Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Predicting Latent Narrative Mood Using Audio and Physiologic Data

Tuka AlHanai and Mohammad Mahdi Ghassemi*
Massachusetts Institute of Technology, Cambridge MA 02139, USA
tuka@mit.edu, ghassemi@mit.edu

Abstract

Inferring the latent emotive content of a narrative requires consideration of para-linguistic cues (e.g. pitch), linguistic content (e.g. vocabulary) and the physiological state of the narrator (e.g. heart-rate). In this study we utilized a combination of auditory, text, and physiological signals to predict the mood (happy or sad) of 31 narrations from subjects engaged in personal story-telling.

We extracted 386 audio and 222 physiological features (using the Samsung Simband) from the data. A subset of 4 audio, 1 text, and 5 physiologic features were identified using Sequential Forward Selection (SFS) for inclusion in a Neural Network (NN). These features included subject movement, cardiovascular activity, energy in speech, probability of voicing, and linguistic sentiment (i.e. negative or positive). We explored the effects of introducing our selected features at various layers of the NN and found that the location of these features in the network topology had a significant impact on model performance.

To ensure the real-time utility of the model, classification was performed over 5 second intervals. We evaluated our model's performance using leave-one-subject-out cross-validation and compared the performance to 20 baseline models and a NN with all features included in the input layer.

Introduction

Human communication depends on a delicate interplay between the emotional intent of the speaker, and the linguistic content of their message. While linguistic content is delivered in words, emotional intent is often communicated though additional modalities including facial expressions, spoken intonation, and body gestures. Importantly, the same message can take on a plurality of meanings, depending on the emotional intent of the speaker. The phrase "Thanks a lot" may communicate gratitude, or anger, depending on the tonality, pitch and intonation of the spoken delivery.

Given it's importance for communication, the consequences of misreading emotional intent can be severe, particularly in high-stakes social situations such as salary negotiations or job interviews. For those afflicted by chronic social disabilities such as Asberger's syndrome, the inability to read subtle emotional ques can lead to a variety of negative consequences, from social isolation to depression (Müller, Schuler, and Yates 2008; Cameron and Robinson 2010). Machine-aided assessments of historic and real-time interactions may help facilitate more effective communication for such individuals by allowing for long-term social coaching and in-the-moment interventions.

In this paper, we present the first steps toward the realization of such a system. We present a novel multi-modal dataset containing audio, physiologic, and text transcriptions from 31 narrative conversations. As far as we know, this is the first experimental set-up to include individuals engaged in natural dialogue with the particular combination of signals we collected and processed: para-linguistic cues from audio, linguistic features from text transcriptions (average postive/negative sentiment score), Electrocardiogram (ECG), Photoplethysmogram (PPG), accelerometer, gyroscope, bio-impedance, electric tissue impedance, Galvanic Skin Response (GSR), and skin temperature.

The emotional content of communication exists at multiple levels of resolution. For instance, the overall nature of a story could be positive but it may still contain sad moments. Hence, we present two analyses in this paper. In the first analysis, we train a Neural Network (NN) to classify the overall emotional nature of the subject's historic narration. In the second analysis, we train a NN to classify emotional content, in real-time. We also show how the optimization of network topology, and the placement of features within the topology improves classification performance.

Literature Review

Cognitive scientists indicate that emotional states are strongly associated with quantifiable physical correlates including the movement of facial muscles, vocal acoustics, peripheral nervous system activity, and language use (Barrett, Lindquist, and Gendron 2007). The detection of a latent emotional state by a machine agent is strongly influenced

*Both authors contributed equally to this work. Tuka AlHanai would like to acknowledge the Al-Nokhba Scholarship and the Abu Dhabi Education Council for their support. Mohammad Ghassemi would like to acknowledge the Salerno foundation, The NIH Neuroimaging Training Grant (NTP-T32 EB 001680), the Advanced Neuroimaging Training Grant (AMNTPT90 DA 22759). The authors would also like to acknowledge Hao Shen for aiding with data labeling, and Kushal Vora for arranging access to the Simbands.

by both the number of physical correlates available (e.g. audio alone, versus audio and visual), the context in which the correlates are observed (e.g. enhanced heart rate during fear versus excitement) (Barrett, Mesquita, and Gendron 2011), as well as the social context in which the conversations take place (e.g. a sports stadium versus a meeting room) (Rilliard et al. 2009).

