A Novel Sequential Monte Carlo Framework For Predicting Ambiguous Emotion State

Presenter: Jingyao Wu

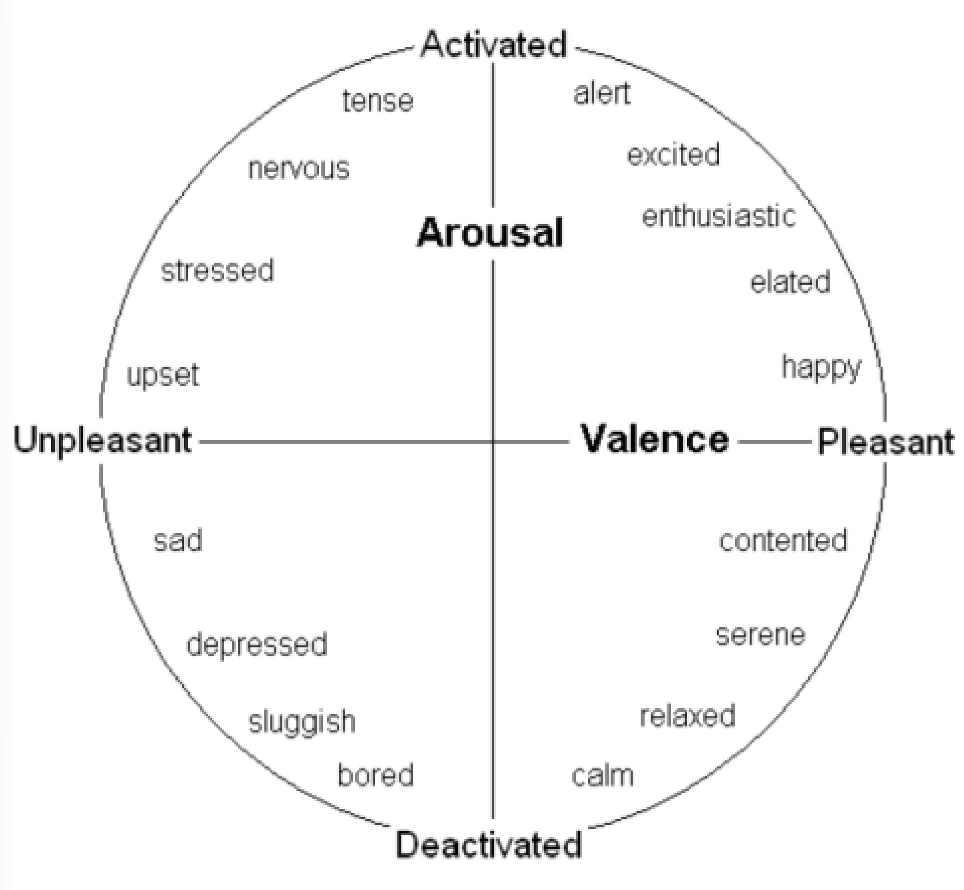
Authors: Jingyao Wu*, Ting Dang*^ , Vidhyasaharan Sethu*, Eliathamby Ambikairajah*
*School of Electrical Engineering and Telecommunications, University of New South Wales, Australia
^Department of Computer Science and Technology, University of Cambridge, UK

ICASSP 2022

Emotion Representations

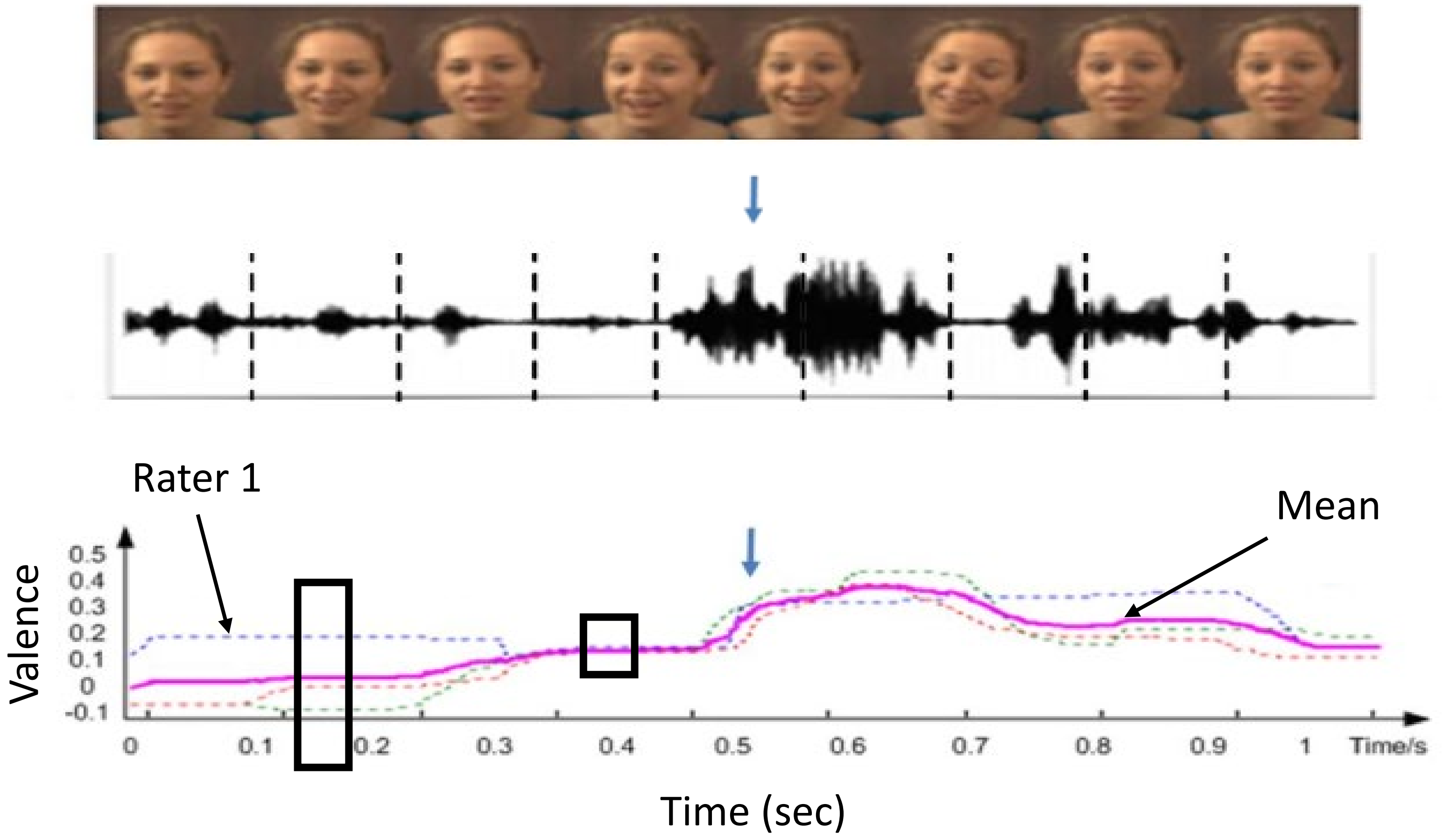
Categorical Representation

Dimensional Representation



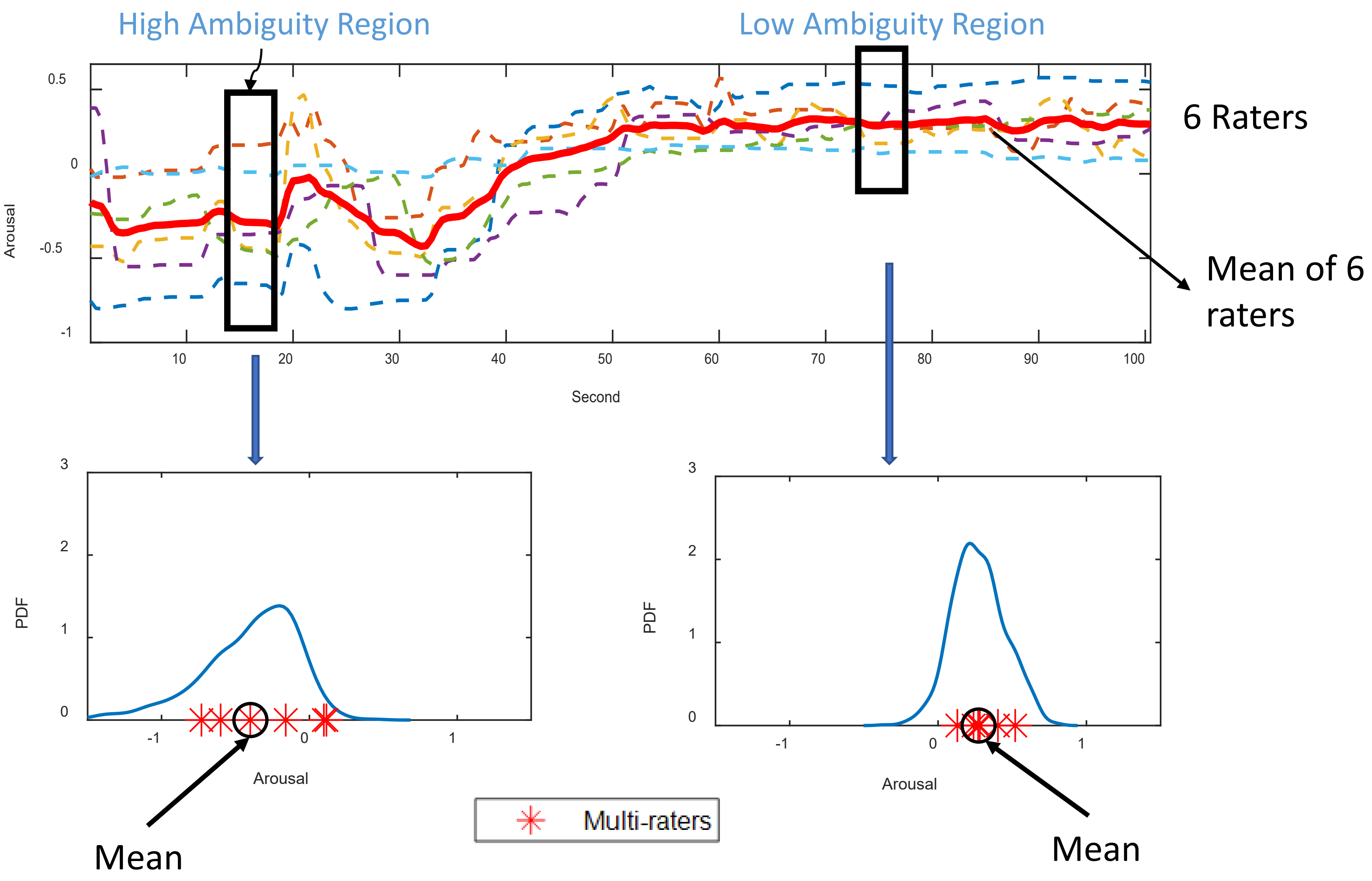
Emotion Labels and Inter-rater Ambiguity

- Typically, emotion ratings are collected from multiple human raters.
- Emotions are not perceived uniformly across individuals.
- Most of the existing works take the average or weighted average of multiple ratings as ‘gold standard’.
- The inter-rater ambiguity which contains emotion subtlety information is ignored.



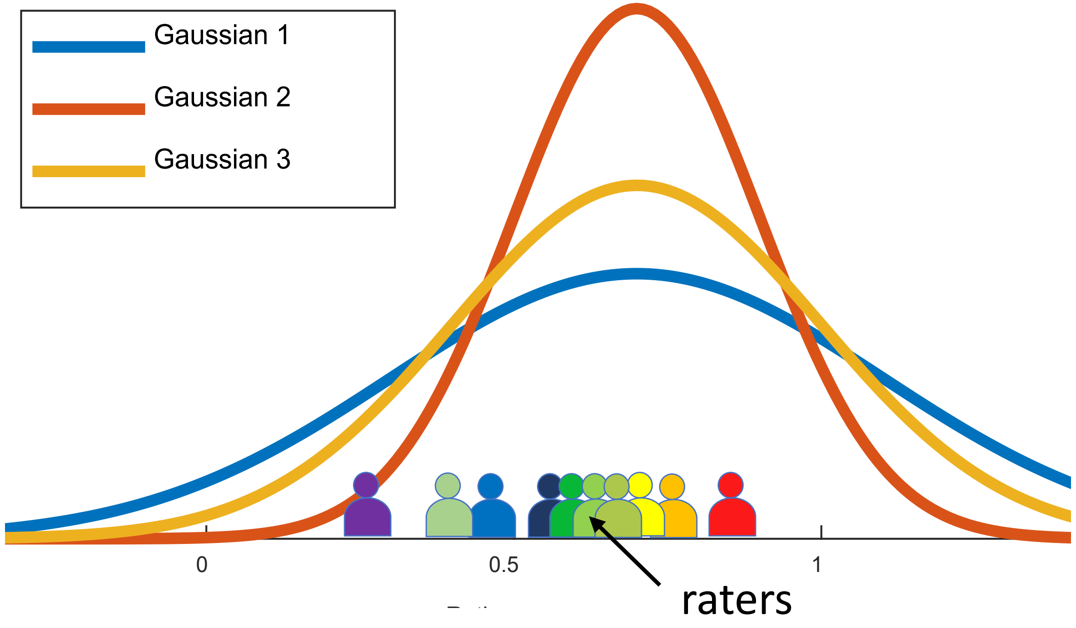
A speech emotion prediction system that is able to model both the emotion state, as well as the ambiguity in the state.

