DYNAMIC MULTI-RATER GAUSSIAN MIXTURE REGRESSION INCORPORATING TEMPORAL DEPENDENCIES OF EMOTION UNCERTAINTY USING KALMAN FILTERS

Ting Dang[1, 2], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1, 2]

[1] School of Electrical Engineering and Telecommunications, UNSW, Sydney, Australia
[2] DATA61, CSIRO, Sydney, Australia
ting.dang@student.unsw.edu.au, v.sethu@unsw.edu.au, e.ambikairajah@unsw.edu.au

ABSTRACT

Predicting continuous emotion in terms of affective attributes has mainly been focused on hard labels, which ignored the ambiguity of recognizing certain emotions. This ambiguity may result in high inter-rater variability and in turn causes varying prediction uncertainty with time. Based on the assumption that temporal dependencies occur in the evolution of emotion uncertainty, this paper proposes a dynamic multi-rater Gaussian Mixture Regression (GMR), aiming to obtain the emotion uncertainty prediction reflected by multi-raters by taking into account their temporal dependencies. This framework is achieved by incorporating feedforward and backward Kalman filters into GMR to estimate the time-dependent label distribution that reflects the emotion uncertainty. It also provides the benefits of relaxing the label distribution of Gaussian assumption to that of a Gaussian Mixture Model (GMM). In addition, a new measurement to estimate emotion uncertainty from GMM as the local variability is adopted. Experiments conducted on the RECOLA database reveal that incorporating temporal dependencies is critical for emotion uncertainty prediction with 17% relative improvement for arousal, and that the proposed framework for emotion uncertainty prediction shows potential in conventional emotion attribute prediction.

Index Terms— continuous emotion prediction, inter-rater variability, Kalman filter, uncertainty, Gaussian Mixture Regression, probabilistic uncertainty volume

1. INTRODUCTION

Predicting emotion from speech signals in terms of several affective attributes (i.e. arousal, valence) has attracted increasing interest in the last few decades. Conventional speech based emotion prediction systems aim to develop a regression model that captures the relationship between features extracted from speech and affective attributes. These attributes are generally annotated by several raters and the 'ground truth' is typically assumed to be the average or weighted average among multiple raters. However, discrepancy between raters is ignored though it may carry some informative insights.

Several studies [1-5] have showed the importance of taking information from multiple raters into account. It is claimed that hard labels may not be able to model natural emotion variability [1, 3]. In addition, inter-rater variability represented by the standard deviation among raters has been considered in a multi-task system [4, 5], and was proven to be beneficial for emotion prediction system. Our previous work [6] developed a multi-rater GMR that incorporated multi-rater information to predict emotion uncertainty, under the assumption that multi-ratings reflect the uncertainty of speech frames. However, these methods all assumed that label distribution obtained from multi-raters is a single Gaussian, which may not always be true in reality. Though our work [6] estimated the label distribution as a GMM, final estimation was still carried out by taking the dominant Gaussian mixture component of GMM.

While most studies investigated emotions' evolving nature in terms of hard labels of emotion attributes, only limited literature has taken the temporal dependencies of the emotion uncertainty prediction into account. Long Short-Term Memory-Neural Networks [7, 8] and Output-associate Relevance Vector Machines [9, 10] are the two most widely adopted techniques in emotion prediction that do take into account the emotion temporal dependencies of hard labels. However, they cannot be directly used to explore the temporal dependencies of the emotion uncertainty, since the uncertainty is generally captured by a distribution. Thus exploring the temporal dependencies of emotion uncertainty aims to reveal the evolving process of label distributions.

This paper proposes a dynamic multi-rater GMR taking into account the temporal dependencies of the emotion uncertainty prediction. The main contributions of this paper are: (1) incorporation of both feedforward and backward Kalman filters into multi-rater GMR to account for the temporal dependencies of label distributions; (2) estimating label distribution as a GMM instead of single Gaussian assumption; (3) adoption of a new measurement to estimate uncertainty prediction from GMM by the probabilistic uncertainty volume.

2. RELATED WORK

Only a limited number of papers have considered emotion uncertainty prediction, and even fewer studies have consid-

ered temporal dependencies of emotion uncertainty prediction. Kalman filters are one of the most widely adopted technique in time series analysis [11]. They have been explored as a multi-modal or multi-subsystem fusion technique in emotion prediction in recent years, since they are ideally suited for continuous state tracking. Good performance for predicting arousal and valence was observed [12-14]. However, this was only carried out for hard labels of emotion attributes. The work presented here explores the temporal dependencies of emotion uncertainty using Kalman filters, which are applied to the emotion label distributions instead of hard labels of emotion attributes. The feed-forward and backward Kalman filters are adopted to take into account both the past and future information.