Importantly, the interactions between these physical correlates also have unique associations with the latent emotional states. For instance, a combination of high heart rate and voicing is associated with excitement although neither are independently discriminative. It follows that emotive estimation is aided by (1) access to multiple data modalities, and (2) the utilization of techniques which can account for the complex interactions between those modalities.

Existing data in the domain of emotion detection have been collected using a variety of tasks including the presentation of images, video clips, and music. Data has also been collected though problem solving tasks, facial expressions exercises, acting, and scripted speech (Douglas-Cowie et al. 2003; Calvo and D'Mello 2010). The highly controlled nature of these studies enhances the ability of investigators to identify features which relate to emotive state at the cost of practical utility. That is, there is relatively little work on emotive description of spontaneous human interactions in a natural setting (Koolagudi and Rao 2012).

With respect to data analysis, existing studies have applied techniques ranging from regression and Analysis of Variance (ANOVA) to Support Vector Machines (SVM), Clustering and Hidden Markov Modeling (HMM). With the recent interest in 'Deep' learning, Neural Networks are also increasingly utilized in emotion detection for speech (Stuhlsatz et al. 2011; Li et al. 2013; Han, Yu, and Tashev 2014), audio/video (Kahou et al. 2013; Wöllmer et al. 2010), audio/video/text (Wöllmer et al. 2013), and physiologic data (Haag et al. 2004; Wagner, Kim, and André 2005; Walter et al. 2011; Katsis et al. 2008). Many of the surveyed studies, however, utilize only a single modality of data (text, audio, video, or physiologic) for the inference task. While these studies succeed in enhancing knowledge, a real-world implementation will require the consideration of multiple data modalities, with techniques that account for multiple levels of interaction between the modalities in real-time, just as humans do (D'Mello and Kory 2012).

Methods

Data Collection

Subject Recruitment Twenty individuals from the Massachusetts Institute of Technology (MIT) were invited to participate in this study. Prior to data collection, all participants were informed of the experimental protocol, the goals of the study, and were told that they could opt-out of the study at any time, for any reason. Participating subjects were asked to provide both written and oral consent for the use of their de-identified data. 50% of the invited individuals agreed to participate in the study. The average age of the 10 participating individuals was 23. Four participants identified as male, and six participants identified as female. All ten individuals listed English as their primary language.

Data	*Quantity*
Subjects	10 (4 male, 6 female)
Narratives	31 (15 happy, 16 sad)
Average Duration	2.2 mins
Total Duration	67 mins
5 Sec Segments	804
Features	*Quantity* *Total (selected)*
Physiologic	222 (5)
Audio	386 (4)
Text	2 (1)

Table 1: Under the Data heading we display information on the total number of subjects, narratives, and samples used in the study. Under the Features heading we provide information on the total number of features collected in each of the modalities (physiologic, audio and text), and the proportion of the features selected for inclusion in our model.

Experimental Venue and Approach The experimental venue was a 200 square foot temperature and light controlled conference room on the MIT campus. Upon arrival to the experimental venue, participants were outfitted with a Samsung Simband, a wearable device which collects high-resolution physiological measures. Audio data was recorded on an Apple iPhone 5S.

Next, participants were provided with the following experimental prompt: *"In whatever order you prefer, tell us at least one happy story, and at least one sad story."*. If subjects asked for additional clarification about story content, they were informed that there was *"no correct answer"*, and were encouraged to tell any story they subjectively found to be happy or sad. A summary of the collected data is presented under the *Data* heading in Table 1.

