A NOVEL SEQUENTIAL MONTE CARLO FRAMEWORK FOR PREDICTING AMBIGUOUS EMOTION STATES

Jingyao Wu[⋆], *Ting Dang* [⋆†], *Vidhyasaharan Sethu*[⋆], *Eliathamby Ambikairajah*[⋆]

[⋆] School of Electrical Engineering and Telecommunications, University of New South Wales, Australia
[†] Department of Computer Science and Technology, University of Cambridge, UK

ABSTRACT

When continuous emotion labelling of natural (non-acted) data is desired, it is typically collected from multiple annotators. However, most automatic emotion recognition systems trained on such data ignore disagreement between annotators and only models the average rating, despite the observation that the degree of disagreement would reflect the ambiguity and subtlety in every expression of emotions. In this paper, we propose a novel Sequential Monte Carlo framework that models the perceived emotion as time-varying distributions that allows for ambiguity to be incorporated. Additionally, we present alternative measures that consider both the similarity of prediction to the multiple labels, as well as whether the degree of ambiguity in the prediction and labels. The proposed system was validated on the publicly available RECOLA dataset.

Index Terms— Continuous emotion prediction, inter-rater variability, Sequential Monte Carlo, Gaussian Process, Gaussian Mixture Model

1. INTRODUCTION

In affective computing systems, the complexity and richness of emotions have increasingly led to them being represented using a dimensional representation, with *arousal* and *valence* being the two most commonly employed affect dimensions. Additionally, since human emotion is dynamic, there is also increasing recognition that the response of automatic emotion recognition systems must also be continuous in time [1, 2]. However, this continuity in both affect dimensions as well as time leads to some interesting challenges when attempting to develop *Continuous Emotion Recognition* (CER) systems. Affect labels are typically obtained from multiple annotators, by asking them to record the emotion they perceive as time-varying arousal and valence values. However, ratings from different annotators differ from each other, and furthermore, the degree of disagreement varies over time. Figure 1(a) shows arousal ratings obtained from 6 annotators (dotted lines) corresponding to the same 100 second interval and it is clear that there is significantly greater inter-rater disagreement in the first 30 secs compared to the last 30 secs.

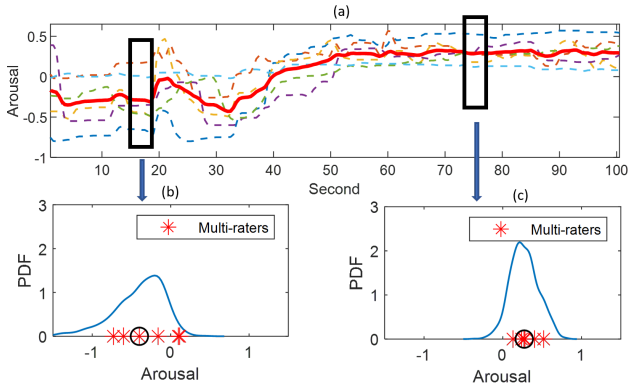

Fig. 1. (a) Arousal ratings from individual raters (dash lines) and the mean rating (solid red line), over a 100s interval from RECOLA; (b) Individual and mean rating and a distribution fit to the rating at an instance with high inter-rater disagreement; (c) Individual and mean rating and a distribution fit to the rating at an instance with low inter-rater disagreement.

In almost all current works on automatic emotion recognition, only the mean rating (solid red line in Fig 1(a)) is considered and inter-rater differences are ignored [2, 3]. However, this is a crude approximation at best and ignores important information about the emotion state, namely, the subtlety of natural emotions that prompted the use of dimensional emotion representations in the first place. Periods of high inter-rater disagreement reflect greater ambiguity in the perceived emotion and can be expected to correspond to subtle expressions of emotion. Conversely, greater agreement amongst the raters can be expected when the expressed emotion is unsubtle and there is little ambiguity in their perception. When ambiguity (inter-rater disagreement) is low, the mean rating would be representative of all individual ratings, however, when ambiguity is high this is not true (see Fig 1(b) and Fig 1(c)).

In this paper, we present two novel contributions that address the modelling of ambiguity in CER systems: (a) we propose a novel approach for modelling both emotion state (arousal and valence), as well as the ambiguity in the state in the context of speech based continuous emotion prediction; and (b) we propose measures to quantify the accuracy of ambiguity aware emotion predictions. The ambiguity aware approach represents arousal and valence as distributions and

the ratings from each annotator are assumed to be samples drawn from this distribution at each point in time. The emotion recognition problem is then treated as the inference of a time-varying distribution based on the input signal (speech in this paper). Within this formulation, the central tendency of the distribution (mean or mode) can be expected to represents the most likely emotion state and the spread of the distributions denotes the level of ambiguity.

