Frustratingly Easy Personalization for Real-time Affect Interpretation of Facial Expression

Samuel Spaulding
MIT Media Lab
samuelsp@media.mit.edu

Cynthia Breazeal
MIT Media Lab
cynthiab@media.mit.edu

***Abstract*—In recent years, researchers have developed technology to analyze human facial expressions and other affective data at very high time resolution. This technology is enabling researchers to develop and study interactive robots that are increasingly sensitive to their human interaction partners' affective states. However, typical interaction planning models and algorithms operate on timescales that are frequently orders of magnitude larger than the timescales at which real-time affect data is sensed. To bridge this gap between the scales of sensor data collection and interaction modeling, affective data must be aggregated and interpreted over longer timescales.**

In this paper we clarify and formalize the computational task of *affect interpretation* in the context of an interactive educational game played by a human and a robot, during which facial expression data is sensed, interpreted, and used to predict the interaction partner's gameplay behavior. We compare different techniques for affect interpretation, used to generate sets of affective labels for an interactive modeling and inference task, and evaluate how the labels generated by each interpretation technique impact model training and inference.

We show that incorporating a simple method of personalization into the affect interpretation process — dynamically calculating and applying a *personalized* threshold for determining affect feature labels over time — leads to a significant improvement in the quality of inference, comparable to performance gains from other data pre-processing steps such as smoothing data via median filter. We discuss the implications of these findings for future development of affect-aware interactive robots and propose guidelines for the use of affect interpretation methods in interactive scenarios.

I. Introduction

Leading roboticists have highlighted perceptual demands as a key challenge for the field of social robotics, in part because modeling and understanding human social signals requires analyzing "extremely detailed, rapid, and nuanced signals...embedded within other activity" [29]. Advances in core pattern recognition technologies continue to improve our ability to analyze the signals with greater detail and speed; this paper focuses on developing tools to better understand and model affective signals within the context of complex social interaction between a human and a social robot.

Common families of models and algorithms for autonomous social interaction (e.g., Dynamic Bayesian Networks and Markov Decision Processes) operate on discrete timesteps or interactive 'turns' that can encompass several seconds of clock time. These models tend to treat affect as a sequence of discrete states, yet the rise of technology for high-frequency detection means that affect is sensed as a stream of continuous values. Robots and other interactive agents that rely on these models and sense real-time affective data can observe several hundred or thousand measurements of sensor data during each model timestep. There are many possible ways to interpret and summarize real-time affect data, but the most common methods typically result in either a scalar summary of an affective metric (commonly valence or engagement, e.g., when affect is used as a real-valued reward signal [14]), or a classification label (e.g. "High/Medium/Low valence") associated with each time window of relevance [25].

In this paper, we outline an information processing pipeline for integrating real-time, autonomously sensed affect data into higher-level interactive models. The pipeline is comprised of three distinct phases and associated computational tasks: affect **detection**, affect **interpretation**, and affective **inference**. Roughly speaking, affect detection involves extracting salient affective *metrics* from sensor data (e.g. video frames or sound waveforms), affect interpretation involves summarizing and synthesizing these extracted metrics over some relevant window of time into affect *feature labels*, and affective inference involves using these derived *affect feature labels* to train a model to predict some relevant aspect of a user's interactive behavior.

Though researchers have tended to focus more on studying affect detection and affective inference, understanding affect *interpretation* is of significant practical importance. Researchers of affective computing and interactive educational media have noted that "fine-grained temporal resolution affords a process-level account of engagement, disengagement, and re-engagement [...] estimates obtained [...] could be aggregated across longer periods to construct time series that reflect moments when interest was first captured, periods of maintained interest, when interest appears to diminish, etc" [7]. For interactive educational agents, this aggregation across longer time periods is critical to properly interpreting user's affective data and, consequently, providing emotionally appropriate feedback, curricular sequencing, or emotional support.

In this paper, we present results from an investigation of different techniques for affect interpretation, using data collected during a human-robot tutoring interaction, exploring the impact of smoothing and thresholding methods for

This research was supported by NSF Grant IIS-1734443 and FutureWei Technologies Inc.