Time-Series Segmentation

Collected data was time-aligned and segmented using 5 second non-overlapping windows. Our window size was selected such that the minimum number of spoken words observed within any given segment was two or more. This criteria was necessary to evaluate the effects of transcribed text features. Smaller window sizes (e.g. 1 second) resulted in segments which contained only partial words, making the analysis infeasible. Importantly, we do not anticipate large fluctuations in speaker emotional state within the 5 second windows as (Schuller, Rigoll, and others 2006) reported that the emotive content of speech is stable within windows sizes as large as 8 seconds.

Each 5 second window of our data was manually transcribed, and annotated for emotional content by a research assistant. For emotional annotation, the research assistant was asked to rate the emotive content of each 5 second audio segment according to the four axis of a simplified Plutchiks emotion wheel: happy-sad, interested-bored, admiration-loathing and angry-afraid (Plutchik 2001). The coding scheme is illustrated in Table 2.

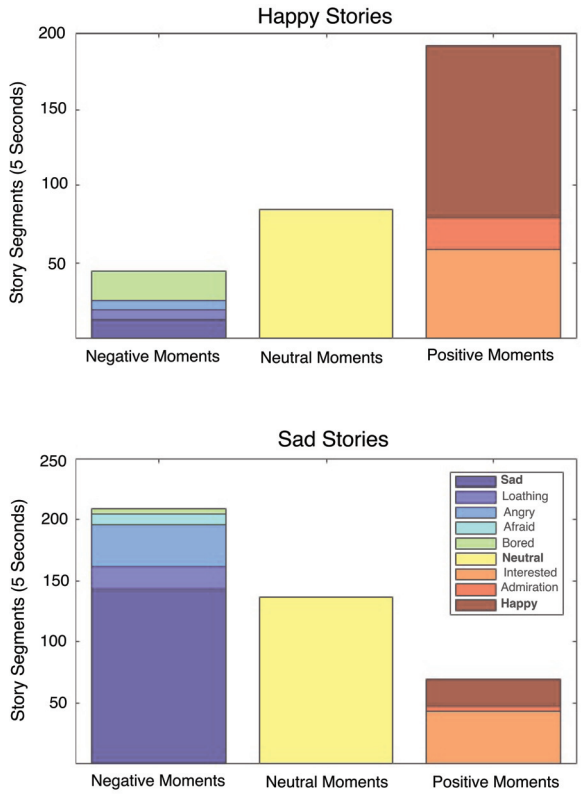

Figure 1: Distribution of emotional content within 5 second windows of collected happy and sad narrations. Negative segments refer to those which were found to contain sadness, loathing, anger, fear or boredom. Positive segments refer to those which were found to contain happiness, interest, or admiration. Neutral segments refer to those which were neither positive nor negative.

The segment-level annotations were grouped into three more general classes: positive, negative, and neutral. Negative segments refer to those containing sadness, loathing, anger, fear or boredom. Positive segments refer to those containing happiness, interest, or admiration. Neutral segments refer to those which were neither positive nor negative. See Figure 1 for a breakdown of our collected data by story type, segment class, and segment value.

Coding	**-1**	**0**	**1**
Happy-Sad	Sad	Neutral	Happy
Interested-Bored	Bored	Neutral	Interested
Admiration-Loathing	Loathing	Neutral	Admiration
Angry-Afraid	Afraid	Neutral	Angry

Table 2: The coding scheme used to assessing the emotional content of 5 second segments of the collected audio.

Feature Extraction

Within each 5 second segment, a set of candidate features for the model were extracted from the Simband sensor data stream, as well as the recorded audio data. All features were normalized to a zero-mean and unit variance representation. All features with zero-variance, or those that were missing in 1% or more of the total observations were excluded. The total number of features in each of the modalities are summarized under the *Features* heading in Table 1. We also provide additional detail on the extracted features below.