Previous studies adopting similar formulations have either assumed the ratings from each annotator to be noisy estimates of an underlying 'true' rating [4], or that the distribution is Gaussian [5, 6], or treated the prediction at each instance to be independent of all other instances [7]. One of the earliest approaches modelling ambiguity in the perception of emotions employed a multi-task learning approach which trains a bidirectional long short-term memory (BLSTM) based emotion recognition system to predict both the mean and standard deviation of annotator ratings [5]. Another approach developed in parallel employed a richer model to represent arousal and valence distributions using Gaussian mixture models but did not model temporal continuity [7]. A later enhancement to this systems added a Kalman filter to model temporal variations of the parameters of arousal/valence distributions as a linear dynamical system [8]. Additionally, non-linear dynamical models using Gaussian Process-Particle Filters have also been proposed and shown to have compelling performances [9]. Finally, it has been proposed that time-varying arousal and valence labels should be treated as Gaussian processes and explicitly modelling the temporal dynamics leads to better models [10]. However, this approach still assumes that the distribution at each time instance is Gaussian, an assumption that was found to be sub-optimal in a recent quantitative analysis of the suitability of a parametric distribution to model arousal/valence distribution [11].

2. PROPOSED SEQUENTIAL MONTE CARLO APPROACH

In the proposed framework, we frame the continuous emotion recognition (CER) problem as a Bayesian tracking problem. Given a sequence of speech features, $\mathbf{x}_{1:t}$, we seek to infer arousal/valence at time t as a probability distribution $P(y_t|\mathbf{x}_{1:t})$, from the previous arousal/valence distribution $P(y_{t-1}|\mathbf{x}_{1:t-1})$ and the current observed features $\mathbf{x}_t$. This is implemented via Sequential Monte Carlo (SMC) approach, using a set of N random samples (particles), $y_t^{(1)}, \ldots, y_t^{(N)}$, and a corresponding set of weights $w_t^{(1)}, \ldots, w_t^{(N)}$ to represent the arousal/valence distributions $P(y_t|\mathbf{x}_{1:t})$. At each time step, the SMC algorithm involves two stages: (i) drawing a set of particles corresponding to the next time step based on a transition model of label dynamics, $P(y_t|y_{1:t-1})$; and (ii) determining the corresponding set of weights based on a suitable observation model of the relationship between observed features and emotion labels, $P(x_t|y_t)$.

2.1. Transition Model

As part of the SMC framework, a model of $P(y_t|y_{1:t-1})$ is implemented based on the assumption that arousal and valence ratings can be modelled as a Gaussian Process (GP) [10]:

$$\tilde{y}_t^{(r)} \sim \mathcal{GP}(m_t, c_{t,t'}), \qquad r = 1, \ldots, R \tag{1}$$

where, $\tilde{y}_t^{(r)}$ denotes the arousal/valence label at time t from annotator r; R is the total number of annotators; $m_t = \frac{1}{R}\sum_r \tilde{y}_t^{(r)}$ denotes the mean of the GP; and $c_{t,t'}$ is the covariance kernel, which is assumed to be a squared exponential kernel, as in [10]:

$$c_{t,t'} = a^2 \exp\left(-\frac{(t-t')^2}{2l^2}\right) \tag{2}$$

with its parameters a and l obtained as maximum likelihood estimation based on annotations from all raters across the training set.

The transition model then assumes that the mean of the label distributions, $\bar{y}_t = E\left[P(y_t|\mathbf{x}_{1:t})\right]$, is also modelled by the same GP, $\bar{y}_t \sim \mathcal{GP}(m_t, c_{t,t'})$, inherently assuming that temporal variations of the distribution mean and the temporal variations of the labels from all the annotators must have the same degree of 'smoothness'. Consequently, at each time step, t, the means of the predicted distributions from the previous time steps, $\bar{y}_{1:t-1}$, are used to obtain an estimate of the current mean, $\bar{y}_t^*$, using Gaussian Process extrapolation. For a detailed discussion of Gaussian processes, the reader is referred to [12]. Note that the previous means, $\bar{y}_{1:t-1}$, would be computed from the weighted particles as per,

$$\bar{y}_k = \sum_{i=1}^{N} w_k^{(i)} y_k^{(i)}, \qquad k = 0, \ldots, t-1 \tag{3}$$

and, we use the ' $*$ ' in the superscript to distinguish between an estimate of the distribution mean obtained from the transition model, $\bar{y}_t^*$, and the distribution mean computed from weighted particles, $\bar{y}_t$, which takes into account the observation model.