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

personalization. We evaluated several candidate techniques and compared their performance on a representative affective inference task — using affect feature labels to train predictive models of children's pronunciation — to determine which methods of affect interpretation provide the best such feature labels for the inference task by evaluating the resultant models' classification performance.

A. A Frustratingly Easy Method for Personalized Affect Interpretation

Over a decade ago, Daumé III introduced a 'frustratingly easy' feature-augmentation method for domain adaptation [6]; More recently, Sun, Feng, and Saenko presented a similarly easy pre-processing method for reducing covariate shift [26] in a computer vision domain adaptation scenario. In this spirit, we detail an equally straightforward method for personalizing the affect interpretation process. In our interactive scenario, an affective metric's feature label for a given interaction window is represented as a binary indicator variable, whose value is determined by whether the maximum measurement of that metric exceeded a threshold value during the interaction window (see Sec. V-B for details on conditions). Typically these threshold values are chosen somewhat arbitrarily; in prior work researchers have selected thresholds to divide the range of metric measurements into equal bins ([9], [14]), or have chosen a threshold empirically based on validation data [18].

We dynamically calculate the mean value and standard deviation of each metric based on the detected affect data from all previously analyzed interaction windows. The threshold is set at the mean value plus twice the standard deviation of the previously sensed metric values. In the non-personalized approaches to affect interpretation, this mean and standard deviation is derived from the metric measurements from the entire participant population. In the personalized approaches, we instead derive this mean and standard deviation from just the metric measurements of the currently detected individual.

First we validate our implementation of a complete interactive system for detecting, interpreting, and performing inference on affective data by showing that a wide variety of machine learning models trained on non-personalized affect feature labels can predict a human player's word pronunciation performance significantly better than chance. We then show that dynamically calculating and using a *personalized* threshold yields affect feature labels that further improve the resultant models' classification performance. This improvement is approximately equal to the performance gain from other preprocessing methods for high-frequency affect data, such as applying a median smoothing filter.

II. Related Work

Here we review work related to socially interactive, affect-aware robots and personalized approaches to modeling affect.

A. Affect-aware interactive robots

Developing interactive robots that can detect, interpret, and make inferences from affective signals has been an active research topic for many years. However, as the ability to detect affect has improved dramatically, the core challenges have moved from detection to interpretation and interactive inference [17]. The two are often treated as a single challenge, with interpretation taking a backseat to development and evaluation of richer computational models of interaction that incorporate affective signals. With recent advances in sensor quality and computational resources, we have seen more examples of complete affect-aware interactive robot systems.

Castellano et al. [4] were able to label children's affect with a system that combined both sensed (smile detection) and annotated features. Their smile detector ran in real-time, and the probability of smile was averaged over the previous 6s window. The average probability of a smile, in combination with annotated gaze features and 10 contextual game features, was used as input to a classification system that could successfully label turns as having either $High$, $Neutral$ or Low valence, engagement, and interest.

Gordon et al. used high-frequency, median-smoothed estimates of engagement and valence from facial expressions as a reward signal for a reinforcement learning agent that personalized its supportive responses to students during a long-term tutoring interaction [9]. Park et al. developed a successor system using a similar approach [10]. Both systems incorporated (though did not highlight as a distinct step) an affective interpretation process, applying a median-filter to each detected metric vector and subsequently computing the mean metric value over the time-window (each approximately dozens of seconds). The final smoothed-average values for valence and engagement were used as input to the RL agent's reward function, and a binned version of the same measures (into High/Medium/Low measurements) were used to determine the state of the user. This smooth-and-average method has the benefit of filtering out artifacts or dropped frames in the data, but may also prevent genuine but subtle facial expressions from being detected [25]. Both systems are examples of integrated interactive systems that personalize at the level of affective inference, rather than affect interpretation.