Physiologic Features Within each 5 second window we extracted the mean, median, and variance of the 91 available streams of Simband data (for a total of 222 Simband candidate features). Streams included Electrocardiogram, Photoplethysmogram, accelerometer, gyroscope, bio-impedence, electric tissue impedance, Galvanic Skin Response, and skin temperature. See the Supplementary Materials for an overview of the Simband data streams.[1]

Audio Features Within each 5 second segment we also extracted para-linguistic audio features. Para-linguistic features were extracted based on the feature set presented at the INTERSPEECH 2009 Emotion Challenge, and using the openSMILE Toolkit (Schuller, Steidl, and Batliner 2009; Eyben, Wöllmer, and Schuller 2010). Low level descriptors were the RMS energy, Mel Frequency Cepstral Coefficients (MFCCs) 1-12, Pitch (F0), Zero Crossing Rate (ZCR), and voicing probability. A total of 384 features defined as functionals were derived from low level descriptions of the speech signal. These functionals were the mean, standard deviation, skewness, kurtosis, maximum value, minimum value, range, absolute position of minimum/maximum values, linear regression coefficients slope/offset, and linear regression Mean Squared Error (MSE).

Text Features Within each 5 second segment, linguistic features were extracted to capture the emotive content of words. More specifically, the audio was manually transcribed and the average positive and negative sentiment of all words in each 5 second window were calculated using the SentiWordNet Lexicon (Esuli and Sebastiani 2006).

Feature Selection

We used the sequential forward selection algorithm to identify the subset of candidate features with the greatest importance for predicting the narrative class (happy or sad) in a logistic regression model. The performance criteria of the forward selection algorithm was improvement in classification performance on a held out validation set. To ensure the robustness of the identified features, the forward selection algorithm was performed on ten folds of our dataset (90% training, 10% validation) and a feature was marked for inclusion in our models only if it was selected in 5 or more of the folds.

Experimental Approach

Using the selected features, we performed two analyses. In the first analysis, we trained a Neural Network model to classify the overall nature of a narration, as reported directly by the subject: happy or sad. In the second analysis, we trained

[1] http://people.csail.mit.edu/tuka/aaai-17-supp-materials.pdf

Selected Features	Signal Type	β	Odds Ratio [95% CI]	p-value
Accelerometer Y-axis (mean)	Physiologic	-1.75	0.17 [0.11- 0.25]	1.82e-18
Accelerometer Z-axis (mean)	Physiologic	-2.28	0.10 [0.06 - 0.17]	1.22e-16
MFCC 3 Differential (min)	Para-linguistic	0.63	1.88 [1.37 - 2.57]	7.26e-05
MFCC 12 (kurtosis)	Para-linguistic	-0.47	0.63 [0.44 - 0.89]	7.82e-03
Negative Sentiment	Linguistic	-0.50	0.61 [0.45 - 0.81]	5.15e-04
PPG Sensor #0	Physiologic	2.84	17.12 [7.36 - 39.8]	4.21e-11
PPG Sensor #1	Physiologic	-2.38	0.093 [0.04 - 0.20]	2.58e-09
PPG Signal (mean)	Physiologic	-1.80	0.17 [0.08 - 0.36]	3.75e-06
RMS Energy (mean)	Para-linguistic	1.82	6.17 [3.85 - 9.88]	1.17e-13
Voicing Probability (linear regression MSE)	Para-linguistic	0.66	0.16 [1.41 - 2.65]	2.57e-05

Table 3: The ten features chosen by the Sequential Forward Selection Algorithm from the 610 candidate features, over 10 folds. Features from all three modalities (physiologic, para-linguistic, and linguistic) were found to be significantly associated with happy/sad narrations. β: Estimate of coefficients

a NN model to classify the segment-level emotional class annotated by the research assistant: positive, negative, or neutral.

Neural Network Optimization

While the power of NNs as a classification tool is a well-established, the performance of NNs are dependent on both the initial settings of network weights, and the topology of the network itself (number of hidden layers, and nodes within those layers). To account for this dependence, we optimized both the network topology, and the location of our selected features within the topology. More specifically, we trained all possible configurations of a NN with a number of hidden layers between 0 and 2 (where 0 hidden layers corresponds to a logistic regression). We also explored all possible configuration of the nodes across the hidden layers such that the ratio of network weights to training points was less than or equal to ten (Friedman, Hastie, and Tibshirani 2001). We also explored a random 10% of all 3^{10} possible locations of our selected features within the NN structure

For the segment-level classification task, we also explored the effects of accounting for state-transition effects by estimating a state transition probability matrix (Markov assumption), and adjusting the probabilistic outputs of our model, given the prior state estimate.