Finally, the particles at time t are drawn from a Gaussian distribution,

$$y_t^{(i)} \sim \mathcal{N}(\bar{y}_t^*, \sigma^2) \tag{4}$$

where the variance, σ^2, was empirically chosen to be equal to the highest variance across all annotator ratings in the training set (i.e., the inter-rater variance at the instance corresponding to the highest inter-rater disagreement).

2.2. Observation Model

In the second stage, at each time step of the SMC algorithm, the weights, $w_t^{(1,\ldots,N)}$, corresponding to the particles drawn in the first stage, $y_t^{(1,\ldots,N)}$, are estimated based on

an observation model that links emotion state, y_t, to the observed features, $\mathbf{x}_t$. Specifically, the ideal weights would be $w_t^{(i)} = P(\mathbf{x}_t|y_t^{(i)})$. However, noting that y_t is one dimensional (arousal/valence) while $\mathbf{x}_t$ would typically be a high dimensional vector, any direct attempt to model $P(\mathbf{x_t}|y_t^{(i)})$ from the training data is unlikely to succeed.

Instead, we adopt a two part observation model that uses one Gaussian mixture model, λ_1, to map the observed feature space ($\mathcal{X}$) to the emotion label space ($\tilde{\mathcal{Y}}$); and a second model, λ_2, to estimate the conditional probability of the inferred label given a particle.

Given the observed features, $\mathbf{x}_t$, and the set of M annotator labels, $\tilde{Y}_t = \{\tilde{y}_t^1, \ldots, \tilde{y}_t^M\}$, at each time step, we train a Gaussian mixture model (λ_1) of the joint distribution $P(\tilde{\mathcal{Y}}, \mathcal{X})$. This model is then used to infer an M-dimensional label vector, $\tilde{Y}_t$, given a feature, $\mathbf{x}_t$, via Gaussian mixture regression:

$$\hat{Y}_t = \mathcal{G}_1(\mathbf{x}_t) = E\left[P(\tilde{Y}_t|\mathbf{x}_t, \lambda_1)\right] \tag{5}$$

This now allows the weight, $w_t^{(i)}$, corresponding to the particle, $y_t^{(i)}$, to be set as

$$w_t^{(i)} \triangleq P(\hat{Y}_t|y_t^{(i)}) = P\left(\mathcal{G}_1(\mathbf{x}_t)\middle|y_t^{(i)}, \lambda_2\right) \tag{6}$$

using a second Gaussian mixture model, λ_2, trained to fit all pairwise combinations of $\hat{Y}_t \times \tilde{Y}_t$ to approximate $P(\hat{\mathcal{Y}}, \tilde{\mathcal{Y}})$. Note that both sets $\hat{Y}_t$ and $\tilde{Y}_t$, comprise of M elements each, therefore the set of pairwise combinations, $\hat{Y}_t \times \tilde{Y}_t$ comprises of M^2 elements. This helps ensure that the spread of the distribution given by the particles is representative of the level of disagreement in the ratings.

3. PROPOSED MEASURES FOR EVALUATING PREDICTIONS

Typical measures employed to evaluate the continuous emotion prediction systems, such as Concordance Correlation Coefficient (CCC), Pearson's Correlation Coefficient (CC) and Mean Squared Error (MSE), compare two time series. However, this ignores ambiguity in the emotion labels. Given the aim of continuous prediction of ambiguous emotion states, evaluation measures must account for two principles: (a) when ambiguity in the labels (inter-rater disagreement) is low, the mean rating is representative of the emotional state, and variance of the predicted distribution must also be low and the central tendency of the predicted distribution must be close to the mean rating; and (b) when ambiguity in the labels is high, the mean rating is not representative of the emotion state and difference between the central tendency of the predicted distribution and the mean rating is less important, but the variance of the predicted distribution must be large.