These types of interactive social affect systems (among others) are important predecessors to our current work. They highlight how affect interpretation itself is not typically evaluated as a distinct step with respect to the overall inferential goal of an affect-aware agent. Instead the evaluative focus is frequently on the system as a whole, or only on the final step: how well the system performs affective inference, i.e., uses affective data in a model to facilitate interactive goals.

B. Personalized affect

Personalized approaches to affect detection and interpretation have been gaining prominence in recent years. In a recent survey of efforts to build 'emotionally sentient agents', McDuff and Czerwinski write "Incorporating context and personalization into assessment of the emotional state of an individual is arguably the next big technical and design challenge for [...] systems that wish to recognize the emotion of a user" [19].

Researchers have repeatedly demonstrated the benefits of a personalized approach to affect detection. In 2015, Jaques et al. trained a population level-classifier that achieved 70% accuracy in predicting student mood (happy/sad) based on affective time-series data [12]. Two years later, the authors reduced errors by 13-22% by applying a personalized multi-task learning approach to the same problem [13].

Chu, de la Torre, and Cohn introduced a personalized variant of a Support Vector Machine for facial expression analysis called a 'Selective Transfer Machine', that more heavily weighted population training samples 'close' to a new individual to learn a classifier for facial expressions [5]. Rudovic et al. demonstrated that a personalized approach to feature sharing and layer-training in a deep neural network can improve estimates of valence, engagement, and arousal during an interactive autism therapy session between a child and a robot [23]. These results and many others demonstrate the importance of personalization in affect analysis, though most focus on personalized models of affect *detection*.

Our work builds upon prior research in two ways. First, we specifically focus on personalization techniques for affect interpretation, rather than for affect detection or affective inference. Second, our evaluation centers on how the interpreted labels can be used in an interactive inference task: predicting children's word pronunciation from affective data only during an educational game. Most evaluations of personalized affect models interpretation focus on inter-class correlation with human-coder ratings. Although human coding is a useful benchmark for evaluating affect detection and interpretation modules, expressions in an interactive context play a deeper role than merely signaling internal affective state, they communicate information that is intricately tied to the interaction (i.e., information that establishes affective ground [15]), which even trained human coders may not fully pick up on without being an active participant. Our ultimate research goal is to develop affective interpretation methods for the purpose of *modeling and generating interactive behavior*, thus our evaluation focuses on validating how useful the interpreted affect labels are to interactive agent algorithms and models that predict or estimate the probability of user actions.

III. An Information-Processing Pipeline for Situated, Interactive Affect Data

Emotional and affective displays are a key channel for social communication. Yet due to challenges in sensing, interpreting, and modeling these signals, this channel remains underutilized in human-robot interactions compared to human-human interactions. Recent advances in pattern recognition and multimodal sensing, largely fueled by the impressive performance of deep neural networks, have helped tools for high-frequency affect sensing to become more widely used in research and commercial settings.

The high time-resolution nature of modern tools for analyzing facial expression data enables researchers to detect very subtle changes in expression (e.g., facial 'microexpressions' [21]), but simultaneously suggests an additional challenge: how best to summarize the large amount of high-frequency expression data over an order-of-magnitude difference in time scale? Recent projects have generally accomplished this task by summary statistic heuristics, such as the mean value of a metric over the relevant window [14], whether the mean value of a metric over the relevant window exceeded a threshold value [1], or whether at any point in the window the value of a metric exceeded some threshold [30].

A priori, each of these heuristics appears to be a reasonable way to determine if an interaction partner displayed some expression during an interaction time window. However, there has not yet been any systematic investigation or comparison of how these heuristic methods for affect interpretation impact the interaction models and algorithms that they feed into. We introduce a pipeline for affect detection, interpretation, and inference used to examine the effect of smoothing and personalized thresholding on affect interpretation.

Much of our pipeline and terminology overlaps with the approach detailed by D'Mello, Kappas, and Gratch [8]. Their approach outlines a process for computing machine-sensed affect *estimates* or *annotations*, binary classifications of whether complex affective states (e.g. boredom, confusion, delight, etc) were present over some window of time, with ground-truth typically determined by human annotations. These affect annotations are analogous to the *affect feature labels* that are the output of the affect interpretation step in our pipeline.