Model Validation

To ensure the robustness of our approach, we compared the optimized narrative-level NN to several other supervised machine learning classifiers including: Linear and Quadratic Discriminant Analysis, Decision Trees, Naive Bayes, k-Nearest Neighbors, Support Vector Machines, Bagged Trees and Boosted Trees. The optimized segment-level NN was compared to a multinomial logistic regression baseline (Our reasons for selecting this baseline are explained in the results section).

All models were validated using leave-one-subject-out cross-validation, where the models were trained on all but one subject's data segments, and then tested on the remaining individual's segments. We compared the performance of our optimized NNs to the baseline approaches on the held-out subject, in each fold. The performance of narrative-level models was compared using the Area Under the Receiver Operating Curve (AUC), because we wanted to present the performance of the model across all potential misclassification costs. The performance of the segment-level models was compared using Accuracy given that the overall incidence of positive, negative, and neutral moments was near balanced (262 positive, 257 negative, 285 neutral).

(A) NARRATIVE	**AUC** (μ)	**AUC** (σ)	**Percentile [25th 75th]**
Quadratic Disc.	0.83	0.06	[0.79 0.89]
Gaussian SVM	0.86	0.07	[0.80 0.93]
Linear Disc.	0.90	0.07	[0.83 0.95]
Subspace Disc. Ens.	0.90	0.06	[0.85 0.95]
NN (0 hidden layers)	0.92	0.05	[0.88 0.95]
(B) SEGMENT	**Acc. (%)** (μ)	**Acc. (%)** (σ)	**Percentile [25th 75th]**
MLR	40.8	7.36	[34.1 46.0]
NN (2L-6x3N)	45.3	8.10	[38.5 49.0]
+ Feature Opt.	47.3	8.72	[39.9 55.1]

Table 4: A comparison of the NN model to several baseline approaches for (A) narrative-level classification and (B) segment-level classification. KNN: k-Nearest Neighbors. SVM: Support Vector Machine. Disc: Discriminant Analysis. Ens: Ensemble. MLR: Multinomial Logistic Regression. NN: Neural Network. Opt: Optimized.

Results

The cross-validated forward selection algorithm identified 10 features for inclusion in the NN model. The selected features spanned all three modalities; Simband physiologic features (mean accelerometer and mean PPG signal) as well as audio para-linguistic (MFCC 2 differential, MFCC 12, RMS energy, and voicing probability) and linguistic (negative sentiment) features.

Narrative-Level

In Table 4(A), we present the results of best performing NN alongside the 20 baseline approaches for the narrative-level classification. After optimization, the narrative-level NN converged to a logistic regression model with a mean