Challenges with Inter-rater Ambiguity

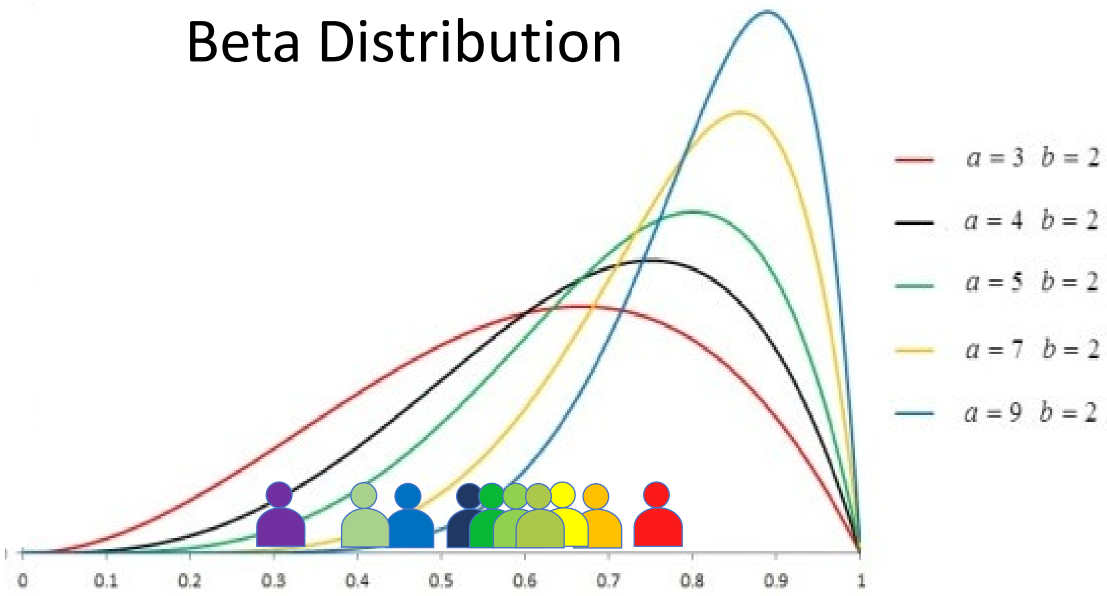


Existing Work - Modelling Emotion Ambiguity

Parametric Distributions



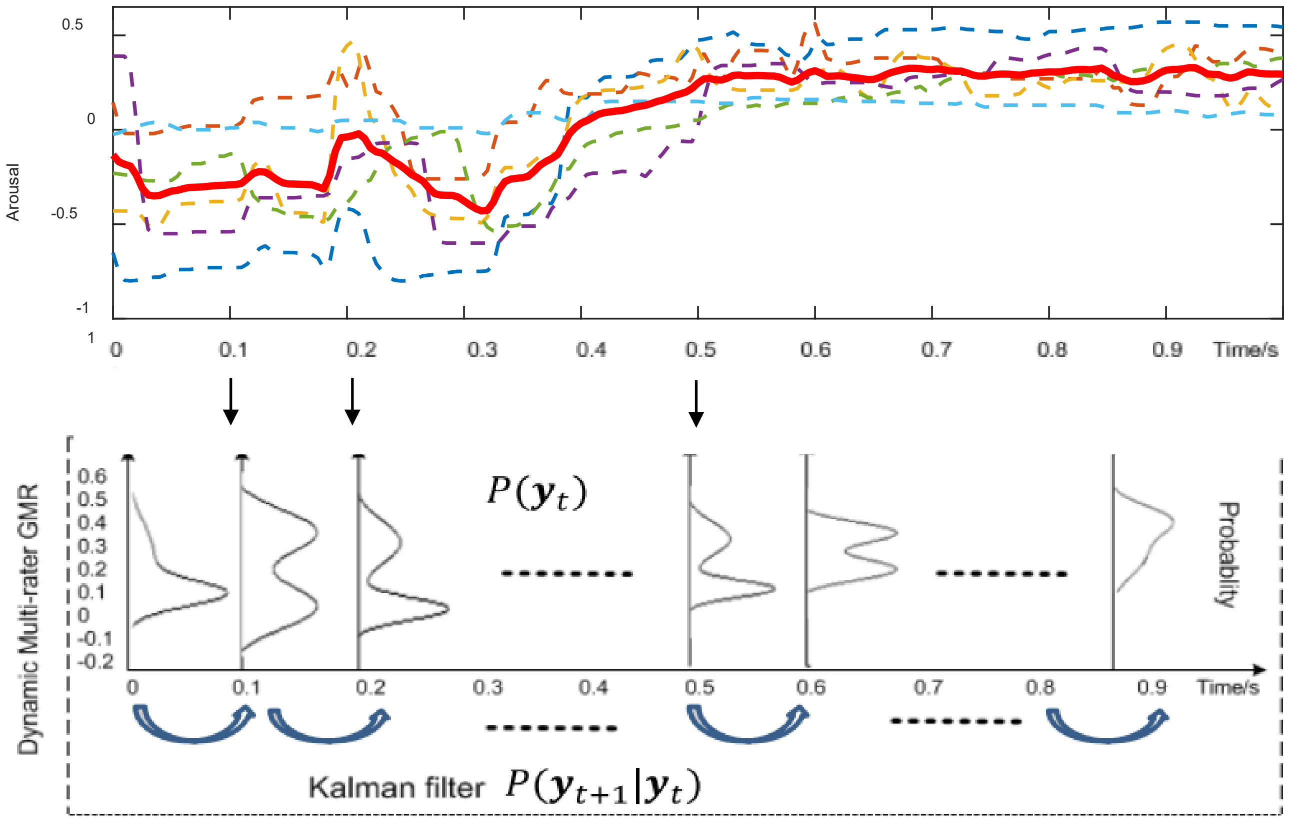
(Zhang, et al., 2018)



(Bose, et al., 2021)

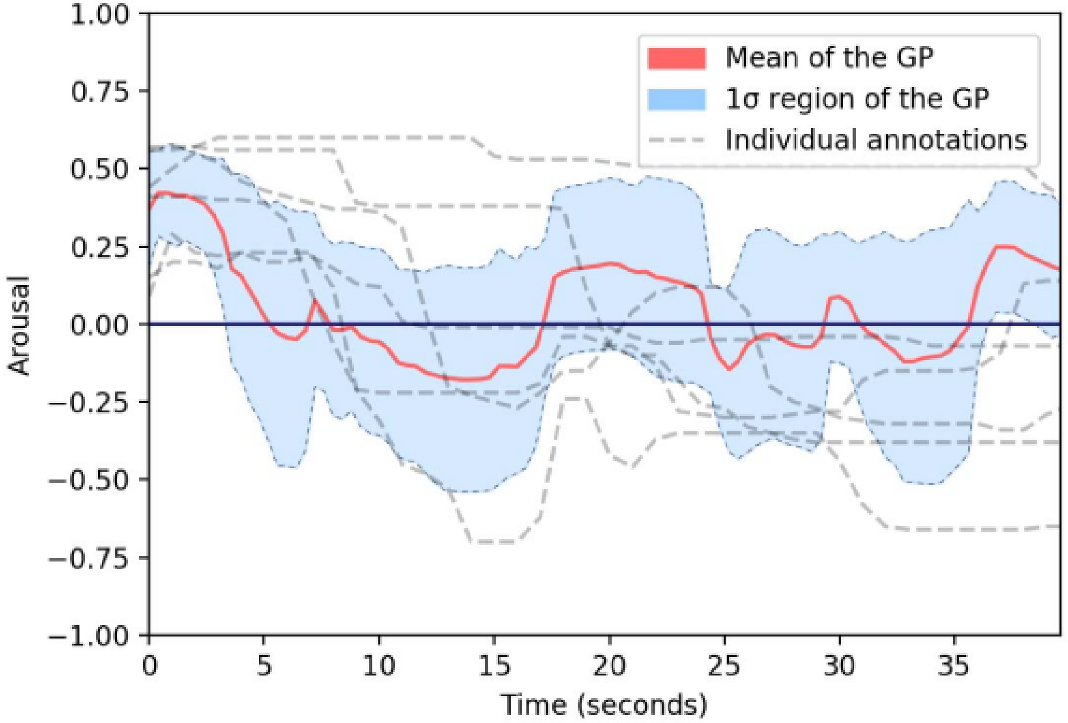
Existing Work - Modelling Emotion Ambiguity

Linear Dynamical Systems



Gaussian mixture model (GMM) with Kalman filter capturing time variations of emotion ambiguity (Dang , et al., 2018).

Gaussian Process

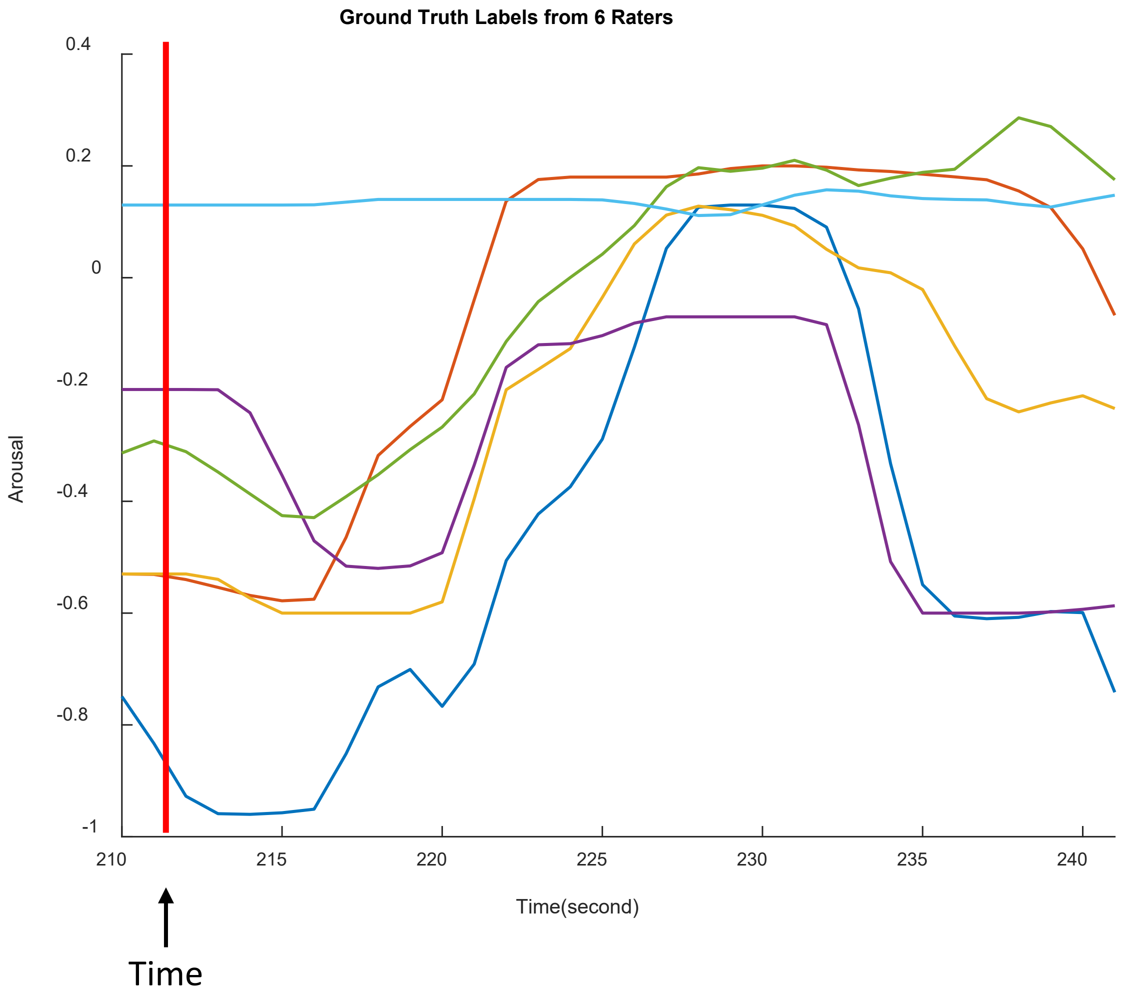


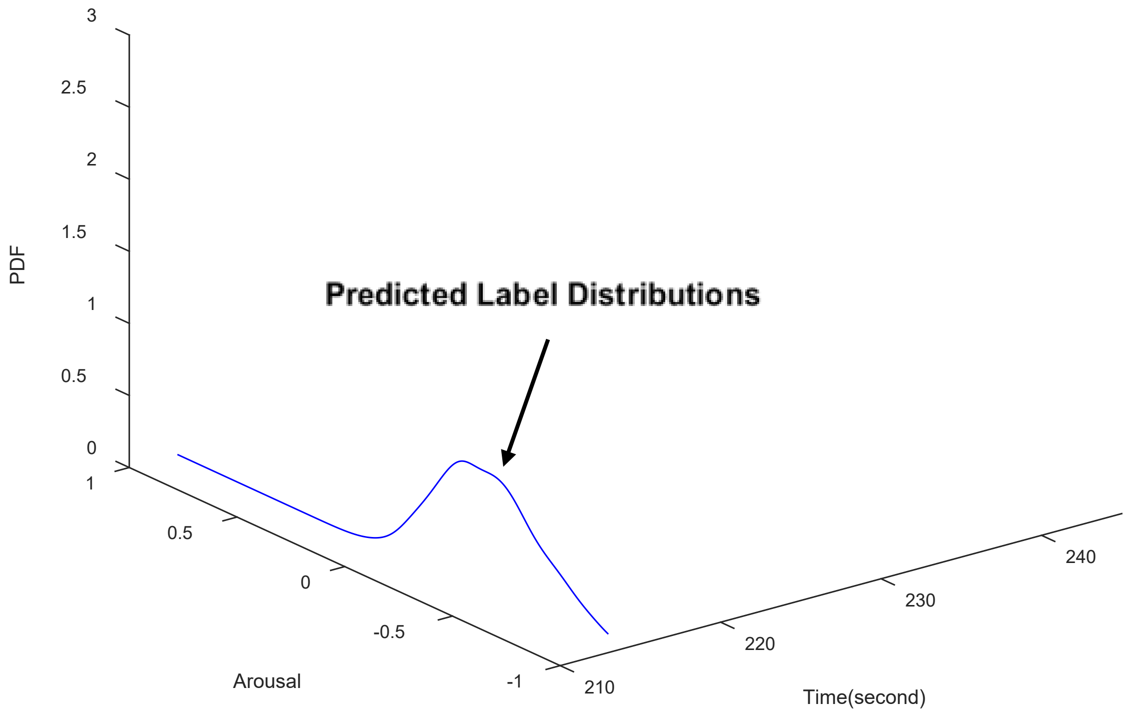
Gaussian Process modelling the ambiguity that captures emotion temporal dynamics (Atcheson , et al., 2019).

Objective

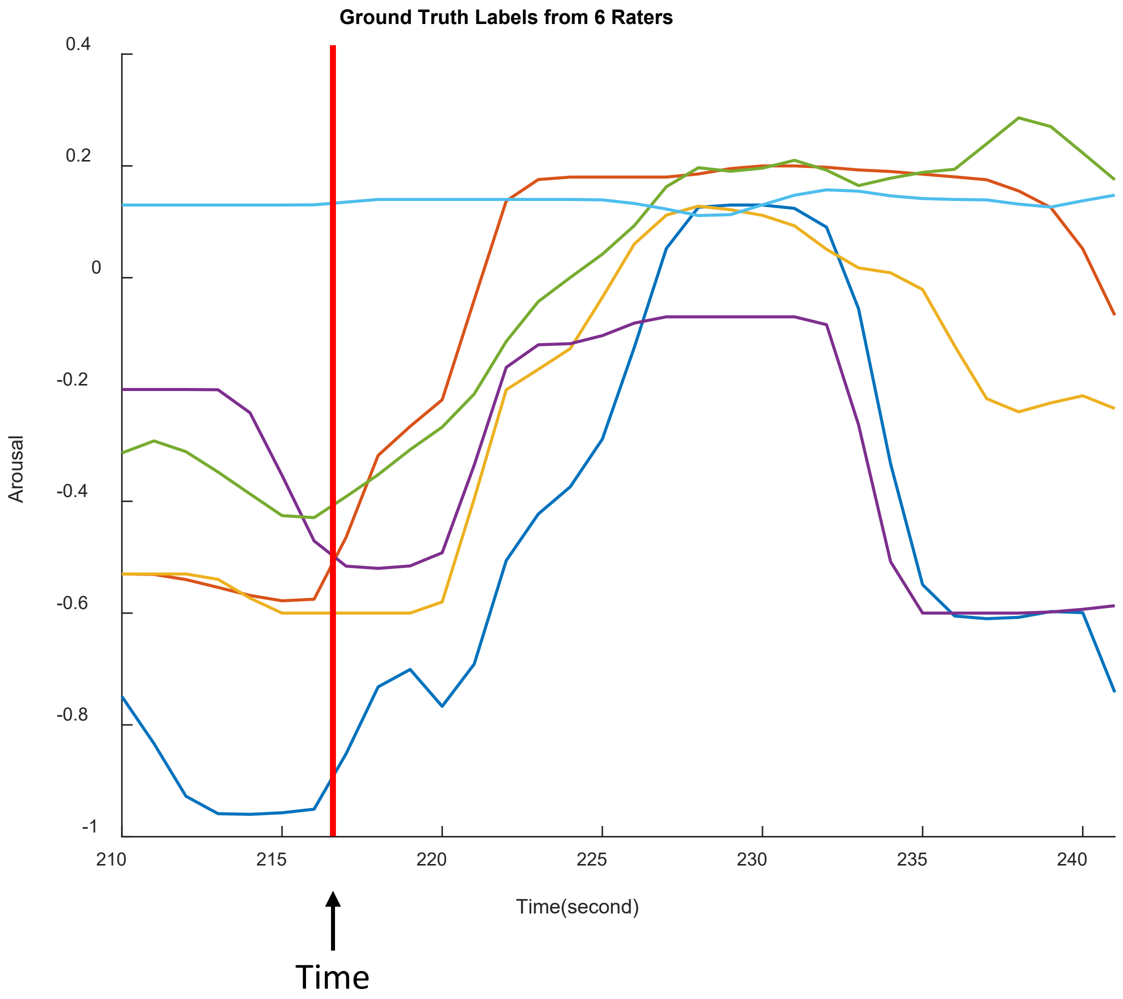
Develop an ambiguity aware emotion prediction framework that models time-varying emotion state (arousal and valence) as well as the ambiguity in the perceived emotion, with non-parametric and non-linear dynamical model.