However, our implemented approach differs in a few key respects. Partly due to the difficulty of obtaining reliable human annotations for complex affective states [11], we restrict our evaluation of affect interpretation to methods that do not require outside human expertise. Therefore, rather than use the detected and interpreted features to predict a discrete emotion label (such as 'confused' or 'cogitating'), then use the discrete label to predict behavior, we consider the interpreted affect feature labels a latent representation of emotion, and attempt to predict behavior directly from the affective features themselves. Consequently, rather than evaluating the interpreted affect feature labels by how well they correspond to human-annotated labels, we evaluate the interpreted affect feature labels by how well various models trained on the feature labels predict interaction behavior. This design choice is further intended to highlight the evaluation requirements of interactive *agents*, for which interpreted affective feature labels are useful insofar as they can be used to make inferences about an interaction partner or about what action the agent should take. The recognition of affective *interpretation* as a distinct computational task within the pipeline and the evaluation of that task with respect to the quality of affective *inference* are key principles of our analysis.

A. *Definitions and Terminology*

Affect *interpretation* is the process by which high-frequency sensed affective data over a given time period is aggregated and summarized for some further purpose. In this paper, we consider the domain of facial expression analysis, in which modern detection algorithms can produce estimates of the degree to which multiple facial descriptors of affective states (joy, sadness, frustration) are expressed in a single frame. We restrict our analysis to the interpretation of facial expression data, but emphasize that many of the salient issues are relevant to other modalities, such as nonverbal body posture, in which high-frequency analysis is enabled by pose-detection tools such as OpenPose [3], or analysis of affect from speech or physiological signals.

The full data pipeline, with relevant terminology highlighted is as follows: Raw data, $D = d_o, d_1, ...d_n$, is sensed over an interaction time ***window*** W (in seconds), where $|D| >> W$. The data, D, is analyzed by an affect ***detector*** (e.g. Affdex or Openpose), which outputs real-valued feature vectors of higher-level ***metrics***, $M = m_0, m_1, ...m_n$ (e.g., the degree to which 'smile' 'surprise', 'brow furrow' etc. are expressed), for each data point (in this case, a single video frame). The affect ***interpreter*** analyzes this set of metric vectors over the ***window***, and the interpreter outputs a vector-valued ***feature label***, l, for that window. This affect feature label is then used to train a predictive model for *affective inference*, an inferential task in which affective data is used to make predictions about some aspect of an individual or interaction.

With these definitions in place, we now state our research contributions more concretely. We developed an interactive social robot system that implements this pipeline, specifically focused on detecting, interpreting, and using facial affect to infer students' pronunciation ability during an interactive game with a robot (Fig. 3). We used Affdex [20] as the affect detector, different affect interpretation methods to derive interpreted affect feature labels, and evaluated these interpreted affect feature labels based on how well they predict correct word pronunciation.

We studied 3 different methods for affect interpretation (see Sec. V-B), and trained 5 common machine learning models on the affect feature labels produced by each method to evaluate the impact of personalization on affect interpretation. We show our pipeline can successfully predict children's pronunciation ability from affective features at above-chance rates, and that even simple methods of personalization significantly improve the quality of the interpreted labels with respect to their use as training data for an interactive inference task.

IV. Task and Dataset Collection

In this section we describe the data collection process, experimental task, and other details. 6 (5 Male, 1 Female, Age: 6±1.1) of the participants completed the experiment in the lab, and another 9 (5 Male, 4 Female, Age: 5±0) completed the same experiment at a local school approximately 6 weeks later. 10 of the 15 children spoke another language in addition to English. Other than the location, the experimental protocol, game system, and experimental setup were identical.

A. Game Interaction Overview

The facial expression dataset was derived from video footage during an interactive game called WordRacer, intended to assess children's pronunciation skills. Children sat across from a small-sized expressive social robot with a tablet in between them. At the start of each round of gameplay, a printed word and a corresponding picture of that word (e.g., a picture of a goat if the word were 'GOAT') would appear in the center of the tablet.