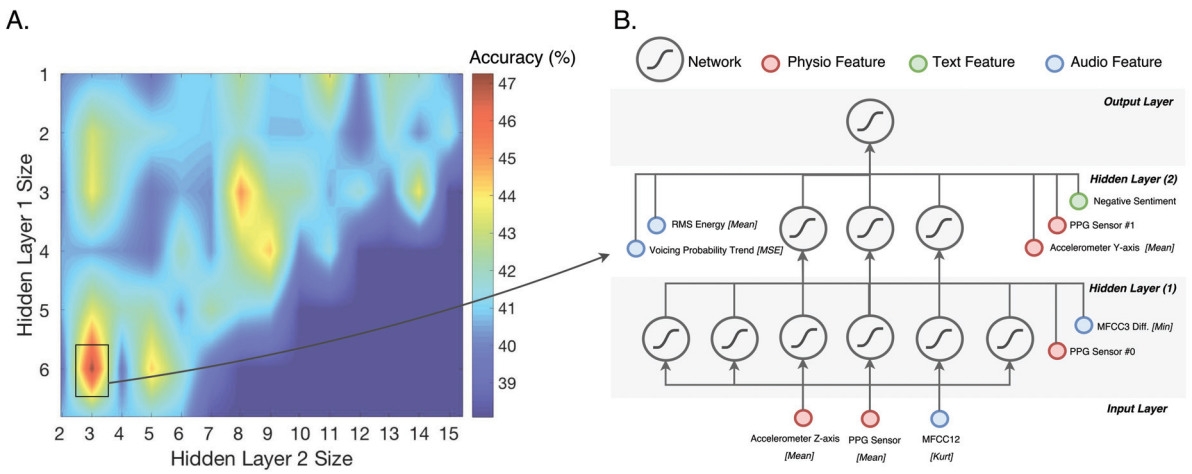

Figure 2: (A) A heatmap of segment-level NN accuracy for a variety of topological settings. Hotter colors correspond to higher accuracy. (B) A depiction of the segment-level NN after optimization of feature location within the NN topology. Lower level features such as the accelerometer signal was placed in lower levels of the network while more abstract features, such as negative text sentiment, was placed higher in the network.

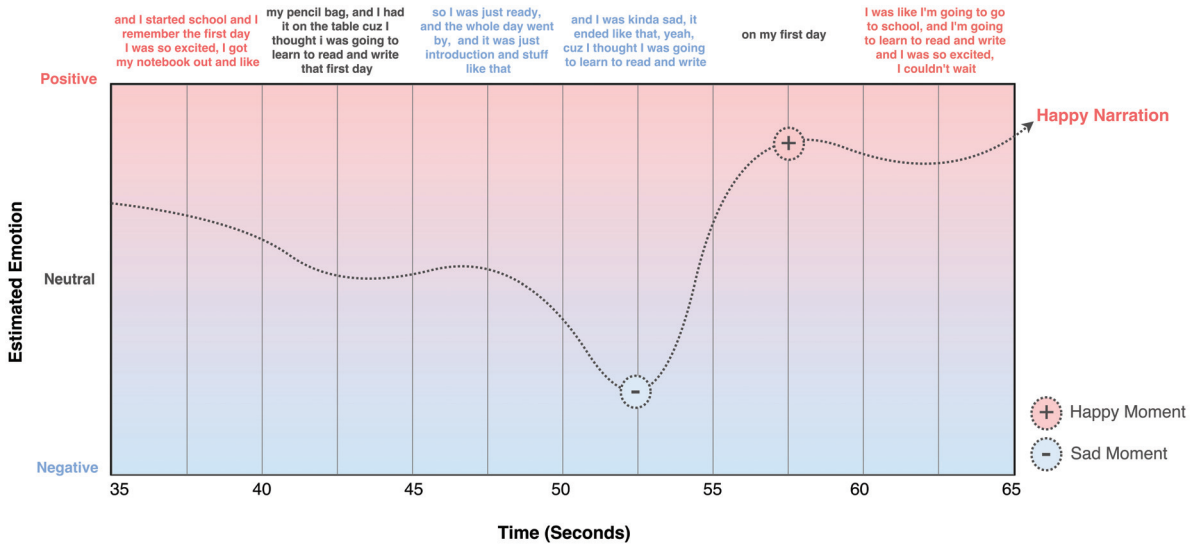

Figure 3: Real-time estimation of the emotional content in 30 seconds of collected data, using our optimized NN. The color of the text at the top of the plot reflects the ground truth labels generated by the research assistant (blue for negative, red for positive, black for neutral). The predictions of the network (y-axis) reflect the underlying emotional state of the narrator.

model AUC of 0.92 (0.02 better than the next best performing model). The selected features, model coefficients, odds ratio and p-values of the logistic regression model are illustrated in Table 3. All selected features exhibited a statistically significant association ($p < 0.001$) with the outcome.

The coefficients of the logistic regression indicate that increased movement (Accelerometery), and cardiovascular activity (PPG Signal) are associated with negative narratives. We also observed that differing postures (As reflected by the PPG sensor number) are associated with different emotional states.