In addition, we assume the emotion label distribution as a GMM instead of single Gaussian distribution, which approaches the problem more realistically. Based on this assumption, a measurement to estimate the uncertainty prediction from GMM is required. It is supposed that a broad GMM indicates a high uncertainty prediction corresponding to high disagreement among multiple raters and vice versa. Therefore the broadness of the GMM referred as probabilistic uncertainty volume (PUV), which measures the local variability of the GMM, is utilized as the uncertainty prediction. The PUV is identical to the probabilistic acoustic volume proposed in [15, 16]. To the best of the author's knowledge, this is the first paper to incorporate the temporal dependencies of emotion uncertainty prediction, and to adopt PUV to estimate the emotion uncertainty.

3. DYNAMIC MULTI-RATER GMR

3.1. Multi-rater GMR

The overall distribution of the inter-rater variability is proven to reflect the uncertainty of speech frames based on our previous multi-rater GMR [6]. It incorporates multi-rater variability in the feature concatenation level, and a GMR is developed to capture the label variability.

In order to obtain the uncertainty prediction for test speech, the conditional distribution $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ of label $\boldsymbol{y}_t$ for each frame t is estimated as a GMM, where λ represents the joint model. An approximation with the dominant mixture component of $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ is adopted, as shown in Figures 1(a) and 1(b). Figure 1(a) displays the ratings from 6 raters of one speech segment. Figure 1(b) shows the prediction $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ approximated as a Gaussian distribution for each frame. This allows for a time-varying indicator of uncertainty prediction as the standard deviation of each frame-wise Gaussian distribution. It is expected that a small standard deviation reflects a low inter-rater variability.

3.2. Incorporating temporal dependencies of uncertainty

Emotion uncertainty is generally captured by a distribution, thus incorporating the temporal dependencies of emotion

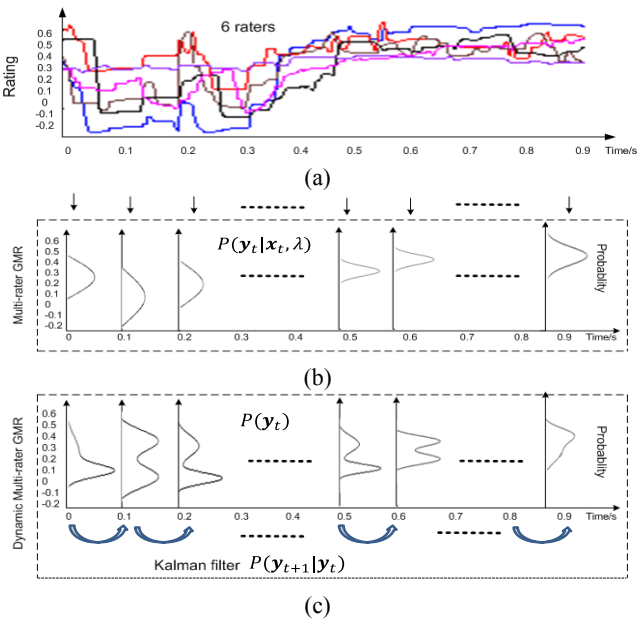


Figure 1: Comparison of multi-rater GMR and dynamic multi-rater GMR; (a) 6 ratings for one speech segment; (b) multi-rater GMR based prediction $\boldsymbol{P(y_t|x_t,\lambda)}$; (c) proposed dynamic multi-rater GMR based prediction $\boldsymbol{P(y_t)}$.

uncertainty is modelled as an evolving process of label distributions $P(\boldsymbol{y}_t)$. Kalman filters are used to estimate the hidden state $P(\boldsymbol{y}_t)$ based on the previous states $P(\boldsymbol{y}_{1:t-1})$ and current observation which is adopted as the predicted conditional distribution $(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$, serving as a noisy observation of $P(\boldsymbol{y}_t)$. This framework also provides the flexibility of the assumption on label distribution. Instead of approximating the $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $P(\boldsymbol{y}_t)$ by Gaussian distribution, the proposed dynamic multi-rater GMR treats $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $P(\boldsymbol{y}_t)$ as GMM. The vector representation $\boldsymbol{v}_t$ of $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $\boldsymbol{s}_t$ of $P(\boldsymbol{y}_t)$ can be generated by concatenating their GMM parameter weights $\overline{w}_{mt}/w_{mt}$, means $\overline{\boldsymbol{u}}_{mt}/\boldsymbol{u}_{mt}$ and vectored covariance $\overline{\boldsymbol{\Sigma}}_{mt}/\boldsymbol{\Sigma}_{mt}$ of each mixture component m respectively:

$$\boldsymbol{v}_t = [\overline{w}_{1t},\cdots\overline{w}_{M_1t},\overline{\boldsymbol{u}}_{1t}^T,\cdots\overline{\boldsymbol{u}}_{M_1t}^T,Vec(\overline{\boldsymbol{\Sigma}}_{1t}),\cdots Vec(\overline{\boldsymbol{\Sigma}}_{M_1t})]^T \quad (1)$$

$$\boldsymbol{s}_t = [w_{1t},\cdots w_{M_2t},\boldsymbol{u}_{1t}^T,\cdots\boldsymbol{u}_{M_2t}^T,Vec(\boldsymbol{\Sigma}_{1t}),\cdots Vec(\boldsymbol{\Sigma}_{M_2t})]^T \quad (2)$$

where M_1 and M_2 represents the number of mixture components for $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $P(\boldsymbol{y}_t)$. Prediction of the hidden states $\boldsymbol{s}_t$ can be formulated as a Kalman filter:

$$P(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) = N(\boldsymbol{s}_t;\ \boldsymbol{F}\boldsymbol{s}_{t-1},\boldsymbol{Q}) \quad (3)$$

$$P(\boldsymbol{v}_t|\boldsymbol{s}_t) = N(\boldsymbol{v}_t;\ \boldsymbol{H}\boldsymbol{s}_t,\boldsymbol{R}) \quad (4)$$

where matrices $\boldsymbol{F},\boldsymbol{H},\boldsymbol{Q}$ and $\boldsymbol{R}$ are the process matrix, observation matrix, process noise covariance and observation noise covariance, which can be estimated during the training phase. The label distribution $\boldsymbol{s}_t$ can be updated sequentially based on equations (3) and (4). An illustration of the proposed dynamic multi-rater GMR is shown in Figure 1(c). The Kalman filter guarantees that the hidden state $\boldsymbol{s}_t$ is dependent on previous states, which reduces the negative ef-

fect of sudden misleading frames. It should be noted that the Kalman filter is utilized to predict the label distribution $\boldsymbol{s}_t$ instead of hard labels. The uncertainty prediction can be estimated based on the label distribution $\boldsymbol{s}_t$, namely, $P(\boldsymbol{y}_t)$.

3.2.1. Training phase

As in [6], a joint GMM $\lambda = P(\boldsymbol{x}, \boldsymbol{y})$ is developed and the prediction $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)$ is estimated using validation partition. Here $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)$, represented as $\boldsymbol{v}_t$, is regarded as the noisy observation of the hidden states $\boldsymbol{s}_t$. $\boldsymbol{s}_t$ and $\boldsymbol{v}_t$ are required to train the Kalman matrices $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}$ and $\boldsymbol{R}$. Ideally, $\boldsymbol{s}_t$ can be trained directly using the labels from multiple raters at each frame t. However, there are generally a limited number of raters in existing databases (i.e. 3 or 6), thus it is not reliable to directly train $\boldsymbol{s}_t$ as a GMM. Maximum-a-posterior adaptation is used to obtain $\boldsymbol{s}_t$ based on a Universal Background Model trained using all labels in the training partition. $\boldsymbol{v}_t$ can be obtained by predicting $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)$.

Given $\boldsymbol{v}_t$ and $\boldsymbol{s}_t$, the matrices $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}$ and $\boldsymbol{R}$ of the Kalman filter can be estimated as [17]. As suggested by [14], introducing an internal delay d during estimation of the process matrix $\boldsymbol{F}$ benefits emotion prediction systems since $\boldsymbol{F}$ cannot be an identity matrix. This is owing to the fact that emotion is a slowly changing process where two adjacent frames are extremely similar. Let $\boldsymbol{A} = (\boldsymbol{s}_{1:t-1-d})^T$ and $\boldsymbol{B} = (\boldsymbol{s}_{d+1:t})^T$, $\boldsymbol{F}$ and $\boldsymbol{Q}$ can be estimated as:

$$\boldsymbol{F} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{B} \quad (5)$$

$$\boldsymbol{Q} = cov(\boldsymbol{B} - \boldsymbol{A}\boldsymbol{F}) \quad (6)$$

where λ can be determined experimentally. Similarly, let $\boldsymbol{C} = (\boldsymbol{s}_{1:t})^T$ and $\boldsymbol{D} = (\boldsymbol{v}_{1:t})^T$. $\boldsymbol{H}$ and $\boldsymbol{R}$ can be estimated as:

$$\boldsymbol{H} = (\boldsymbol{C}^T\boldsymbol{C} + \lambda\boldsymbol{I})^{-1}\boldsymbol{C}^T\boldsymbol{D} \quad (7)$$

$$\boldsymbol{R} = cov(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{H}) \quad (8)$$

3.2.2 Test phase

During the test phase, the predicted label distribution $\boldsymbol{v}_t$ is estimated. Initial values of the hidden states $\boldsymbol{s}_0$ and the covariance $\boldsymbol{Q}_0$ are given and the Kalman filter is applied to predict the hidden states $\hat{\boldsymbol{s}}_t$ sequentially given $\boldsymbol{v}_t$ and the Kalman matrices. The algorithm used to estimate hidden states $\hat{\boldsymbol{s}}_t$ can be found in [17]. Finally, the predicted distribution $P(\hat{\boldsymbol{y}}_t)$ can be reconstructed by decomposing $\hat{\boldsymbol{s}}_t$ into GMM parameters, and emotion uncertainty can be obtained.

3.2.3. Feedforward and backward Kalman filters

Since the Kalman filter only considers the temporal dependencies on the past information, two Kalman filters, one trained in the feedforward direction (KF1), and another in the backward direction (KF2) are proposed to consider the temporal dependencies of both past and future information.

During the test phase, the label distribution $\hat{\boldsymbol{s}}_t^{KF1}$ and $\hat{\boldsymbol{s}}_t^{KF2}$ were estimated using KF1 and KF2 respectively. A linear combination of $\hat{\boldsymbol{s}}_t^{KF1}$ and $\hat{\boldsymbol{s}}_t^{KF2}$ is used as the final estimation $\hat{\boldsymbol{s}}_t$ in (9), where the linear coefficient α was determined experimentally in training phase by equation (10).

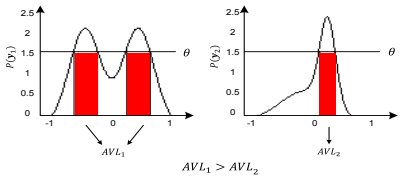


Figure 2: Probabilistic uncertainty volume (***PUV***) of two distributions $\boldsymbol{P}(\boldsymbol{y_1})$ and $\boldsymbol{P}(\boldsymbol{y_2})$. Red area under threshold θ is the ***PUV***.

$$\hat{\boldsymbol{s}}_t = \alpha\boldsymbol{s}_t^{KF1} + (1-\alpha)\boldsymbol{s}_t^{KF2} \quad (9)$$

$$\alpha = \underset{\alpha}{argmin}\|\boldsymbol{s}_t - \hat{\boldsymbol{s}}_t\|_2 \quad (10)$$

3.3. Uncertainty prediction

Uncertainty in predictions are estimated as the probabilistic uncertainty volume $\widehat{PUV}_t$ of $\hat{\boldsymbol{s}}_t$. Figure 2 illustrates the predicted label distribution $P(\hat{\boldsymbol{y}}_1)$ and $P(\hat{\boldsymbol{y}}_2)$ for time t_1 and t_2. Given a threshold θ, the $\widehat{PUV}_t$ of $P(\hat{\boldsymbol{y}}_t)$ is the red area:

$$PUV_t = \int f(\boldsymbol{y})d\boldsymbol{y}, \;\; f(\boldsymbol{y}) = \begin{cases} 1, & P(\boldsymbol{y}_t) > \theta \\ 0, & P(\boldsymbol{y}_t) \le \theta \end{cases} \quad (11)$$

As shown in Figure 2, $\widehat{PUV}_1$ for a broad GMM is larger than $\widehat{PUV}_2$ for a narrow GMM. This is expected to correspond to two frames with high and low inter-rater variability respectively. The inter-rater variability is similarly estimated as the PUV_t of 'ground truth' $P(\boldsymbol{y}_t)$. The correlation between the predicted $\widehat{PUV}_t$ and PUV_t is adopted as the evaluation metric.