In the proposed SMC framework, the predicted distribution at each time step is represented by a set of particles and their corresponding weights. In the analyses reported in this paper, we take the *mean* ($\bar{y}_t$) as the central tendency of interest and the *standard deviation* ($\bar{s}_t$) as an estimate of the spread of the distribution. The mean can be computed from the particles as per Eq. (3), and the standard deviation (SD) can be computed as,

$$\bar{s}_t = \sqrt{\frac{\sum_{i=1}^{N} w^{(i)}(y_t^{(i)} - \bar{y}_t)^2}{\sum_{i=1}^{N} w^{(i)}}} \tag{7}$$

Finally, to evaluate the predictions we propose the use of the following measures: (i) Concordance correlation coefficient (CCC) between $\bar{s}_t$ and $\tilde{s}_t$, the standard deviation of the emotion ratings. A low CCC indicates that the system is unable to predict the degree of ambiguity; (ii) Mean squared error (MSE) between $\bar{y}_t$ and mean label $\tilde{y}_t$ for different frames partitioned into deciles based on the standard deviation of the labels. The MSE at the lower deciles corresponding to less ambiguous labels can be expected to be significantly more salient than the MSE in the upper deciles (where accurately predicting the ambiguity would be more relevant).

4. EXPERIMENTAL SETUP

The RECOLA dataset [13] is a widely used multimodal corpus. There are 9 five-minute utterances each in the training and development sets, identical to the data partition in the AVEC challenge 2016 [14]. Results are reported on the development set since test labels are not publicly available. The dataset contains continuous emotion annotation from 6 annotators (sampled every 40ms), with continuous arousal and valence ratings between -1 and 1.

The Bag-of-audio-words(BoAW) feature representations are extracted with 100 clusters from 20 dimensional MFCCs using OpenXbow [15]. PCA is applied for dimensionality reduction resulting in 40 dimensional features, conserving 98% of the variability. Their first-order derivatives are computed and concatenated with the original feature vectors, leading to a 80 dimensional feature representation. Delay compensation is applied with 4 seconds for arousal and 2 seconds for valence as in [16]. Emotion changes over 40 ms (sampling rate of labels) intervals are likely to be extremely small, and modelling at that temporal resolution is not likely to be beneficial. Therefore, the labels are downsampled to the window-level representations, by averaging sets of 25 frames (1 second) with 50% overlap between consecutive sets.

The SMC is implemented using the MATLAB toolbox listed in [17]. The number of particles was empirically chosen to be 1000 from the set of [100, 200, 500, 1000, 2000]. The initial particles at time $t = 1$ are drawn from a normal distribution $\mathcal{N}(0, 1)$. The variance, σ^2, of the proposal distribution given in Eq (4) was determined to be 0.25 and 0.2025 for arousal and valence respectively, based on the highest variance in the labels within the training partition. GP is im-

plemented using GPML toolbox [12], with its hyperparameters automatically optimized by maximizing the marginal log-likelihood. GMR is implemented using MATLAB toolbox. 8-mixture and 4-mixture GMMs are used for λ_1 and λ_2 respectively, with full covariance matrix for both.

5. RESULTS AND DISCUSSION

As outlined in section 3, two aspects of the emotion prediction system needs to be evaluated - its ability to predict the level of ambiguity in the annotations, and its ability to predict the emotion state represented by the mean label accurately when ambiguity is low. We first analysed the CCC between the predicted $\bar{s}_t$ and standard deviation of the emotion ratings $\tilde{s}_t$ for all frames within the test set, as shown in Table 1. Table 1 also presents the results of the only two comparable analyses that have been reported in literature till date to the authors' knowledge. Overall, the results in Table 1 demonstrate the effectiveness of the proposed SMC in capturing the emotion ambiguity. Given the evidence that arousal is better modelled using audio and valence using video [16], the following analysis is focusing on arousal prediction.

An example of the prediction in Fig. 2 shows the predicted distribution along with the six annotations. The grey shaded areas indicate the range of values corresponding to $\pm 1.5\bar{s}_t$ from the mean of predicted distribution, and the six colored lines show the annotator labels. The predicted emotion ambiguity matches the inter-rater variability, with regions of high predicted ambiguity corresponding to high inter-rater disagreement and vice versa.

Fig. 3 shows the MSE between the means of the predicted distribution, $\bar{y}_t$, and the label means, $\tilde{y}_t$, as discussed in Section 3. The range of label SDs in each decile is indicated on the x-axis, and the MSE within the decile is represented by the red cross. Additionally, the cumulative MSE computed from SD=0.0 to the maximum SD in each decile is also displayed by the bars. It clearly shows that MSE increases as the emotion ambiguity becomes larger, suggesting better arousal predictions when inter-rater agreement is high (low SD) and vice versa. These results confirm that the proposed system can predict the mean of less ambiguous labels accurately, while also providing a rich prediction that can predict the level of ambiguity (both high and low).

Finally, we also computed the CCC between predicted distribution mean $\bar{y}_t$ and ground truth mean $\tilde{y}_t$, which turned out to be 0.702 for arousal and 0.391 for valence. This conventional measure can be compared to those reported in literature, for instance 0.783 for arousal and 0.495 for valence in [5]. However, it should be noted that this completely ignores ambiguity.