At the top and bottom of the tablet screen were two buttons, which children were told were the "buzzers" they and the robot would use to signal their knowledge of the word. Whichever player rang in first would get a chance to read the word on the screen, and a pronunciation that was deemed correct would be awarded a point. Each child played the game until they had given 20 responses or until they indicated their desire to stop playing. The robot was framed as a co-playing peer to the child and played the game, ringing in, pronouncing words, and accumulating points, according to an active learning protocol. For more detailed information on the game task and data collection protocol, see Spaulding et al. (2018) [24].

V. Processing Pipeline and Models

A. Affect Detection

During the game, facial expressions were sensed from a front-facing USB camera stream at 30fps during the interaction. We used a stationary camera, so some frames did not contain a detectable face, a common drawback of fixed-sensor methods. We calculated exactly when each round of the game started (i.e. when a new word graphic was shown on screen) and kept sensing until either the child or robot rang in to give their answer. Thus, we expect the detected affect from each round to be in direct response to the presented content of the current round, not the larger context of game results. Finally, we only analyzed rounds in which the child rang in first, and therefore was expected to pronounce a word. We processed frames from each round via the Affdex Auto-SDK, which detected 12 relevant metrics at approximately 30fps: **'smile', 'anger', 'valence', 'browFurrow', 'noseWrinkle', 'joy', 'surprise', 'browRaise', 'upperLipRaise', 'mouthOpen', 'eyeClosure', 'cheekRaise'**.

B. Affect Interpretation: Smoothing and Personalization

One of the biggest challenges of developing models of affect is the diversity and idiosyncrasy of affective expressions. Individuals' levels of emotional expression are highly varied based on personality, culture, situation, and countless other details. The complexity and variety of this challenge is one of the strongest reasons we advocate for a personalized approach to affect interpretation.

As described in Section III, we use the term *affect interpretation* to denote the process by which high-frequency sensed affective data from a particular time window is collected and converted into a label or feature vector suitable for further use by an interactive agent for higher-level modeling or inference. In this paper we interpret each detected metric separately, and the output of each interpreted metric is a binary indicator variable with a value determined by different smoothing and thresholding techniques.

We analyzed 3 different methods for affect interpretation to examine the impact of two affect interpretation techniques:

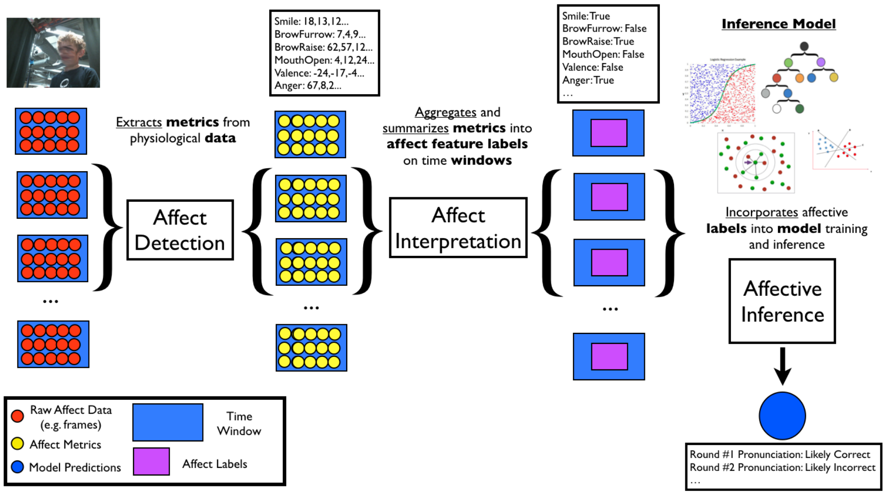

Fig. 1. A pipeline for integrating affective data with interaction models. Affective data is sensed at high-frequency and a Detector extracts metrics from each data point. An Affect Interpreter summarizes the metrics within a given window of time, producing a higher-level label for each timestep. Those labels are used as data for Affective Inference, using interpreted affect features to infer something about the user or interaction, such as cognitive state or skill mastery.