4. EXPERIMENTAL RESULTS

4.1. Database

The RECOLA database [18] is a multimodal database in French containing audio, video and physiological signals. Speech data from 18 speakers was equally divided into training and development partitions, which is identical to the partitions used in the Audio-Visual Emotion Recognition Challenge (AV+EC 2016) [19]. The annotation was performed by six raters for arousal and valence.

4.2. Experimental settings

65 low-level descriptors and their first-order derivatives are extracted using Opensmile [20, 21]. Five functionals are used to calculate the statistic features [6]. Dynamic features and labels are calculated as in [22]. PCA is used to conduct dimensionality reduction in the feature space from 650 to 40 dimensions [6]. Delays of 4s for arousal and 2s for valence are applied. GMMs with 2, 4 and 8 full covariance mixture components were tested to model $P(\boldsymbol{y}_t)$ and GMMs with 8 mixture components were found to be the most suitable, consistent with our previous findings, and are used for the joint distribution λ [6].

Internal delays of 1, 3, 5, 7, and 9 seconds have been tested for the Kalman filter. The fusion coefficients α for these filters are tested in the range of [0,1] with a step increase of 0.1. A regularization term for the filters is optimized in the range $[10^{-10}, 10^{5}]$. PUV_t is estimated by sampling 100000 points based on Monte-Carlo approach. The threshold used to estimate probabilistic uncertainty volume of the $P(\boldsymbol{y}_t)$ over the entire test partition, is optimized in the range of [1,99] percentiles with a step increase of 2.

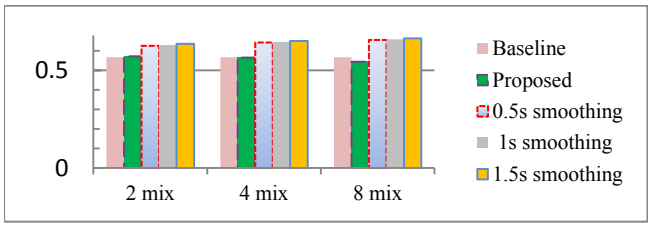
All the experiments are trained and validated using the 9 speakers in the training partitions, and evaluated using the development dataset. Additionally to estimate whether the predicted distribution matches the true distribution, Kullback-Leibler (KL) divergence is numerically estimated as in [23], and the mean and standard deviation of the KL divergence over entire validation partition are reported.

4.3. Analysis of uncertainty prediction

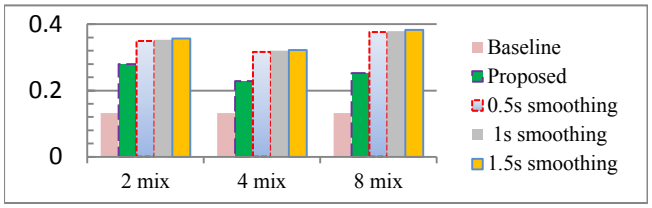
Given the assumption that high inter-rater variability produces a high uncertainty prediction, we aim to investigate the positive correlation between the predicted uncertainty $\widehat{PUV}_t$ computed from Kalman prediction $P(\hat{\boldsymbol{y}}_t)$, and the multi-rater uncertainty PUV_t computed from $P(\boldsymbol{y}_t)$ obtained using test labels only. A moving average filter was used to smooth the uncertainty prediction. It was observed that the 'ground truth' PUV_t is noisy, which may affect the final evaluation. Thus we additionally apply a mean filter with 0.5, 1, and 1.5 seconds to smooth PUV_t but not to over smooth the 'ground truth'. The results are shown in Figure 3.

It can be observed that the proposed method already outperforms the baseline with raw PUV_t, for both arousal and valence, suggesting that incorporating temporal dependencies benefits uncertainty prediction, especially for valence. With the increasing smoothing range, the system performance was further improved. No significant performance difference was observed when using different mixture components to model $P(\boldsymbol{y}_t)$ for arousal, while the model with 8 mixtures outperforms all other configurations for valence, suggesting that predicting valence uncertainty from speech is a more complex problem. Surprisingly, the internal delays of the Kalman filter were not shown to be an influencing factor in uncertainty prediction, which is possibly owing to the complex representations of model parameters. Additionally the KL divergence between ground truth (modelled as a GMM) and predicted label distributions, $P(\hat{\boldsymbol{y}}_t)$, for the proposed systems was compared to that of baseline [6] and the results given in Table 1, indicate that the proposed system leads to more reliable and smoothed distribution prediction.

