Interf.*, vol. 3, nos. 1–2, pp. 33–48, 2010.
[17] H. Gunes and M. Piccardi, "Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration," in *Proc. Affect. Comput. Intell. Interact. (ACII)*, Beijing, China, 2005, pp. 102–111.
[18] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Cybern.*, vol. 39, no. 1, pp. 64–84, Feb. 2009.
[19] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr./Jun. 2012.
[20] T. Senechal *et al*., "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 993–1005, Aug. 2012.
[21] J. Kim, "Robust speech recognition and understanding," in *Bimodal Emotion Recognition Using Speech and Physiological Changes*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech, 2007, pp. 268–280.
[22] A. Savran, R. Gur, and R. Verma, "Automatic detection of emotion valence on faces using consumer depth cameras," in *Proc. IEEE ICCV Workshop Consum. Depth Cameras Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 75–82.
[23] A. Savran, B. Sankur, and M. T. Bilge, "Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units," *Pattern Recognit.*, vol. 45, no. 2, pp. 767–782, 2012.
[24] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 10–21, Jan./Jun. 2011.
[25] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 677–682.
[26] F. Eyben *et al*., "String-based audiovisual fusion of behavioral events for the assessment of dimensional affect," in *Proc. IEEE Autom. Face Gesture Recognit. Workshop.*, Santa Barbara, CA, USA, 2011, pp. 322–329.
[27] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise emotion recognition using concatenated-HMM," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Santa Monica, CA, USA, 2012, pp. 477–484.
[28] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.
[29] B. Schuller *et al*., "AVEC 2011, the audio/visual emotion challenge," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, Berlin, Germany, 2011, pp. 415–424.
[30] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Santa Monica, CA, USA, 2012, pp. 493–500.
[31] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, New York, NY, USA, 2012, pp. 485–492.
[32] (2012). *AVEC 2012, 2nd International Audio/Visual Emotion Challenge and Workshop* [Online]. Available: http://sspnet.eu/avec2012/
[33] 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 Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan./Mar. 2012.
[34] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, New York, NY, USA, 2010, pp. 1459–1462.
[35] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.
[36] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2004, pp. 577–580.
[37] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: Actors, wizards, and human beings," in *Proc. ISCA Workshop Speech Emot.*, New Castle, U.K., 2000, pp. 195–200.
[38] H. Cao, A. Savran, R. Verma, and A. Nenkova, "Acoustic and lexical representations for affect prediction in spontaneous conversations," *Comput. Speech Lang.*, to be published.
[39] B. Vlasenko, B. Schuller, K. T. Mengistu, G. Rigoll, and A. Wendemuth, "Balancing spoken content adaptation and unit length in the recognition of emotion and interest," in *Proc. Interspeech*, 2008, pp. 805–808.
[40] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *J. Amer. Stat. Assoc.*, vol. 74, no. 1, pp. 427–431, 1979.
[41] M. K. Pitt and N. Shephard, *Sequential Monte Carlo Methods in Practice*, N. de Freitas, A. Doucet, and N. J. Gordon, Eds. New York, NY, USA: Springer, 2001.

Arman Savran received the M.S. and Ph.D. degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey.

He is a Post-Doctoral Researcher with the University of Pennsylvania, Philadelphia, PA, USA. His current research interests include facial expression analysis, 3-D face processing, multimodality fusion, and emotion recognition.



Houwei Cao received the B.E. degree from Shenzhen University, Shenzhen, China, the M.S. degree from the University of Surrey, Guildford, U.K., and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, in 2004, 2005, and 2011, respectively.

She is currently a Post-Doctoral Researcher with the University of Pennsylvania, Philadelphia, PA, USA. Her current research interests include speech and language processing, multilingual and cross-lingual speech recognition, and emotion and affect analysis and recognition.

Ani Nenkova received the Ph.D. degree in computer science from Columbia University, New York, NY, USA.

She was a Post-Doctoral Fellow with Stanford University, Stanford, CA, USA, before joining the University of Pennsylvania (Penn), Philadelphia, PA, USA. She is an Associate Professor of Computer and Information Science with Penn. Her current research interests include automatic text summarization, affect recognition, and text style and quality.

Ragini Verma received the master's degree in mathematics and computer applications and the Ph.D. degree in computer vision and mathematics from the Indian Institute of Technology, Delhi, India.

She is an Associate Professor with the Section of Biomedical Image Analysis (SBIA), Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA. She was a Post-Doctoral Fellow at INRIA, Rhone-Alpes, France, and in medical imaging at SBIA. Her current research interests include diffusion imaging, multimodal connectomics, facial expression analysis, with clinical studies in schizophrenia, autism and brain tumors, as well as projects in animal imaging.