The MFCC 12 feature is a measure of static background audio noise which arises when speakers fall silent. That is, when there is less speaking, it is indicative of a more negative narratives. This is also reflected in the voicing probability feature, which captures if the subject was speaking and was associated with a positive narratives.

We may interpret the MFCC 3 feature as an indication of minimum voice energy changes. That is, fluctuations in low frequency energy content are associated with happy narratives. This is reflected more generally by the RMS Energy feature, which indicates that higher energy is associated with positive narratives. Lastly, we observe that negatively charged linguistic content is associated with negative

narratives.

Segment-Level

In Table 4(B), we present the results of the best performing NN for the segment-level classification. Given that logistic regression was the best performing model in the narrative-level classification, multinomial logistic regression was selected as the baseline for this component of our analysis.

As shown in Figure 2A, the optimal topology was a two hidden-layer network with six nodes in the first layer and three nodes in the second layer resulting in an accuracy of 45.3 %, a 4.5% absolute increase over the Multinomial Logistic Regression (40.8 % accuracy). As shown in Figure 2B, the optimal connection of the features was distributed across the different layers of the network. The optimal connection of the RMS Energy (mean), Voicing Probability (linear regression MSE), Negative sentiment, PPG Sensor #1, and Accelerometer Y-axis (mean) was to the output layer. The optimal connection of the MFCC 3 Differential (min), and PPG sensor #0 features was to the second hidden layer. The optimal connection of Accelerometer Z-axis (mean), PPG Signal (mean), and MFCC12 (kurtosis) was to the first hidden layer.

The mean accuracy of the feature optimized NN, was 47.3%, exhibiting a 7.5% absolute, and 17.9% relative improvement to the baseline approach. We found that modulation of network prediction using the state transition probabilities decrease classification performance, indicating that emotional content is stable within the 5 second segments.

In Figure 3, we illustrate the performance of our optimized NN on a 30 second segment of an actual subject's data. The color of the text at the top of the plot reflects the annotations from the research assistant (blue is negative, red is positive, black is neutral), while the amplitude of the signal reflects our model's real-time emotion estimate. The figure demonstrates the ability of our model to perform real-time classification during natural conversation.

Discussion

In this work, we collected a set of non-invasive audio and physiologic signals on subjects engaged in minimally structured narrations and performed two analyses. In the first analysis, we trained an optimized NN to classify the overall nature of a narration, as reported directly by the subject: happy or sad. In the second analysis, we trained an optimized NN to classify segment-level emotional class: positive,negative or neutral.

The novelty of this work lies in both the multi-modal data collected, and the methodological approach utilized for the classification of the emotional state. Unlike prior efforts in this area (Zeng et al. 2009; El Ayadi, Kamel, and Karray 2011), participants in the study were not asked to artificially act-out particular emotive states. Instead, participants were asked to narrate a story of their choosing, that was happy or sad in their subjective opinion. To our knowledge, this type of data is unique in the existing literature. Hence, the models we developed in this study may be more representative of the emotive content one may expect to see during natural human interaction.

There were several features such as those related to ECG that were not marked for inclusion using our feature selection approach. This result is likely due to our feature extraction paradigm and not the informational content in the signals themselves.

The mean AUC of our narrative-level NN was 0.92. This result provides evidence that real-time emotional classification of natural conversation is possible with high fidelity. The convergence of our narrative-level NN to a logistic regression model also indicates that the features which discriminate between happy or sad narrations are easily interpreted and consistent across patients.

While the accuracy of our segment-level classification was 17.9% above chance, and 7.5% better than the baseline approach, there is clearly room for improvement. One reason for the relatively modest performance of the segment-level model may be the small size of our data-set, which limits the ability of any modeling approach to identify the generalized associations between our selected features and the segment-level emotions. Indeed, we take these results as strong motivation for the collection of larger datasets.

Beyond emotion detection, this work demonstrates how to enhance the performance of NNs by optimization of network topology, and the location of features within the network structure. We observed that introducing features into particular layers improved the performance of the NN. Interestingly, lower-level features such as PPG and accelerometer signals were placed lower in the network structure, while higher level features such as negative sentiment of transcribed audio were placed in a higher level of the network structure.