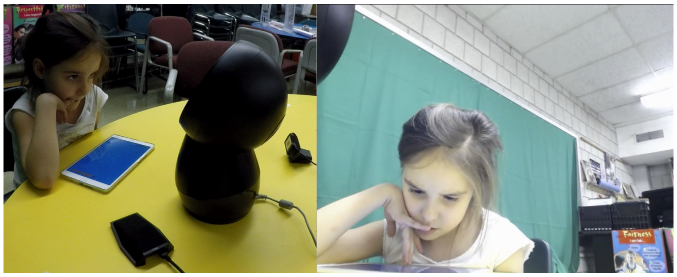

Fig. 2. Two camera angles showing experimental setup. Right image shows camera stream used to analyze facial affect

median-filter smoothing of detected metric values and the use of a personalized threshold for interpretation. The 3 conditions, plus a baseline condition of randomly generated labels are denoted: *Unsmoothed, NPSmoothed, PSmoothed*, and *Random*. E.g., in the PSmoothed condition, if the maximum value of the median-smoothed detected metrics exceeds the personalized threshold, that metric's interpreted value is 1. For each time window, the final interpreted affect feature vector is a multi-hot vector composed of each metric indicator variable, which is then used as training input to a variety of standard machine learning models during the affective inference stage. The threshold value for a given metric is set at the mean value of that metric plus two standard deviations. For non-personalized methods, the mean and standard deviations are based on the detected values from all participants (i.e. the sample mean and variance); the mean and variance for personalized thresholds are computed from only that specific individual's detected metric values. Despite this simple change, we find that labels generated through a personalized threshold can help train significantly better models for affective inference.

C. Affective Inference task: predicting children's pronunciation from facial affect

We evaluated the quality of the interpreted affect feature labels by using them as input to predict children's pronunciation ability during successive rounds of an interactive pronunciation game with a robot. In each turn, children had an opportunity to pronounce a presented word. We recorded their speech and analyzed it via a third party pronunciation analysis software provided by Soapbox Labs, which takes in the expected word and a speech sample and provides a score for how well that word was pronounced. Each child played until they had pronounced 20 words. For each round, the input vector was a binary vector of the indicator variables, the output of the affect interpretation step, and the supervised label was whether or not the child pronounced the word correctly (operationalized as a pronunciation score above 70; 100 is the maximum).

We conducted a leave-one-out cross-fold validation with 5 common machine learning models: Logistic Regression, Random Forest, Naive Bayes, SVM (with an RBF kernel) and K-Nearest Neighbors (k=4), and report the per-fold average area under the ROC curve (AUC-ROC) alongside the standard error of the mean for each class of model and each method of affect interpretation.

VI. Model Evaluation and Results

These results come from a relatively small, complex, dataset, largely collected from untrained subjects in a real classroom. As a result it features many artifacts common to such scenarios, including periodic occlusion of the face by the child's hands, dropped frames from rapid head movements and orientation change, and unpredictable one-time occurrences, such as lights turning on or off. Despite these challenges, we are able to learn an affective inference model that meets the standard described by D'Mello et al. "that a model is *above-chance* accurate" [8]. Having met that initial bar to validate the overall pipeline, we now look at the impact of personalization and smoothing on the quality of the interpreted labels.

Models trained on unsmoothed, non-personalized labels perform slightly better than random labels, with an all-model average absolute improvement of 2.3% (4.7% relative) and higher absolute performance on 3 of 5 tested models. Smoothed, non-personalized models perform much better, matching or outperforming both non-smoothed and random methods on all 5 models, with an average absolute