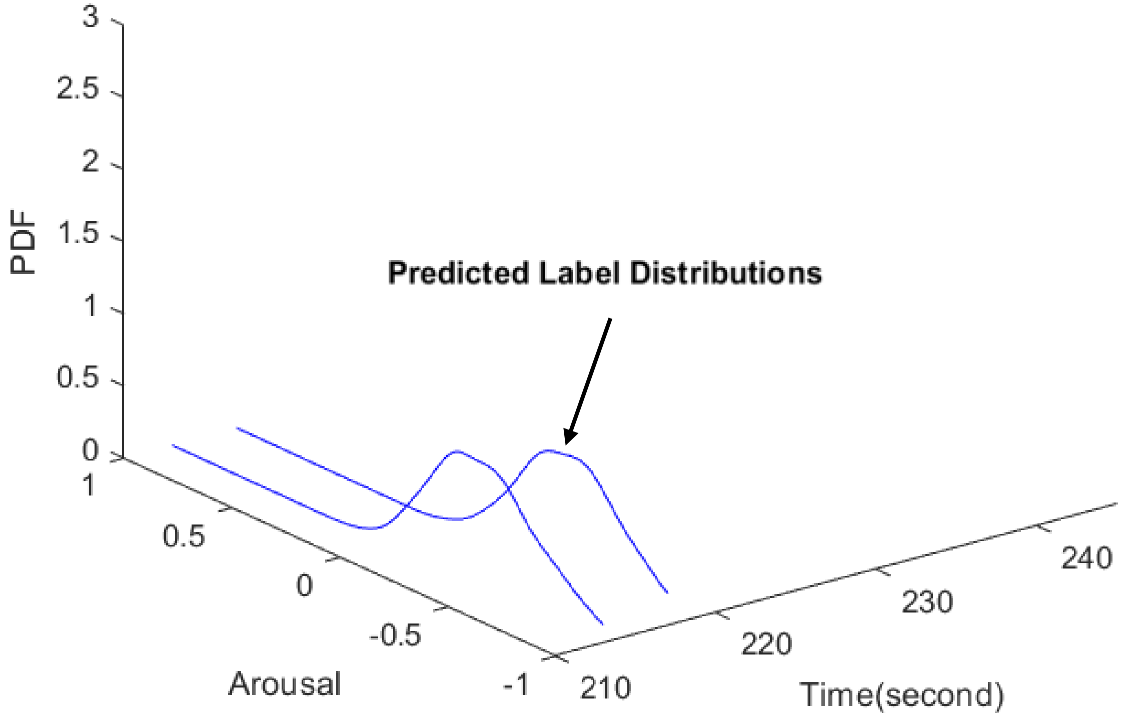
Proposed Sequential Monte Carlo Framework



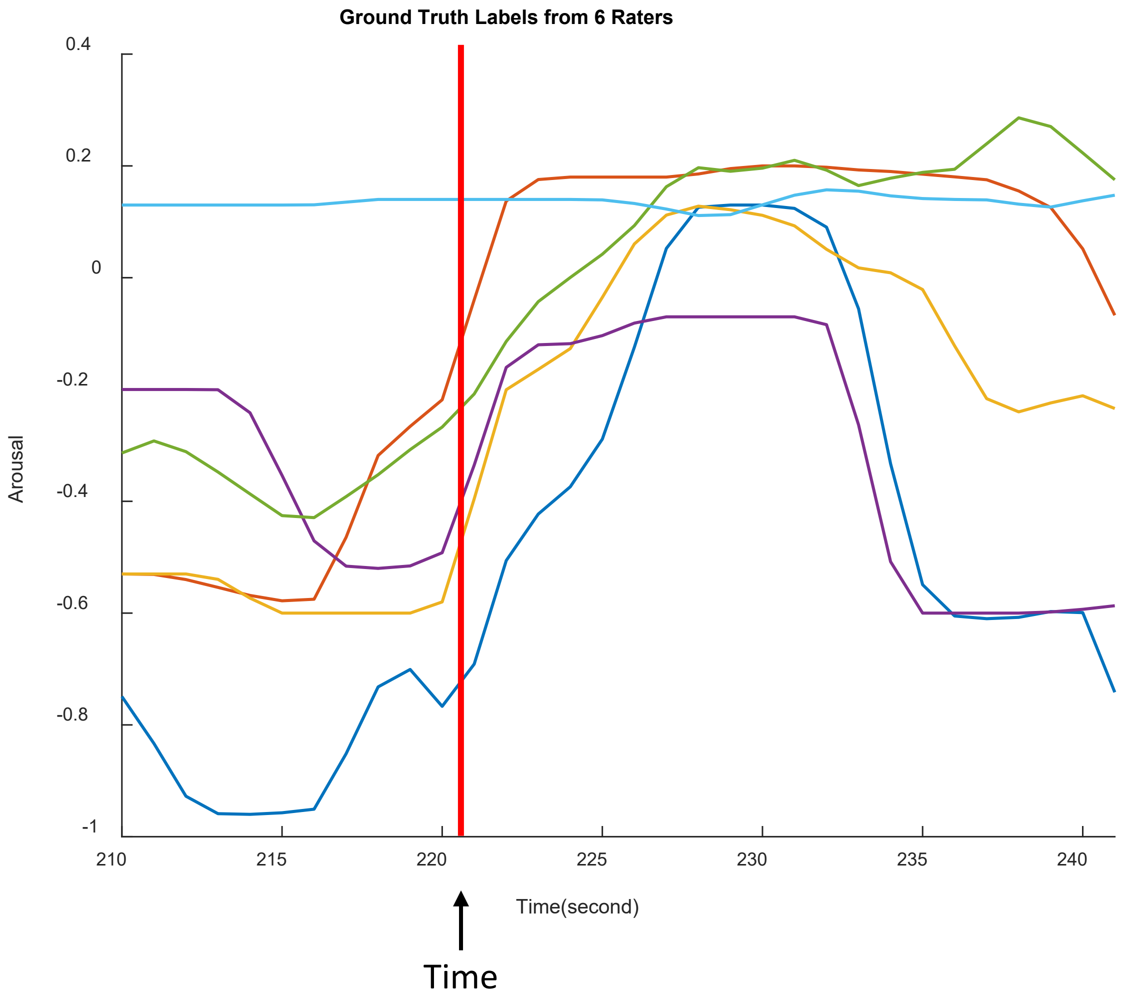


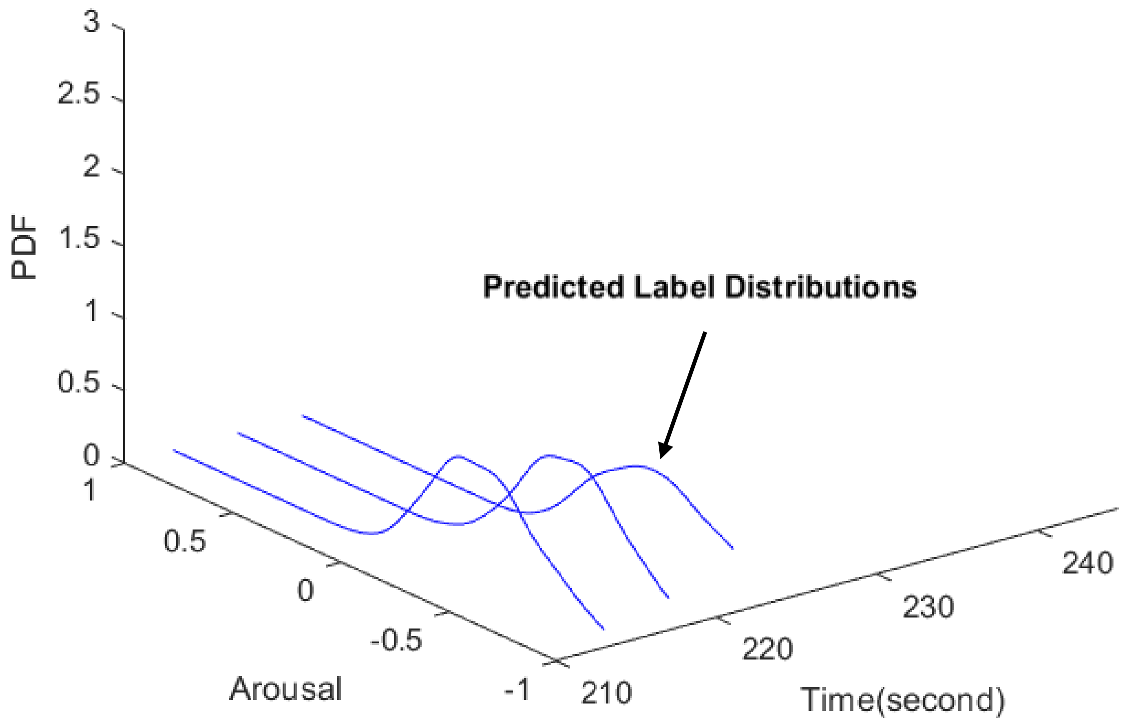
Proposed Sequential Monte Carlo Framework



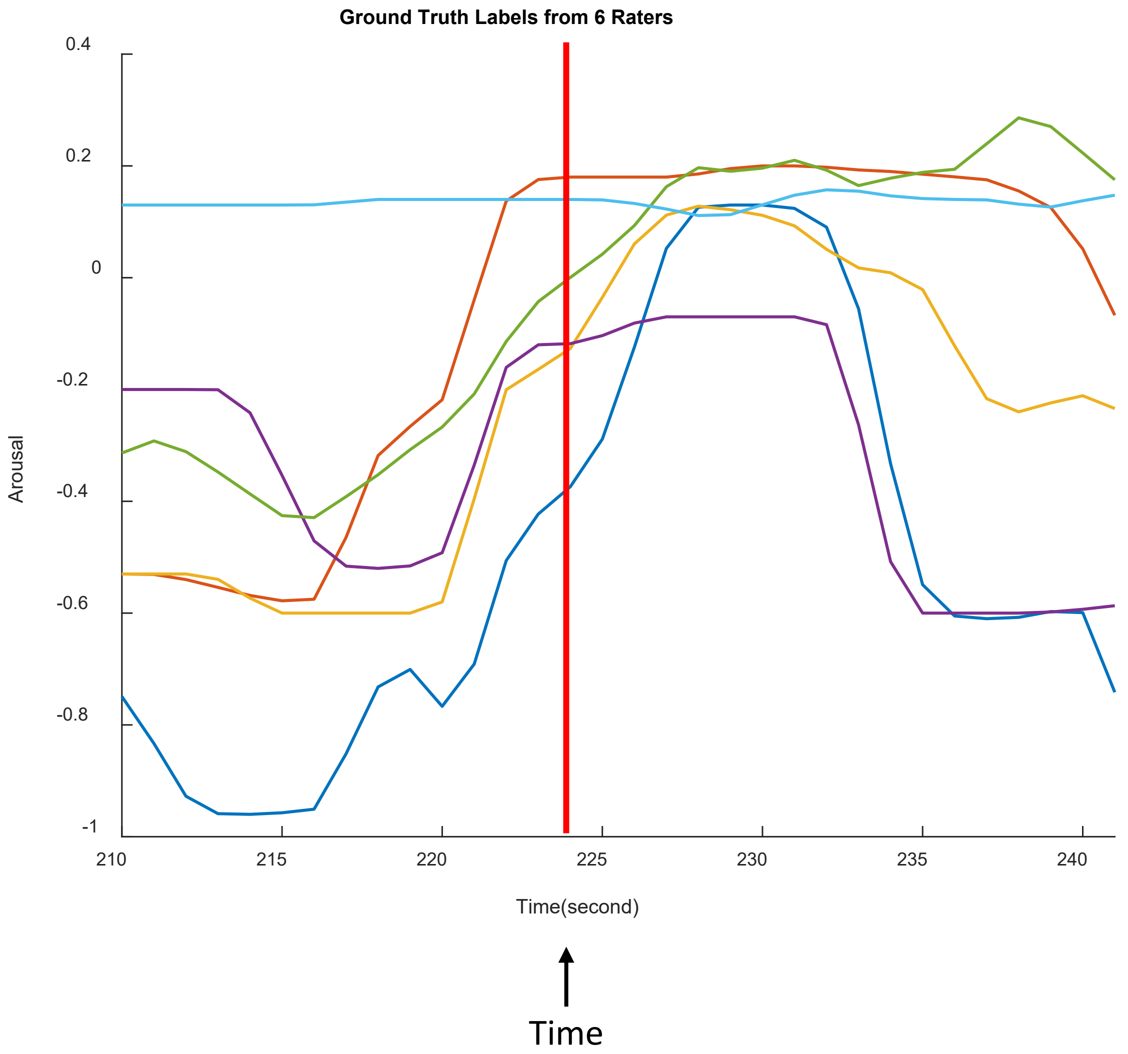


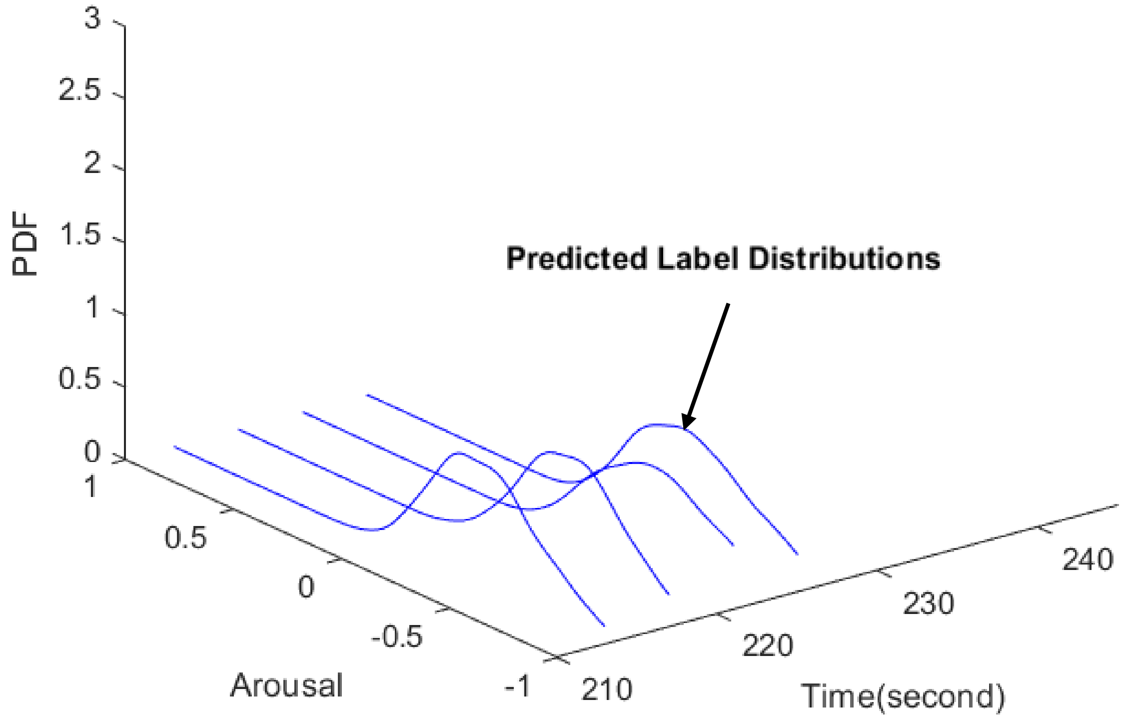
Proposed Sequential Monte Carlo Framework