Additional work on the ultimate utility and usability of an emotion detection algorithm is also needed to understand the value of real-time emotion detection. Additional experiments may include discretely sending text message notifications on the status of the interaction to one of the participants, and monitoring changes in emotive trajectory associated with the intervention. When used in combination, our models could serve as a starting point for a real-time emotion system, which provides historic and real-time assessments of interactions (See Figure 4), allowing for long-term coaching and in-the-moment interventions.

References

Barrett, L. F.; Lindquist, K. A.; and Gendron, M. 2007. Language as context for the perception of emotion. *Trends in cognitive sciences* 11(8):327–332.

Barrett, L. F.; Mesquita, B.; and Gendron, M. 2011. Context in emotion perception. *Current Directions in Psychological Science* 20(5):286–290.

Calvo, R. A., and D'Mello, S. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on* 1(1):18–37.

Cameron, J. J., and Robinson, K. J. 2010. Dont you know how much i need you? consequences of miscommunication vary by self-esteem. *Social Psychological and Personality Science* 1(2):136–142.

D'Mello, S., and Kory, J. 2012. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 31–38. ACM.

Douglas-Cowie, E.; Campbell, N.; Cowie, R.; and Roach, P. 2003. Emotional speech: Towards a new generation of databases. *Speech communication* 40(1):33–60.

El Ayadi, M.; Kamel, M. S.; and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44(3):572–587.

Esuli, A., and Sebastiani, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, 417–422. Citeseer.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462. ACM.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Haag, A.; Goronzy, S.; Schaich, P.; and Williams, J. 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. In *ADS*, 36–48. Springer.

Han, K.; Yu, D.; and Tashev, I. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*, 223–227.

Kahou, S. E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, Ç.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R. C.; et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 543–550. ACM.

Katsis, C. D.; Katertsidis, N.; Ganiatsas, G.; and Fotiadis, D. I. 2008. Toward emotion recognition in car-racing drivers: A biosignal processing approach. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 38(3):502–512.

Koolagudi, S. G., and Rao, K. S. 2012. Emotion recognition from speech: a review. *International journal of speech technology* 15(2):99–117.

Li, L.; Zhao, Y.; Jiang, D.; Zhang, Y.; Wang, F.; Gonzalez, I.; Valentin, E.; and Sahli, H. 2013. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 312–317. IEEE.

Müller, E.; Schuler, A.; and Yates, G. B. 2008. Social challenges and supports from the perspective of individuals with asperger syndrome and other autism spectrum disabilities. *Autism* 12(2):173–190.

Plutchik, R. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist* 89(4):344–350.

Rilliard, A.; Shochi, T.; Martin, J.-C.; Erickson, D.; and Aubergé, V. 2009. Multimodal indices to japanese and french prosodically expressed social affects. *Language and speech* 52(2-3):223–243.

Schuller, B.; Rigoll, G.; et al. 2006. Timing levels in segment-based speech emotion recognition. In *INTERSPEECH*.

Schuller, B.; Steidl, S.; and Batliner, A. 2009. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, 312–315. Citeseer.

Stuhlsatz, A.; Meyer, C.; Eyben, F.; ZieIke, T.; Meier, G.; and Schuller, B. 2011. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 5688–5691. IEEE.

Wagner, J.; Kim, J.; and André, E. 2005. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 940–943. IEEE.

Walter, S.; Scherer, S.; Schels, M.; Glodek, M.; Hrabal, D.; Schmidt, M.; Böck, R.; Limbrecht, K.; Traue, H. C.; and Schwenker, F. 2011. Multimodal emotion classification in naturalistic user behavior. In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*. Springer. 603–611.

Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; and Narayanan, S. S. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *INTERSPEECH*, 2362–2365.

Wöllmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; and Rigoll, G. 2013. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31(2):153–163.

Zeng, Z.; Pantic, M.; Roisman, G. I.; and Huang, T. S. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31(1):39–58.