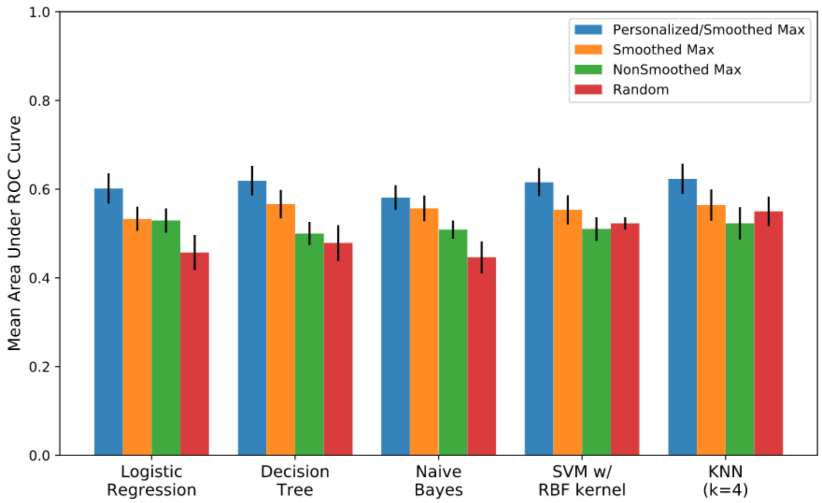

Fig. 3. Area under the ROC curve for a variety of models trained on differently-generated label sets. Error bars represent standard error of the mean

TABLE I

AREA UNDER ROC CURVE (MEAN ± SEM) FOR MODELS TRAINED ON SETS OF AFFECTIVE FEATURE LABELS DERIVED FROM PERSONALIZED (P), NONPERSONALIZED SMOOTHED (NPS), UNSMOOTHED (US), AND RANDOM METHODS FOR AFFECT INTERPRETATION

Condition	LogReg	Random Forest	Naive Bayes	SVM (RBF kernel)	KNN (k=4)
PS	0.6±0.03	0.62±0.03	0.58±0.03	0.62±0.03	0.62±0.03
NPS	0.53±0.03	0.57±0.03	0.56±0.03	0.55±0.03	0.56±0.04
US	0.53±0.03	0.5±0.03	0.51±0.02	0.51±0.03	0.52±0.04
Random	0.46±0.04	0.48±0.04	0.45±0.04	0.52±0.01	0.55±0.03

improvement of 4.0% (7.8% relative) over the unsmoothed interpretation condition. Personalized, smoothed interpretation improves even further, outperforming every other interpretation method on all 5 tested models and showing a 5.3% absolute improvement (9.6% relative) over non-personalized, smoothed condition, exceeding both the absolute and relative performance increase from median smoothing.

VII. Discussion and Conclusions

Based on these results, we suggest that this approach to personalization can and should be incorporated by any researchers who face a choice of affective interpretation techniques when analyzing high-frequency affect data. Our framing of personalization refers to the affective interpretation (i.e., label generation) process; during affective inference the interpreted features are used to learn a general (i.e., non-personalized) model. This is an inversion of the approach of multi-task learning with deep neural networks, in which early layers, typically associated with feature extraction and representation, are shared, and later layers are 'personalized' or 'fine-tuned' separately to be task-specific. This and other methods of personalization rely on weighting of selectively including personalized data in the training of a model (e.g, group-level information such as demographic features or personal data collected from an individual). In contrast, the personalized thresholding method we describe does not preclude other methods of personalization, and is incredibly simple, easy to run in real-time, and does not require re-training a complex model as new data is sensed.

Interestingly, this method for affect interpretation with a personalized threshold can be viewed as a form of within-participant z-normalization, a pre-processing technique for time-series data. Z-normalization has been used to improve affect interpretation on speech signals ([16], [27]) and behavioral data (e.g. mouse trajectories [28]), but has not been closely studied for analysis of facial expression. While normalization of structural facial features is a common pre-processing step for affect detection, we show that z-normalization can be an easily and beneficial method for personalization of facial expression interpretation.

In this paper, we have presented results from one of the first systematic investigations of techniques for affect interpretation. By naming and outlining a process that many have worked on implicitly, we hope to help identify common patterns and techniques (e.g. median smoothing of facial metrics) to standardize and improve results in the field. We also demonstrated a 'frustratingly easy' technique for personalized affect interpretation that is simple to implement and improves the quality of interpreted affect labels for subsequent inference.