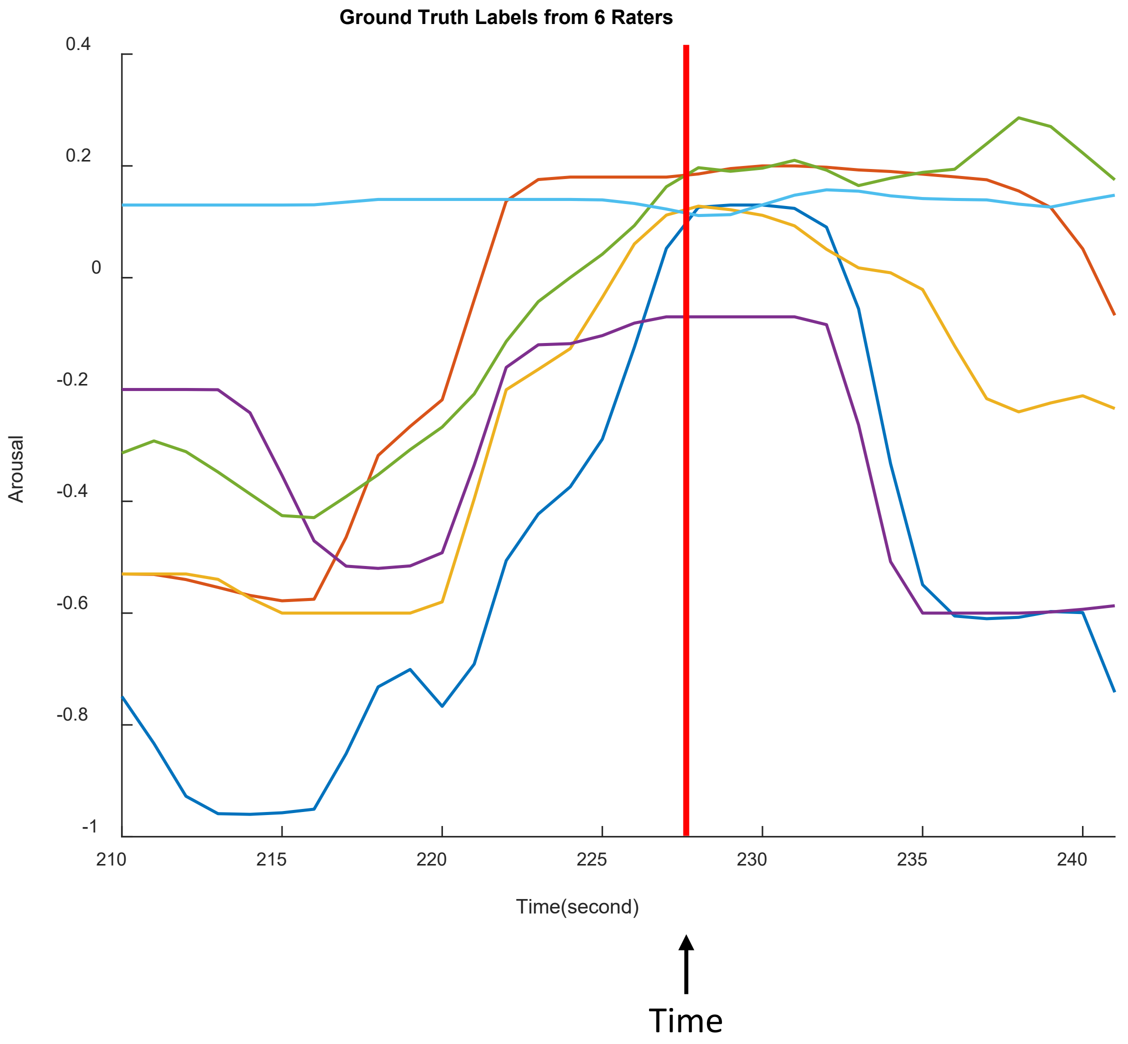


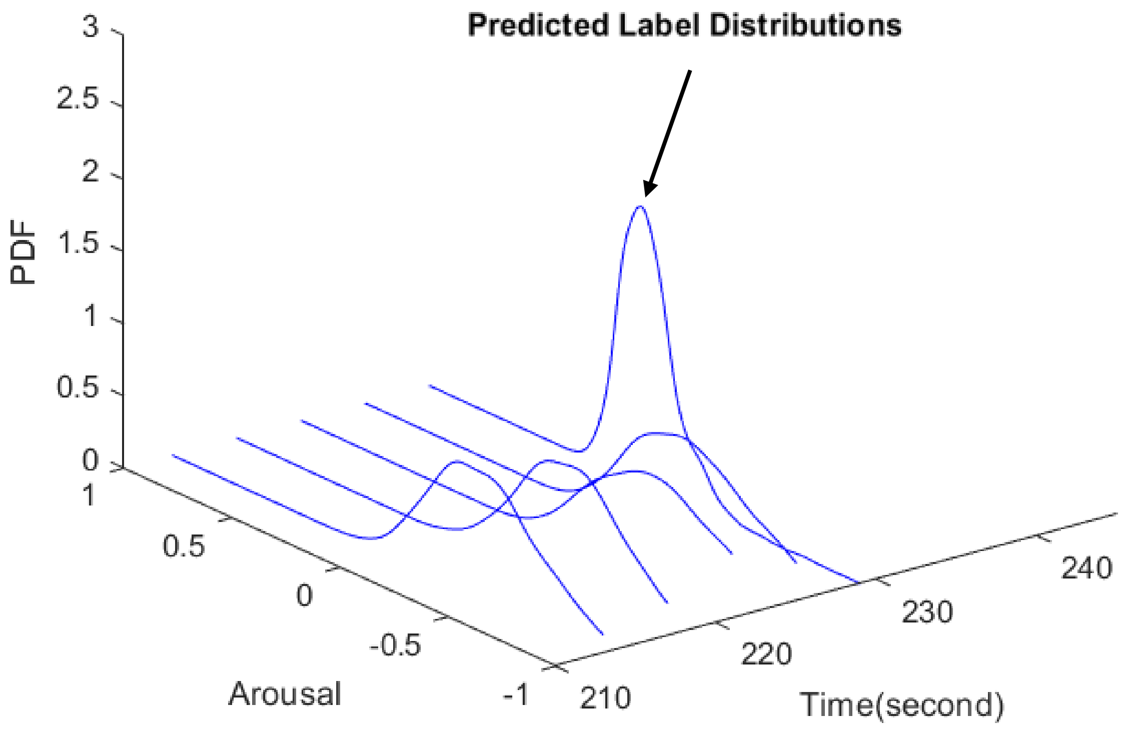
Proposed Sequential Monte Carlo Framework



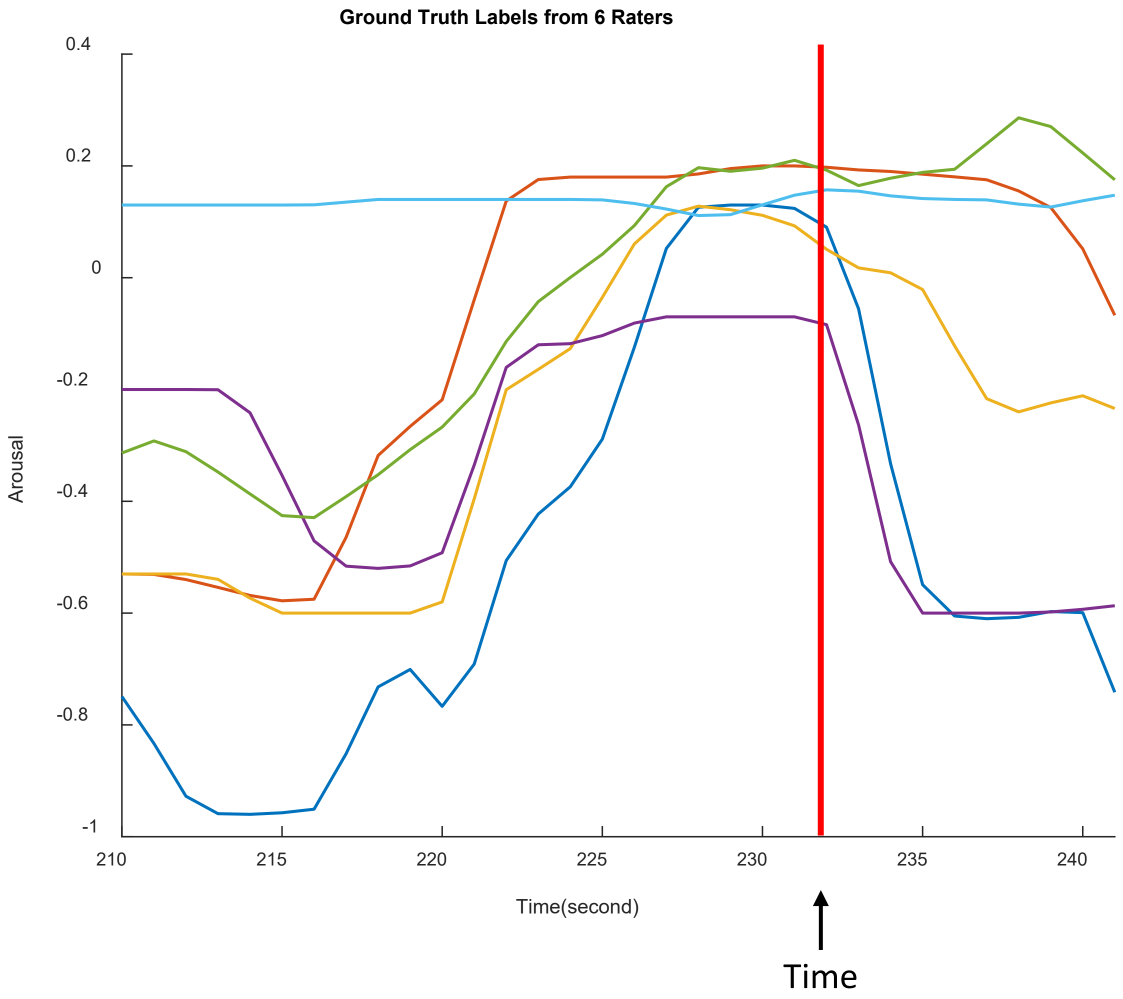


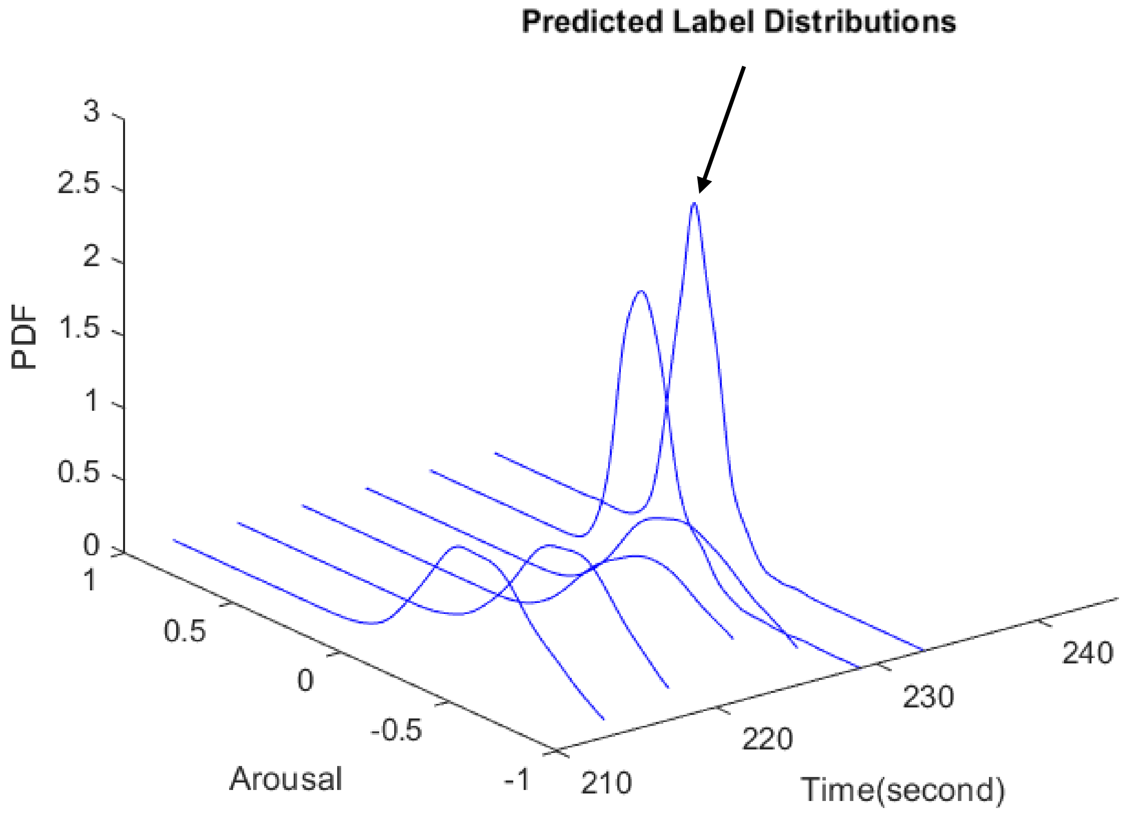
Proposed Sequential Monte Carlo Framework