We also compared the system performance using a single feedforward and bidirectional Kalman filters under the optimal system configurations. Bidirectional Kalman filters showed a slightly better performance of 0.665 over 0.662, and 0.383 over 0.381 for arousal and valence respectively. The optimal fusion coefficient α was found to be 0.5, suggesting an equal influence of each directional filter.



(a) arousal



(b) valence

Figure 3 : Uncertainty prediction performance in terms of correlation coefficient (CC) with x axis indicating the mixture components of $P(\boldsymbol{y}_t)$. Baseline refers to [6]. Proposed means evaluation on raw PUV_t. Smoothing means evaluation on smoothed PUV_t.

Table 1: Comparison of mean and standard deviation (SD) of KL

	Arousal		Valence	
	Proposed	Baseline	Proposed	Baseline
Mean	0.1439	1.6872	0.2085	1.8628
SD	0.1818	7.2714	0.2044	1.1236

In order to investigate the effectiveness of the proposed framework for emotion hard label prediction, the arousal predictions are also estimated from $P(\hat{\boldsymbol{y}}_t)$ by the expectation–maximization algorithm [24]. The performance for arousal prediction achieves 0.70 and 0.43 in terms of CC and Concordance CC respectively, which is calculated between predicted $\hat{\boldsymbol{y}}_t$ and the mean ratings. Though it could not outperform the state-of-the-art arousal prediction system with CCC of 0.796 [19], it still shows potential in predicting emotion attributes without directly using mean ratings. It should also be noted that these measures of CC and CCC compared to mean ratings completely ignore uncertainty in emotion labels.

5. CONCLUSION

This paper proposes a dynamic multi-rater GMR to predict emotion uncertainty by considering the temporal dependencies, which is achieved by applying Kalman filters to the label distributions. The uncertainty predictions are estimated as the probabilistic uncertainty volume of the label distribution. The results indicate a 17% relative improvement in arousal uncertainty prediction by incorporating temporal dynamics. This also doubles the baseline performance for valence uncertainty prediction. As a pioneering study on the temporal dependencies of emotion uncertainty, this paper provides insights into the time-dependent variability introduced by multi-raters. Future work will focus on the non-linear Kalman filters which relax the assumption of the linearly evolving emotion uncertainty.

REFERENCES

[1] E. Mower *et al.*, "Interpreting ambiguous emotional expressions," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1-8: IEEE.

[2] F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters,* vol. 66, pp. 22-30, 2015.

[3] R. Lotfian and C. Busso, "Retrieving Categorical Emotions Using a Probabilistic Framework to Define Preference Learning Samples," in *INTERSPEECH*, 2016, pp. 490-494.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS),* vol. 2, no. 1, p. 6, 2012.

[5] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty," presented at the ACM MM 2017, Mountain View, 2017.

[6] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression," *Proc. Interspeech 2017,* pp. 1248-1252, 2017.

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

[8] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing,* vol. 4, no. 5, pp. 867-881, 2010.

[9] Z. Huang *et al.*, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 41-48: ACM.

[10] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins, "Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 10, no. 3, pp. 461-468, 2016.

[11] M. S. Grewal, "Kalman filtering," in *International Encyclopedia of Statistical Science*: Springer, 2011, pp. 705-708.

[12] K. Somandepalli, R. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. S. Narayanan, "Online Affect Tracking with Multimodal Kalman Filters," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 59-66: ACM.

[13] K. Brady *et al.*, "Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 97-104: ACM.

[14] Z. Huang and J. Epps, "An Investigation of Emotion Dynamics and Kalman Filtering for Speech-based Emotion Prediction," *Proc. Interspeech 2017,* pp. 3301-3305, 2017.

[15] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *INTERSPEECH*, 2014, pp. 1238-1242.

[16] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Communication,* vol. 75, pp. 27-49, 2015.

[17] M. Oveneke, I. Gonzalez, V. Enescu, D. Jiang, and H. Sahli, "Leveraging the Bayesian Filtering Paradigm for Vision-Based Facial Affective State Estimation," *IEEE Transactions on Affective Computing,* 2017.

[18] 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*, 2013, pp. 1-8: IEEE.

[19] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3-10: ACM.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462: ACM.

[21] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[22] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 2288-2291: IEEE.

[23] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models-Analysis and normalisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7522-7526: IEEE.

[24] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.