The pitfalls of a 'one-size-fits-all' approach to human data analysis are especially pronounced when analyzing human faces [2], [22]. Our results suggest a method that can be easily adopted during affect interpretation to improve system performance for individual users. Diverse human interactions with affect-aware systems are becoming more common; careful study of how affective data is treated in each step of a complex system is critical to avoid biased or inaccurate conclusions, and personalization of affect-aware technology to individuals or groups may be one path towards ensuring their benefits will be equitably shared.

Acknowledgment

The authors would like to thank Affectiva for providing tools for affect detection, to Soapbox Labs for providing us an accurate tool for children's pronunciation assessment, and to the reviewers for their helpful comments and suggestions.

References

[1] A. Bernin, L. Müller, S. Ghose, K. von Luck, C. Grecos, Q. Wang, and F. Vogt. Towards more robust automatic facial expression recognition in smart environments. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 37–44. ACM, 2017.

[2] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. Mcowan. Context-Sensitive Affect Recognition for a Robotic Game Companion. *ACM Transactions on Interactive Intelligent Systems*, 4(2):1–25, 2014.

[5] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial expression analysis. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):529–545, 2017.

[6] H. Daume III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, 2007.

[7] S. D'Mello, E. Dieterle, and A. Duckworth. Advanced, analytic, automated (aaa) measurement of engagement during learning. *Educational psychologist*, 52(2):104–123, 2017.

[8] S. D'Mello, A. Kappas, and J. Gratch. The Affective Computing Approach to Affect Measurement. *Emotion Review*, 10(2):174–183, 2018.

[9] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[10] S. S. L. G. Hae Won Park, Ishaan Grover and C. Breazeal. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[11] M. Jaiswal, Z. Aldeneh, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. *arXiv preprint arXiv:1903.11672*, 2019.

[12] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 222–228. IEEE, 2015.

[13] N. Jaques, S. Taylor, A. Sano, and R. Picard. Predicting tomorrow?s mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pages 17–33, 2017.

[14] S. Jeong and C. L. Breazeal. Improving smartphone users' affect and wellbeing with personalized positive psychology interventions. In *Proceedings of the Fourth International Conference on Human Agent Interaction, HAI 2016, Biopolis, Singapore, October 4-7, 2016*, pages 131–137, 2016.

[15] M. F. Jung. Affective grounding in human-robot interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 263–273. ACM, 2017.

[16] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.

[17] J. J. Lee, F. Sha, and C. Breazeal. A bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 487–496. IEEE, 2019.

[18] A. Lopez-Rincon. Emotion recognition using facial expressions in children using the nao robot. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 146–153. IEEE, 2019.

[19] D. McDuff and M. Czerwinski. Designing Emotionally Sentient Agents. *Communications of the ACM*, 2018.

[20] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3723–3726. ACM, 2016.

[21] Y.-H. Oh, J. See, A. C. Le Ngo, R. C. W. Phan, and V. M. Baskaran. A survey of automatic facial micro-expression analysis: Databases, methods, and challenges. *Frontiers in Psychology*, 9:1128, 2018.

[22] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, 2019.

[23] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19), 2018.

[24] S. Spaulding, H. Chen, S. Ali, M. Kulinski, and C. Breazeal. A social robot system for modeling children's word pronunciation: Socially interactive agents track. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1658–1666. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[25] S. Spaulding, G. Gordon, and C. Breazeal. Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 864–872. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[26] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[27] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, 2015.

[28] T. Yamauchi and K. Xiao. Reading emotion from mouse cursor motions: Affective computing approach. *Cognitive science*, 42(3):771–819, 2018.

[29] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, B. J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. Veloso, Z. L. Wang, and R. Wood. The grand challenges of science robotics. *Science Robotics*, 3(14), 2018.

[30] L. Zhang, D. Tjondronegoro, V. Chandran, and J. Eggink. Towards robust automatic affective classification of images using facial expressions for practical applications. *Multimedia Tools and Applications*, 75(8):4669–4695, 2016.