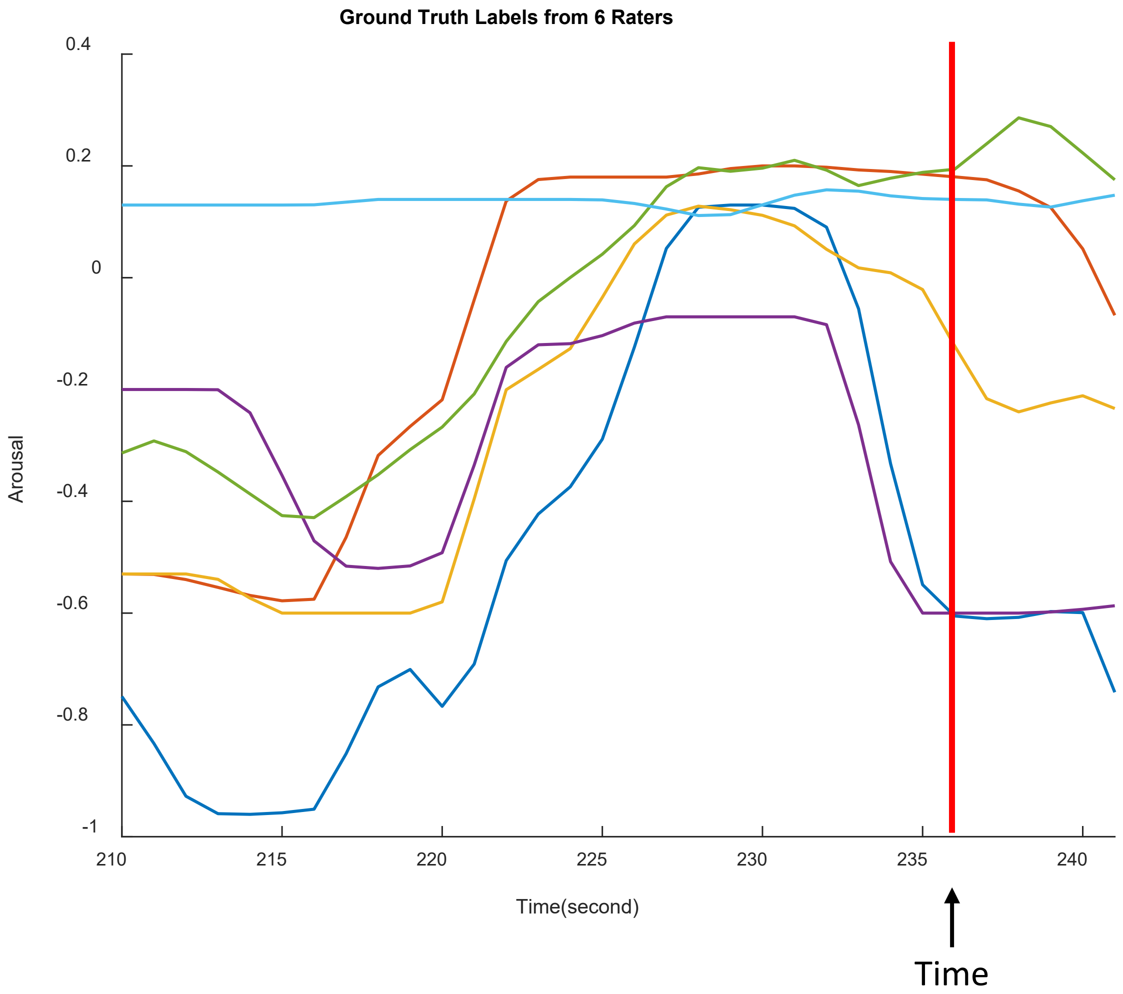


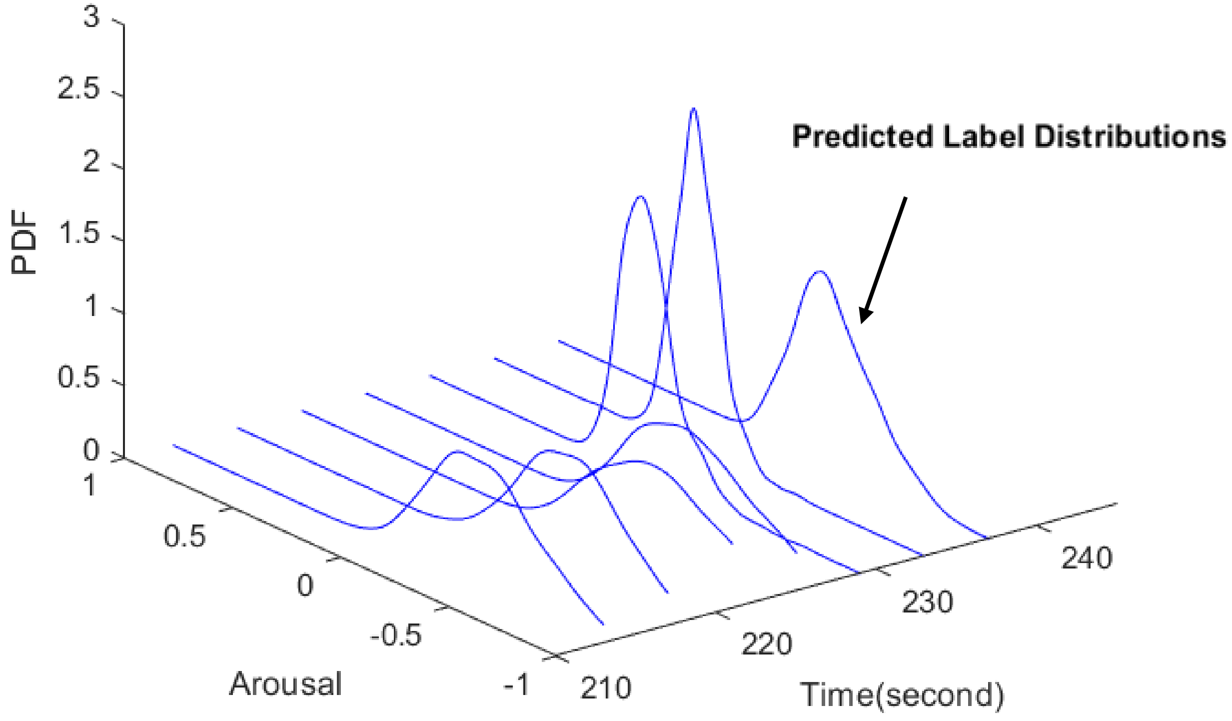
Proposed Sequential Monte Carlo Framework



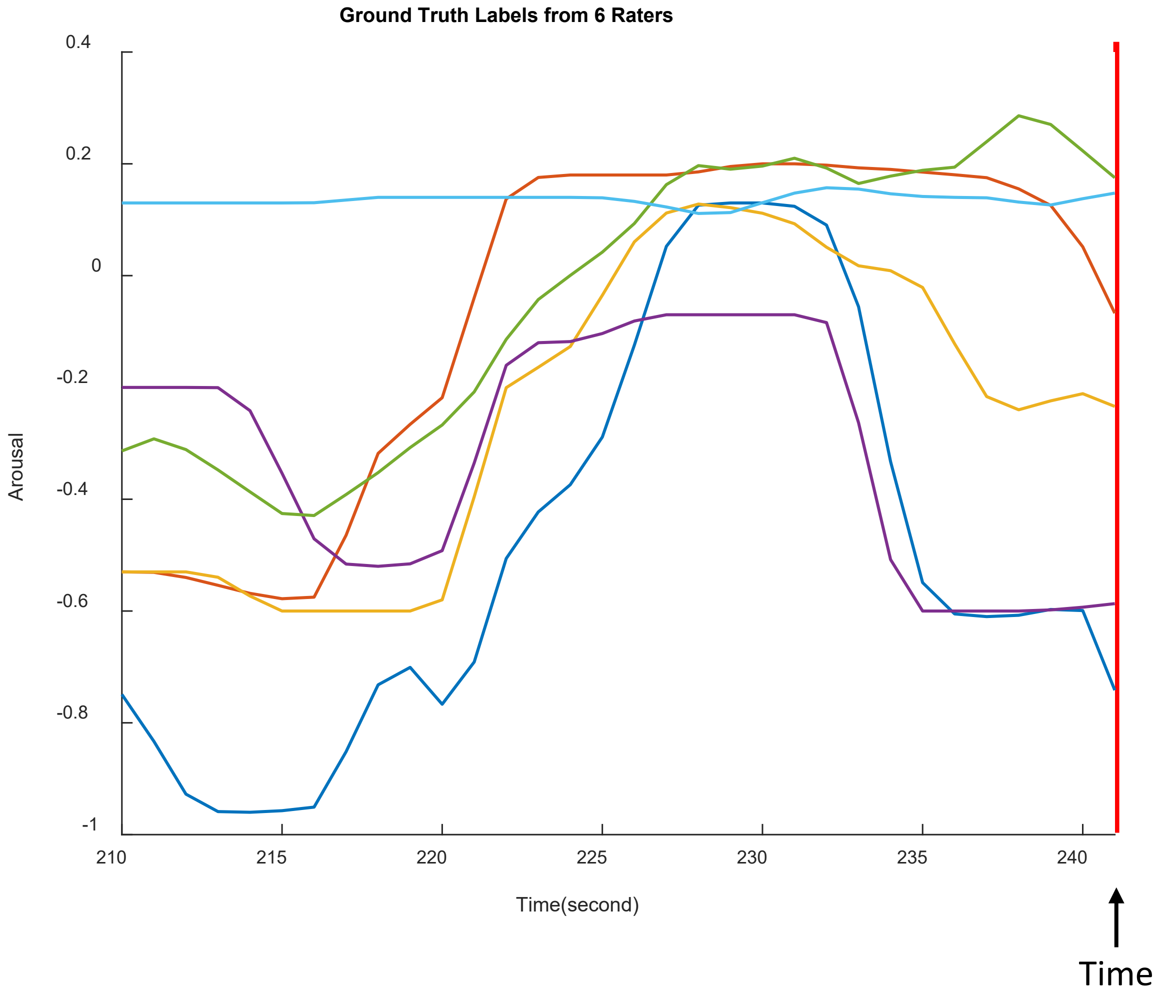


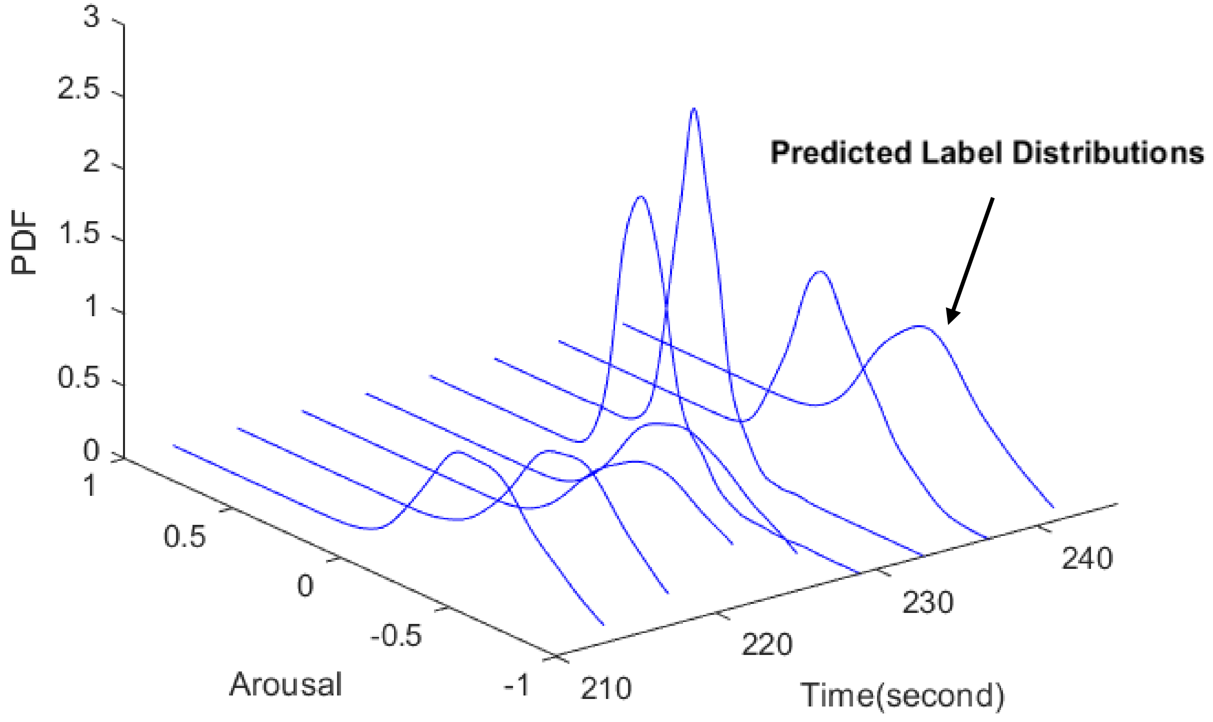
Proposed Sequential Monte Carlo Framework