Table 1. Concordance Correlation Coefficient (CCC) and Pearson's Correlation Coefficient (CC) measure between predicted SD ($\bar{s}_t$) and the SD of the 6 annotation ($\tilde{s}_t$)

	Arousal		Valence	
	CCC	CC	CCC	CC
BLSTM [18]	0.103	-	0.075	-
GMR [7]	-	0.568	-	0.132
Proposed SMC	**0.403**	**0.456**	**0.195**	**0.201**

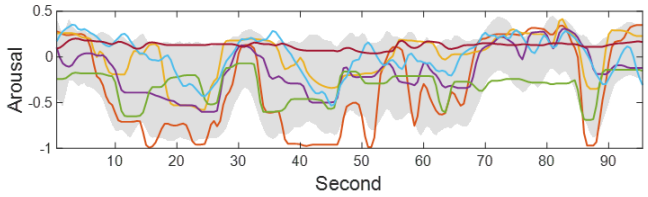

Fig. 2. Arousal ratings from 6 annotators (colored lines) over a 100 second interval. The grey shaded area shows the predicted ambiguity, depicting $\pm 1.5\bar{s}_t$ from the mean of predicted distributions.

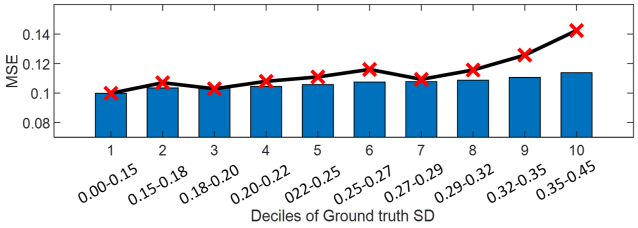

Fig. 3. Mean Squared Error (MSE) at regions with different ambiguity levels determined by deciles of label SD. The SD range corresponding to each decile is shown in the x-axis. The black line depicts the MSE within each decile and bar plot shows the MSE computed over all frames accumulated up to the indicated decile.

6. CONCLUSION

This paper presents a novel ambiguity aware emotion prediction framework that models time-varying emotion state (*arousal* and *valence*) as well as the ambiguity in the perceived emotion. The experiments reported in this paper demonstrate that the inter-rater differences in emotion annotations reflecting ambiguity in the state can be represented as a series of distributions. Furthermore, the paper also illustrates how a Sequential Monte Carlo framework can be employed as a non-parametric, non-linear dynamical model for predicting these ambiguous emotion states. Experimental validation shows that the proposed framework is able to track the level of ambiguity in the labels over time, and predict the emotion state accurately within regions of low ambiguity. This study paves the way for a more nuanced approach to emotion prediction that takes into account the ambiguous and subtle nature of realistic emotion expression.

7. REFERENCES

[1] Hatice Gunes and Maja Pantic, "Automatic, dimensional and continuous emotion recognition," *Interna-*

tional Journal of Synthetic Emotions (IJSE), vol. 1, no. 1, pp. 68–99, 2010.

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

[3] Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[4] Md Nasir, Brian Baucom, Panayiotis Georgiou, and Shrikanth Narayanan, "Redundancy analysis of behavioral coding for couples therapy and improved estimation of behavior from noisy annotations," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1886–1890.

[5] Zixing Zhang, Jing Han, Eduardo Coutinho, and Björn Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1289–1301, 2018.

[6] Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps, "Using gaussian processes with lstm neural networks to predict continuous-time, dimensional emotion in ambiguous speech," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 718–724.

[7] Ting Dang, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah, "An investigation of emotion prediction uncertainty using gaussian mixture regression.," in *INTERSPEECH*, 2017, pp. 1248–1252.

[8] Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, "Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4929–4933.

[9] Konstantin Markov, Tomoko Matsui, Francois Septier, and Gareth Peters, "Dynamic speech emotion recognition with state-space models," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2077–2081.

[10] Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps, "Demonstrating and modelling systematic time-varying annotator disagreement in continuous emotion annotation.," in *Interspeech*, 2018, pp. 3668–3672.

[11] Deboshree Bose, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, "Parametric distributions to model numerical emotion labels," *Proc. Interspeech 2021*, pp. 4498–4502, 2021.

[12] Carl Edward Rasmussen and Hannes Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *The Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

[13] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[14] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "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.

[15] Maximilian Schmitt and Björn Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," 2017.

[16] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps, "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.

[17] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[18] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 890–897.