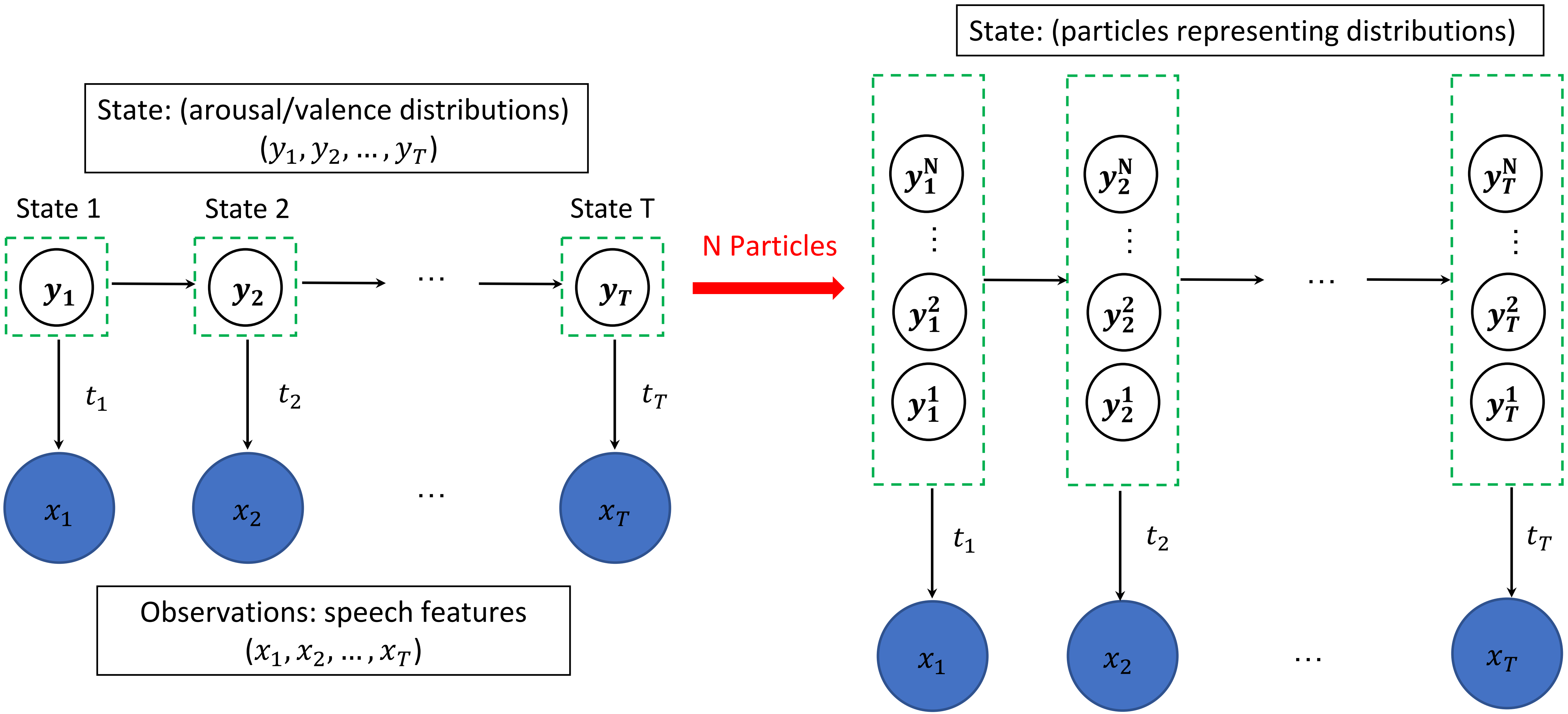


Proposed Sequential Monte Carlo Framework

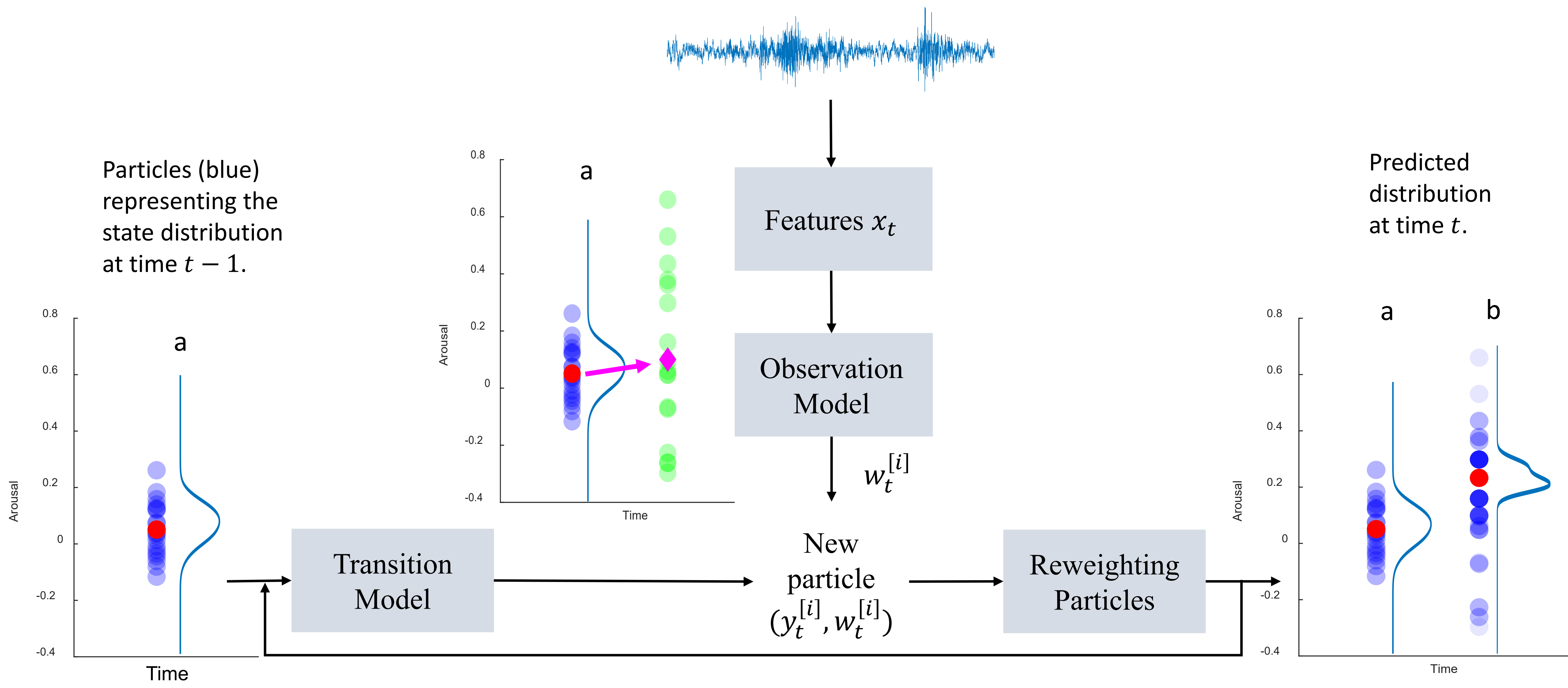




Continuous State Space Model

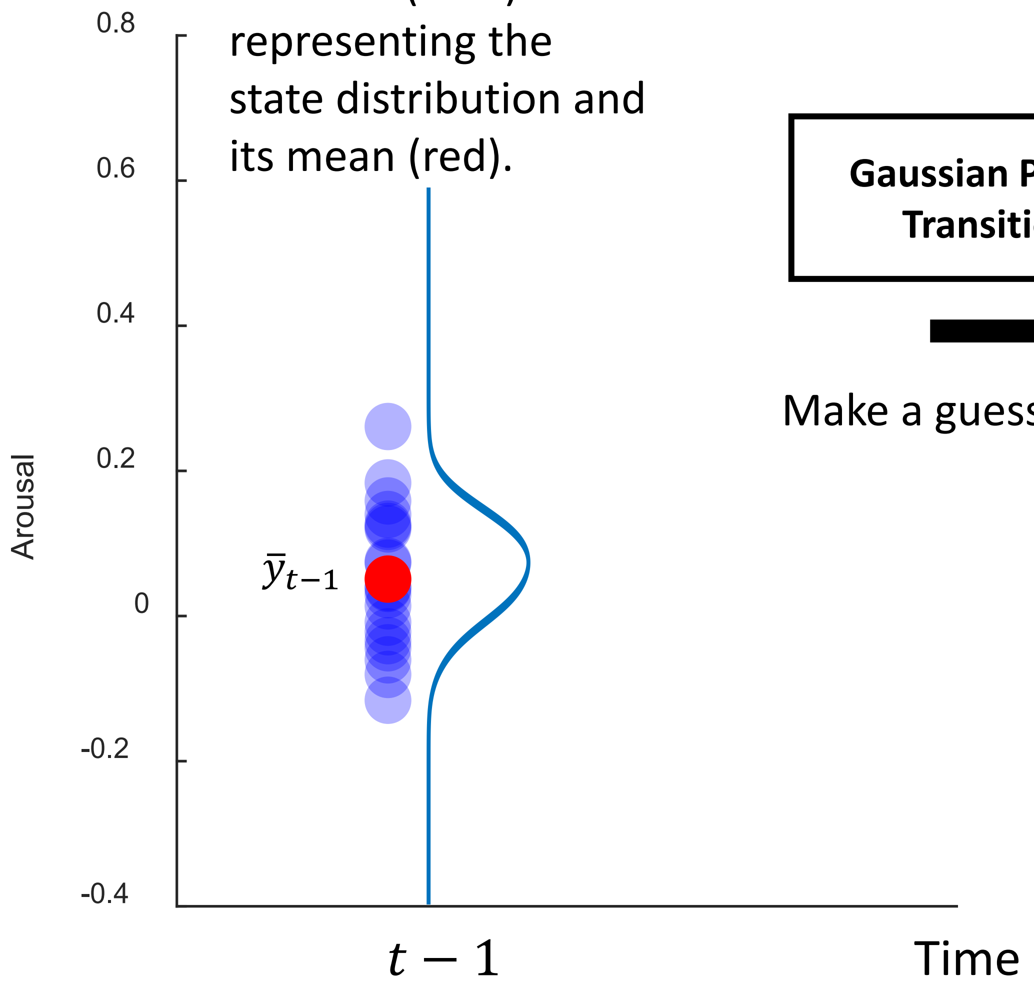


Proposed Sequential Monte Carlo Framework



SMC Processes Breakdown – State Transition

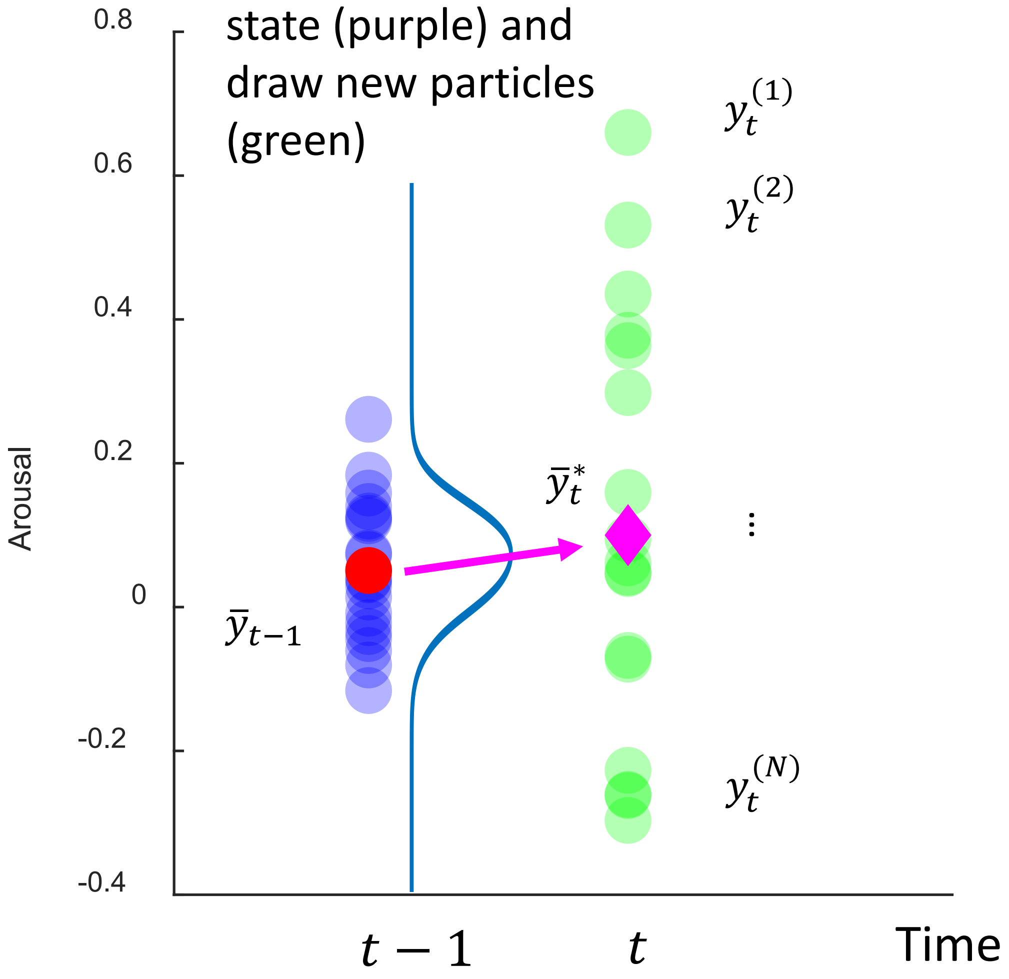
Particles (blue) representing the state distribution and its mean (red).



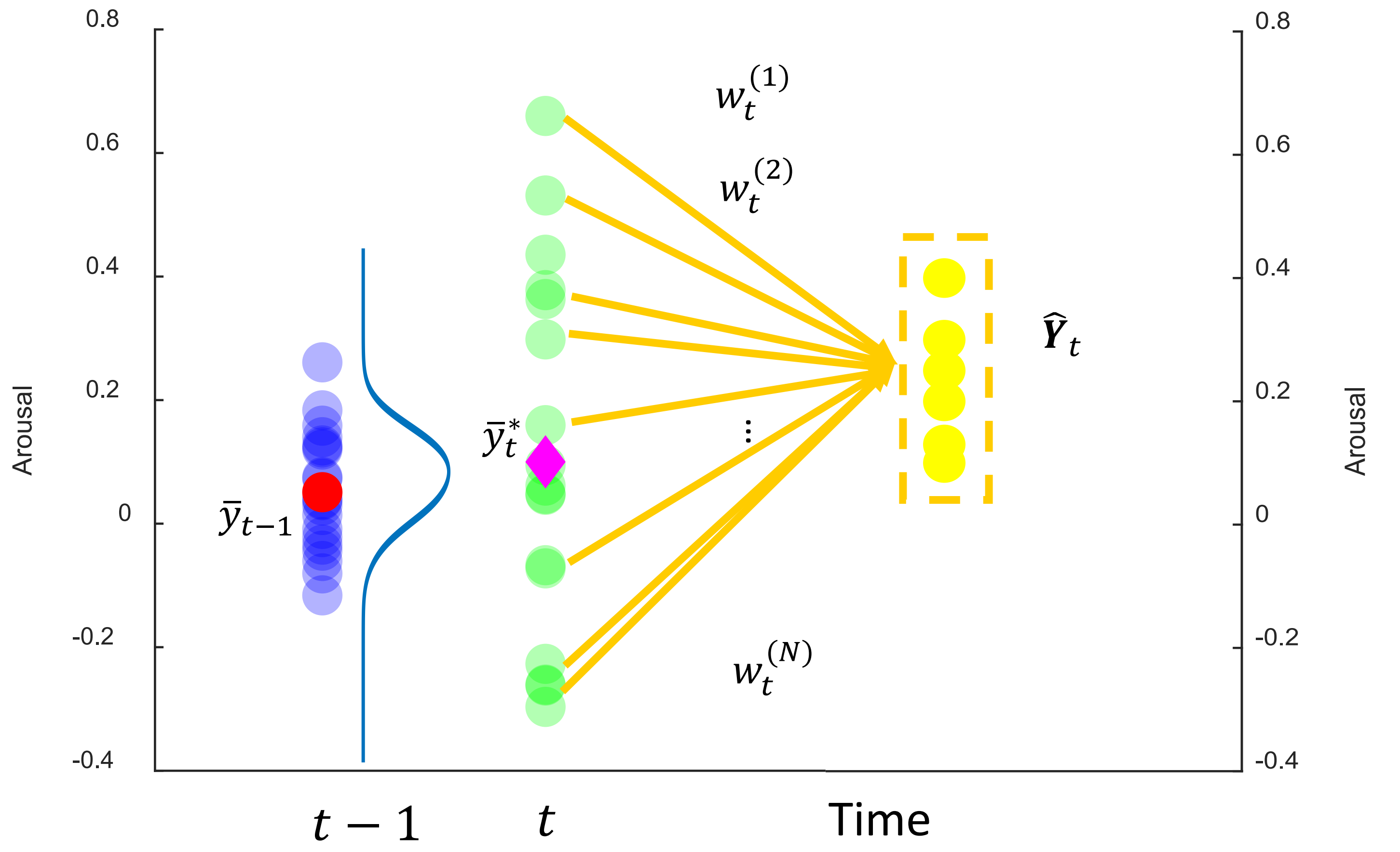
Gaussian Process State Transition model

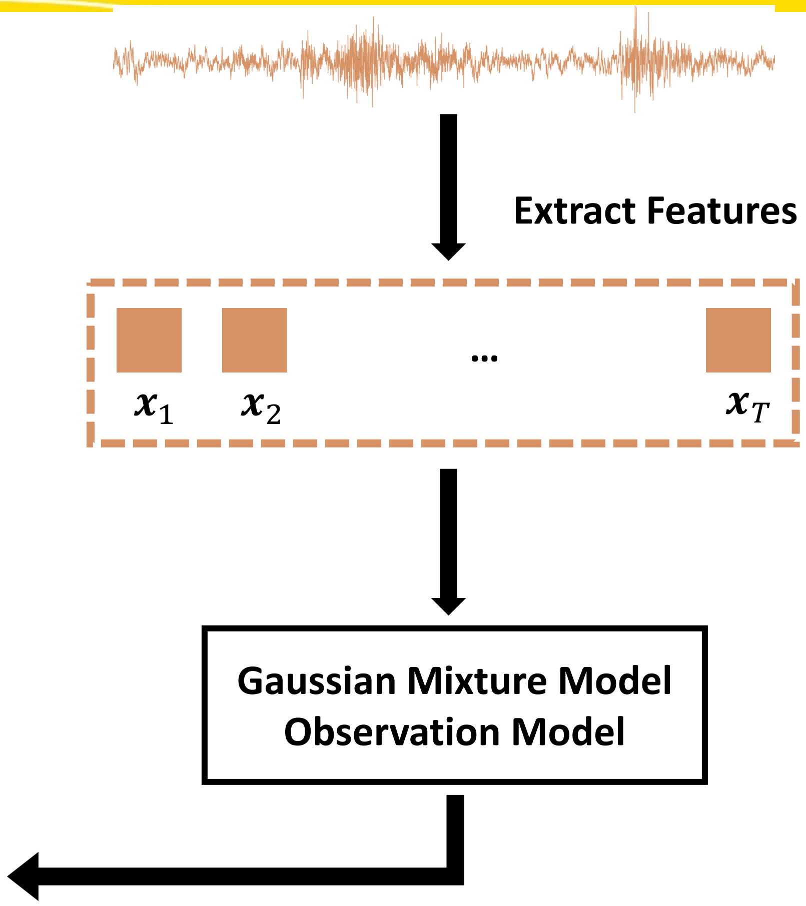
Make a guess of the new state

Transient to the new state (purple) and draw new particles (green)



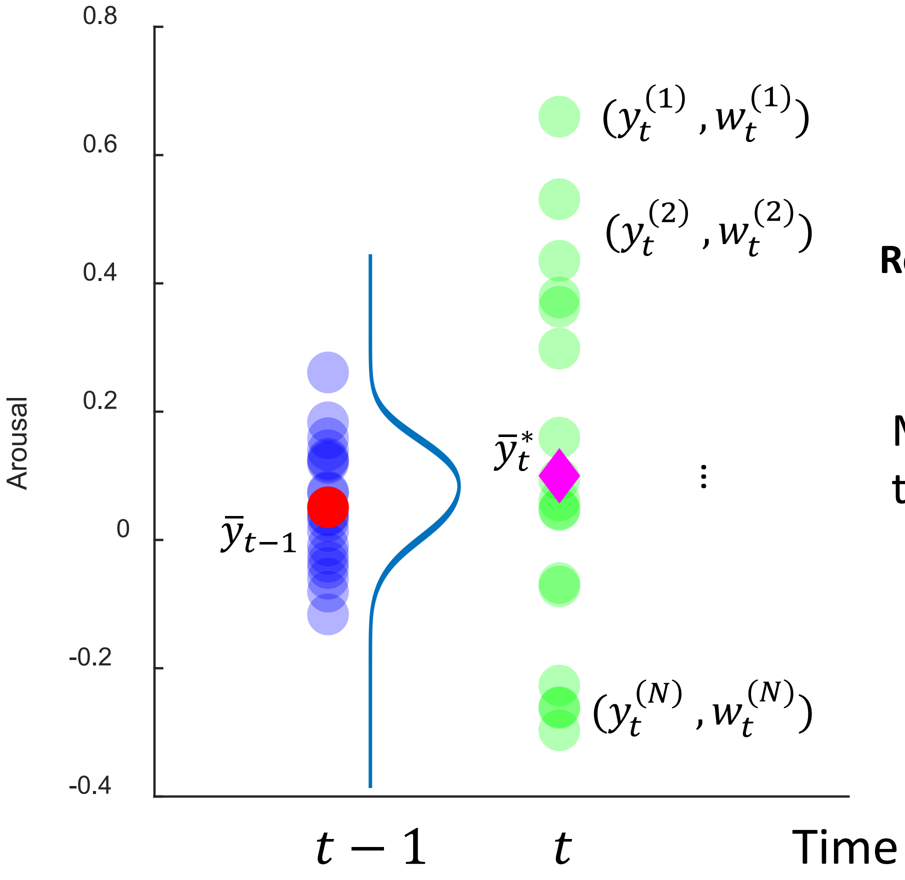
SMC Processes Breakdown – State observation





Link the observations to the estimated new state.

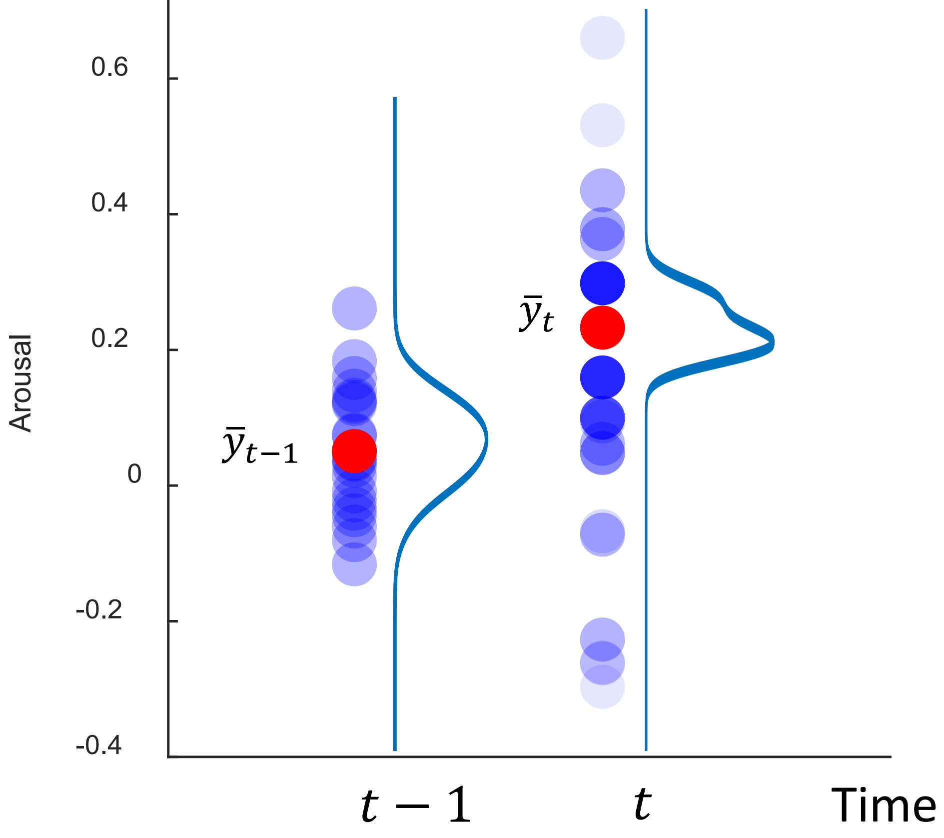
SMC Processes Breakdown – Samples Reweighting



Reweighting the Particles

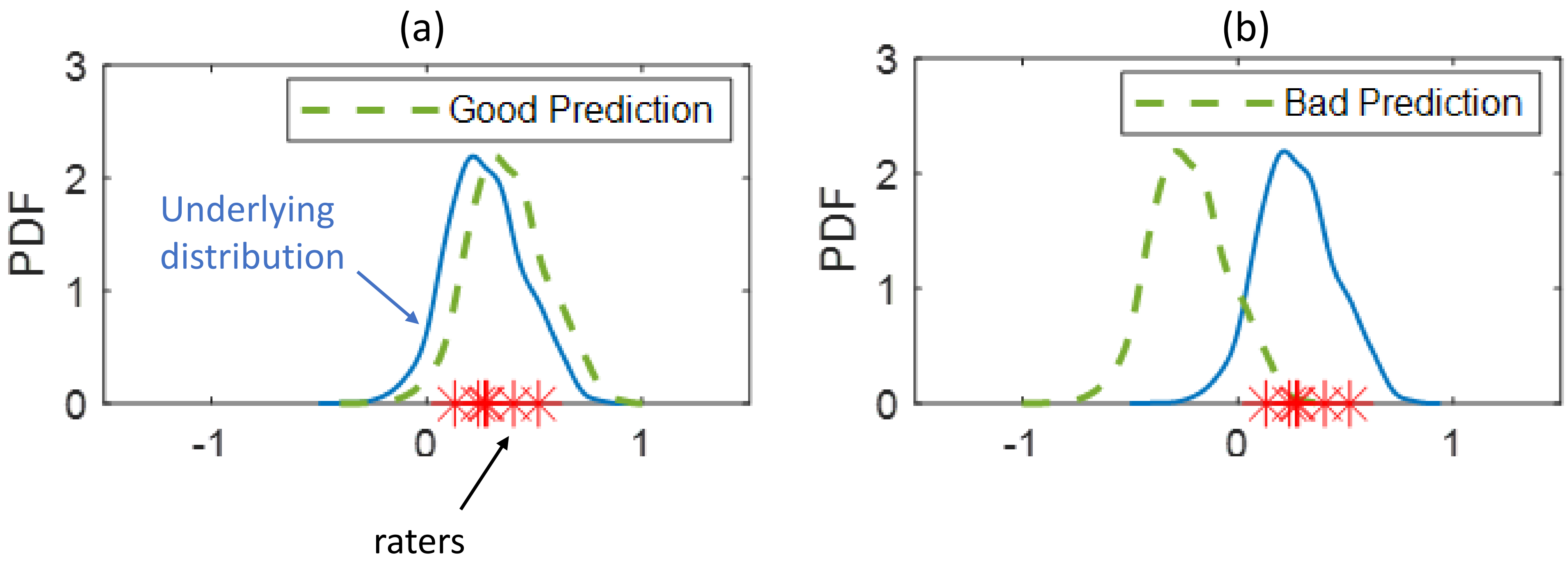
Make corrections of the estimated state.

Final predicted distribution and its weighted mean.



Validation: Proposed Measures

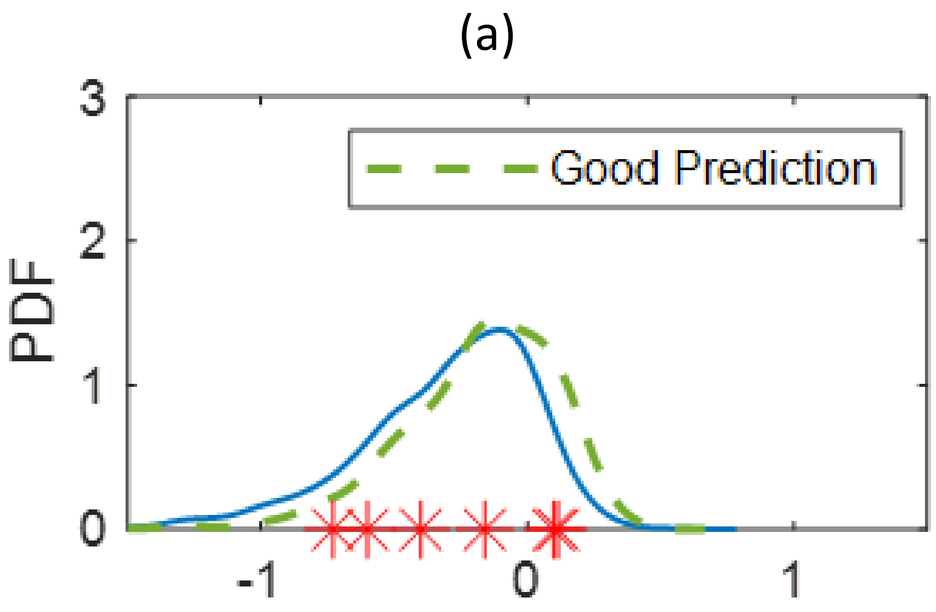
Low Ambiguity Region

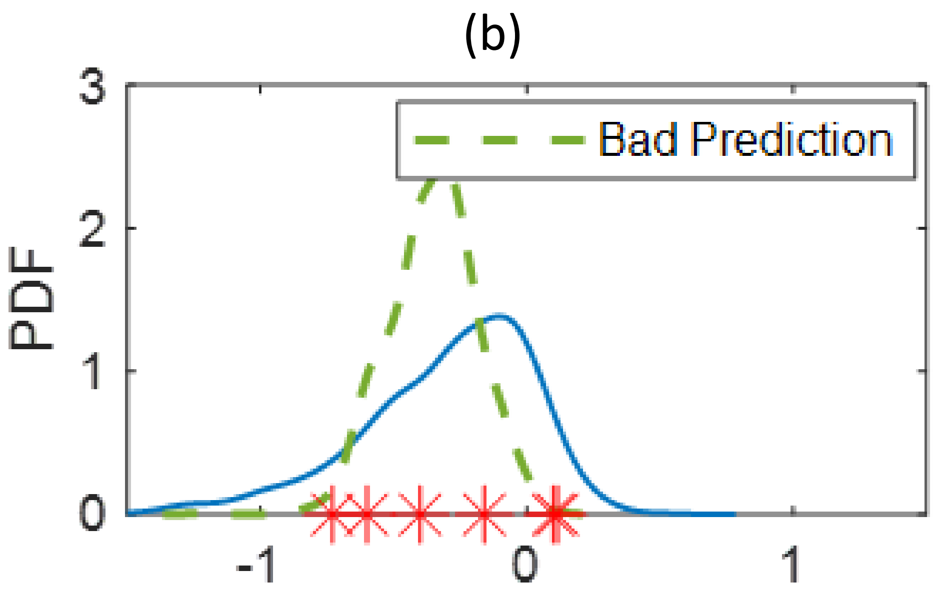


- Distributions should be narrow.
- Predicted mean should be closed to the ground truth mean.

Proposed Measures

High Ambiguity Region





- Mean is less important.
- Distributions should be broad.

Experimental Settings

- Corpus: the RECOLA dataset; 9 training & 9 development utterances.
- Arousal & valence labels; 6 annotators.
- 40ms sampling rate; 1 second window (50% overlap).
- Delay compensation: 4 seconds for arousal and 2 seconds for valence.
- Features: Bag-of-audio-words(BoAW) features with 100 clusters.
- 8 - mixture GMM for λ_1, 4 – mixture GMM for λ_2
- 1000 particles.

Experimental Results

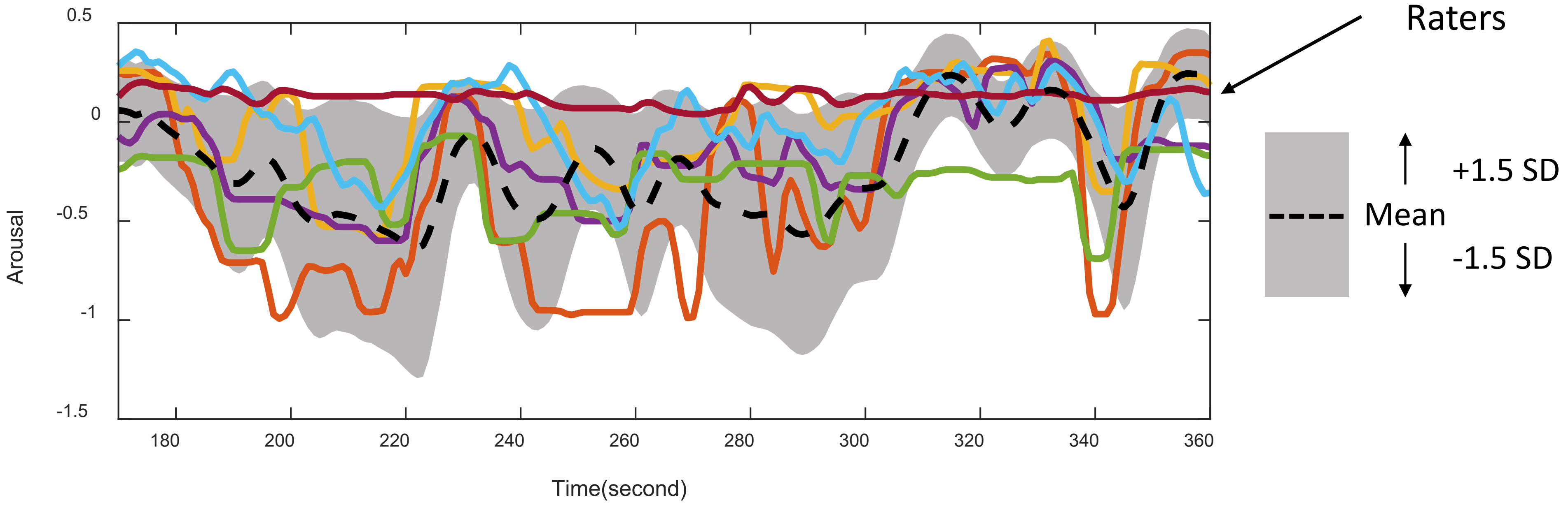
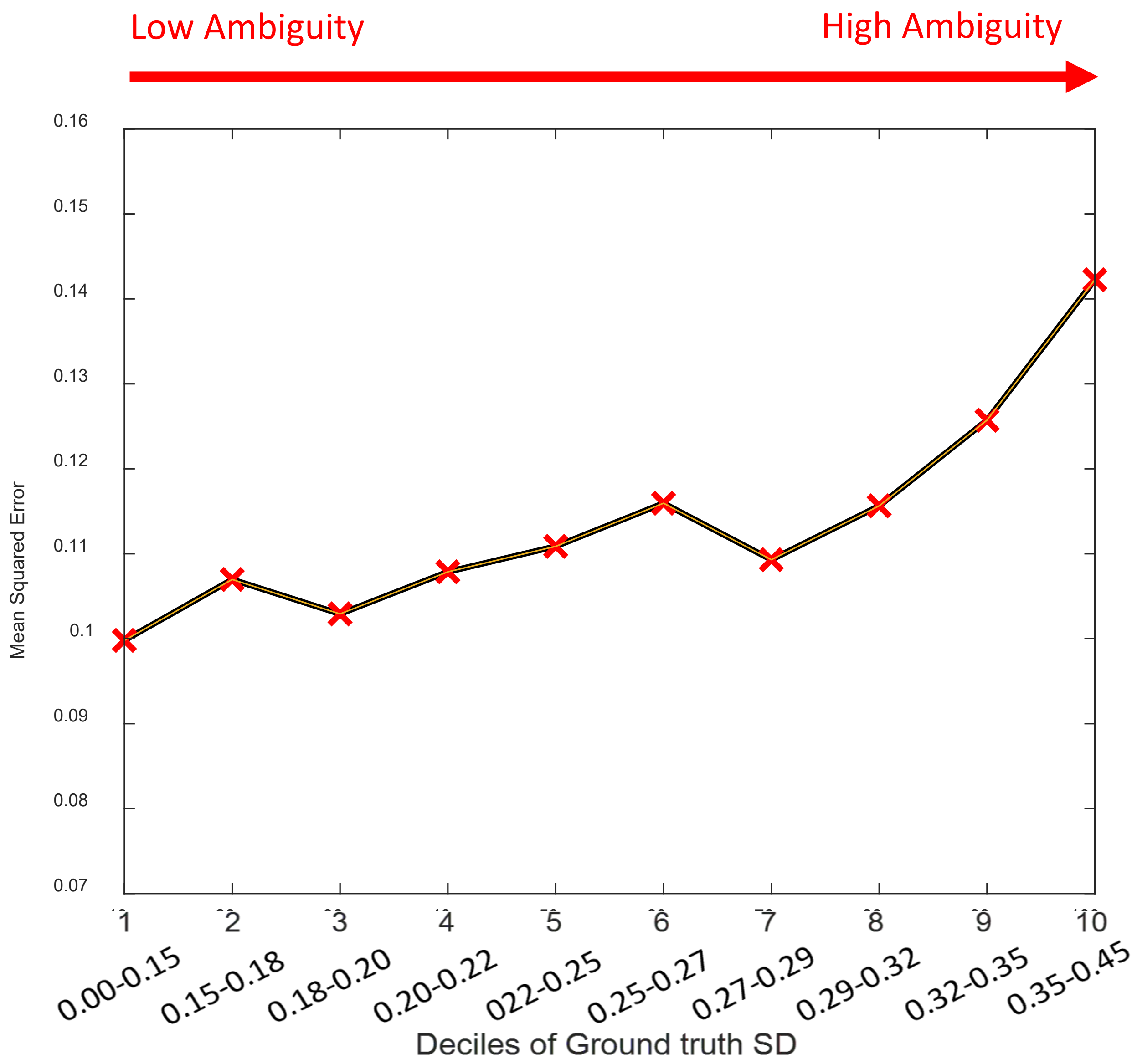


Table 1 CCC and CC measure between predicted SD and SD from 6 annotators

	Arousal		Valence	
	CCC	CC	CCC	CC
BLSTM (Han, et al., 2017)	0.103	-	0.075	-
GMR (Dang, et al., 2018)	-	0.568	-	0.132
Proposed SMC	**0.403**	**0.456**	**0.195**	**0.201**

Experimental Results



CCC between the predicted mean and ground truth mean is 0.702 for arousal and 0.391 for valence.

Conclusions

- We present a novel Sequential Monte Carlo framework that predicts both the emotion state (*arousal* and *valence*) and the ambiguity in the perceived emotion.
- It 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. It predicts the emotion state accurately within regions of low ambiguity, and it identifies the regions of high ambiguity.

Thank you

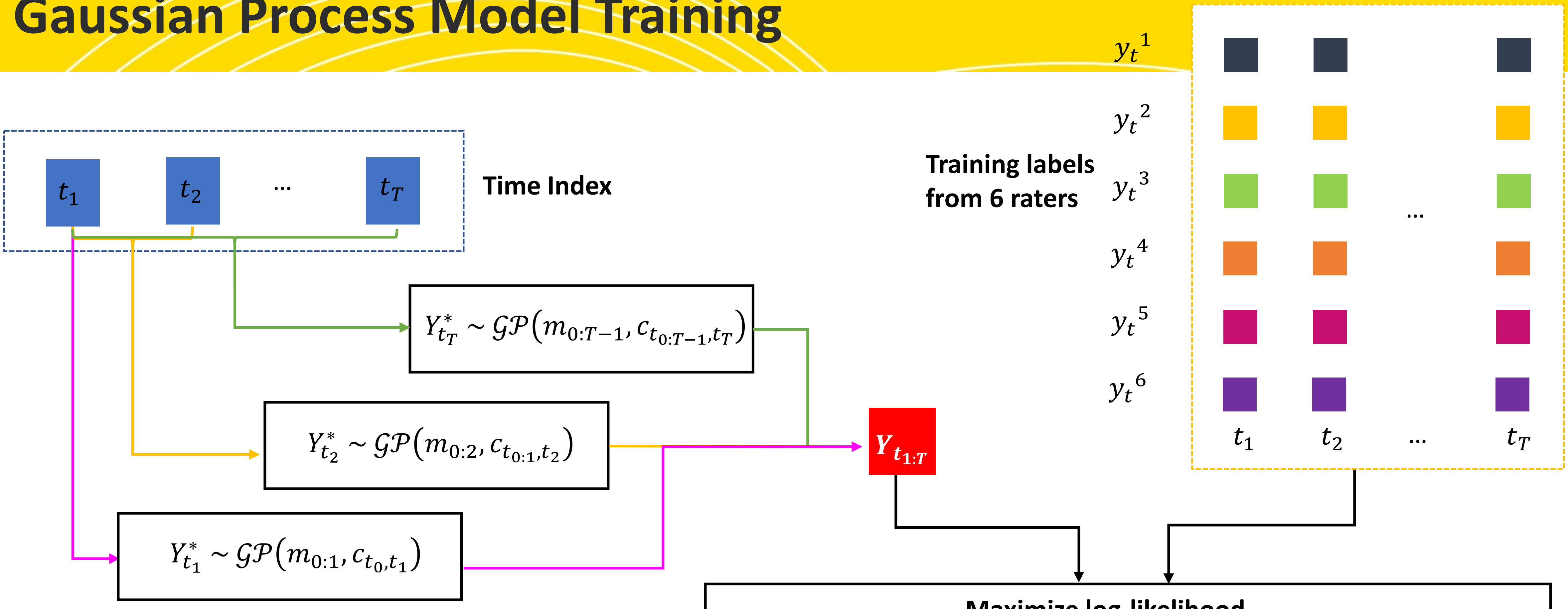
Reference

[1] Hatice Gunes and Maja Pantic, “Automatic, dimensional and continuous emotion recognition,” *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.
[2] Hatice Gunes and Bjorn 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 Bjorn 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.

Reference

[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, Bjorn 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 Bjorn Schuller, “Openxbow: introducing the passau open-source crossmodal bag-ofwords 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 Bjorn 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.

Gaussian Process Model Training



Maximize log-likelihood

$$\theta_{max} = argmax_\theta\{\log(P(\boldsymbol{Y}_{\boldsymbol{0:m}}|\boldsymbol{t}_{\boldsymbol{0:m}}, \theta))\}$$

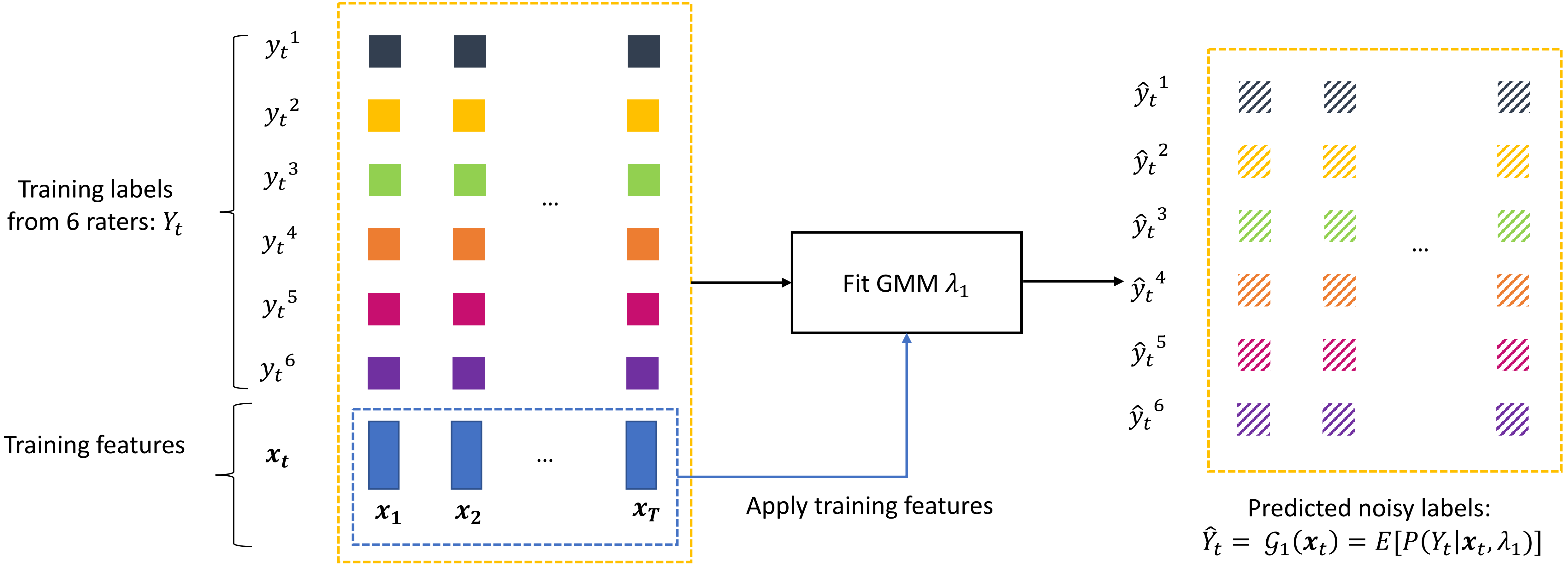
$$P(\boldsymbol{Y}_{\boldsymbol{0:m}}|\boldsymbol{t}_{\boldsymbol{0:m}}, \theta) = \prod_{r=1}^{R} P(y^r_{0:m}|\boldsymbol{t}_{\boldsymbol{0:m}}, \theta)$$

$$\log\left(P(\boldsymbol{Y}_{\boldsymbol{0:m}}|\boldsymbol{t}_{\boldsymbol{0:m}}, \theta)\right) = \sum_{r=1}^{R} \log(P(y^r_{0:m}|\boldsymbol{t}_{\boldsymbol{0:m}}, \theta))$$

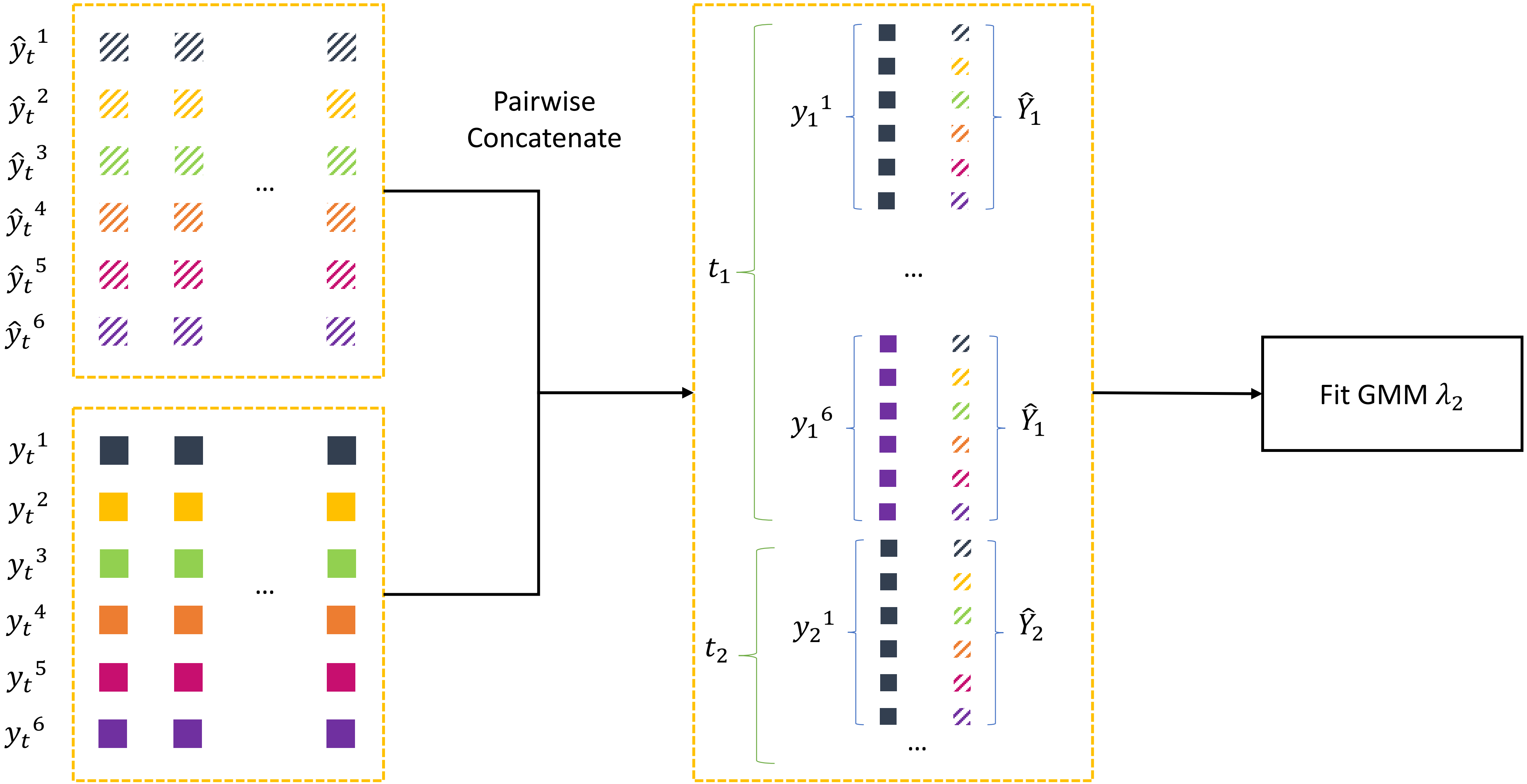
GP Mean: $m_t = \frac{1}{R}\sum_r y^r_t$

GP Covariance Matrix: $c_{t,t'} = a^2 \exp(-\frac{(t-t')^2}{2l^2})$

Gaussian Mixture Model λ_1 Training



Gaussian Mixture Model λ_2 Training



Experimental Results

Low Ambiguity High Ambiguity