Improving predictions of polite and frustrated speech using linguistic features associated with different cognitive states in children

Cindy Chiang (cindyc@usc.edu)
Department of Psychology, 3620 South McClintock Ave.
Los Angeles, CA 90089

Jacqueline Brixey (brixey@usc.edu)
USC Institute for Creative Technologies
12015 E Waterfront Dr, Los Angeles, CA 90094

James Gibson (jjgibson@usc.edu)
Viterbi School of Engineering, 3650 McClintock Ave.
Los Angeles, CA 90089

Morteza Deghani (mdehghan@usc.edu)
Department of Psychology, 3620 South McClintock Ave.
Los Angeles, CA 90089

Abstract

Childrens poor emotional self-regulation is associated with poor mental health outcomes. This study presents methods that improve prediction rates of polite and frustrated speech using linguistic cues. These improvements can be used to help automatically identify characteristics of poor self-regulation in future studies. This work adds to previous research by considering existing computer science, psychology, and psycholinguistics methodologies and findings. More specifically, features associated with childrens cognitive control capacities across age groups are considered to investigate acoustic, semantic, and syntactic features in speech. The current analyses indicate that the features most predictive for polite and frustrated speech differ, a combination of features work best for predicting both speech types, and the predictive quality of features do not vary substantially by age. Further work should be conducted to clarify how well these findings transfer to general and clinical populations as well as to consider the developmental norms of different age groups.

Keywords: self-regulation; linguistic features; machine learning

Introduction

Approximately 13 percent of children and adolescents have been estimated to have clinically significant mental health problems that impair daily life functioning (Jellinek et al., 1999; Semansky, Koyanagi, & Vandivort-Warren, 2003). Many of these mental health problems have been linked to poor emotional self-regulation (Forbes & Dahl, 2005; Hinshaw, 2002; Kuntsche, Knibbe, Engels, & Gmel, 2007; Wyman et al., 2009) and difficulty with regulating emotion during higher levels of distress. Interventions developed to target self-regulation (Wyman et al., 2010) have been shown to be effective in decreasing rates of problematic behavior in schools, and in improving some aspects of functioning in the classroom (Wyman et al., 2010).

Despite the large number of youths affected and the efficacy of these targeted interventions, there is not an automated way to identify children with poor emotional self-regulation. To take initial steps towards developing such a method, this paper works to identify methods of improving prediction of polite and frustrated speech using linguistic features present in child speech.

Several linguistic features have been used to identify emotional states in adults, specifically prosodic and semantic features. These features were linked to various psychological and emotional states and were used to automatically categorize these states. Features akin to semantic cues have been used, and word count methods, such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015), linked emotional states to word distributions in a range of categories. Researchers participating in the annual Interspeech Challenge have also sought to determine acoustic features related to emotional speech (Schuller, Steidl, & Batliner, 2009).

Additionally, some linguistic features have been used to identify polite or frustrated speech. For example, polite speech class prediction performance improved through a fusion of acoustic, lexical, and contextual features for children's speech (Yildirim, Narayanan, & Potamianos, 2011). Boril, Sadjadi, Kleinschmidt,and Hansen (2010) used results of tasks measuring cognitive load to improve prediction rates of frustrated speech in drivers. The study improved prediction rates of frustrated speech using subjects performance on cognitive tests and the acoustic features of their speech.

However, literature in psychology indicated that there are additional linguistic features that differentiate polite and frustrated speech. These features have also been observed to change through development and cognitive load. Developmental changes in language included the comprehension and production of more complex sentences (Gaer, 1969). Cognitive changes across development included a larger capacity to overcome cognitive load difficulties (Hsu & Jaeggi, 2013). The potential impacts of these changes in the linguistic cues of both polite and frustrated speech are further detailed below.

In both an observational (Gleason, Perlmann, & Greif,

1984) and experimental study, Greif and Gleason (1980) found that children use polite speech in structured and formulaic manners. Polite speech was also found to be couched in routine, often prompted by parents, and reinforced by parents (Gleason et al., 1984). As a result, even when polite speech was deliberately elicited in children aged two to five, the frequency of polite speech was very low (Greif & Gleason, 1980). Two-year-olds in this study, thought to be too young to even understand the nuances of polite speech, still produced polite speech. The researchers of both studies attributed this phenomenon to the rote and formulaic nature of speech. The frequency in these studies illustrated the link between the directedness of polite speech and its frequency in adult speech. The authors in both studies hypothesized that these differences were also linked to socioeconomic status and parenting styles, as these types of speech were reflective of input and directions from parents, rather than developmental factors.

Developmental differences in these areas may not have been as evident, since children as young as two formulaically produced polite speech. The scripted nature of polite speech should have impacted the semantic and syntactic cues present. As such, these studies indicated that there were little variability in the types of words, contexts, and language structures that were utilized when producing polite speech. Changes in cognitive load similarly should not influence polite speech.

Linguistic cues in frustrated speech, in contrast, have been found to influence cognitive load. Boril and colleagues (Boril et al., 2010) improved rates of categorizing frustrated speech in adults when cognitive factors and acoustic cues were considered. Cognitive factors should similarly influence childrens frustrated speech.

Previous literature dealing with cognitive factors influence on speech found that poor cognitive control influenced peoples ability to accurately interpret complex sentences (MacDonald, Just, & Carpenter, 1992) and impacted lexical associations (Boudewyn, Long, & Swaab, 2012). These factors could be further modulated by changes in cognitive development, as some aspects of cognitive development continue past young childhood (Munakata, Snyder, & Chatham, 2012).

The current study looks to improve prediction processes of polite and frustrated speech by considering the existing literature in computer science and psychology. Based on the reviewed literature, several factors could improve prediction processes of polite and frustrated speech: using a subset of linguistic features, using combined linguistic characteristics, and using linguistic features known to co-occur with the cognitive load children experience while calm or frustrated. These pieces are addressed in the experimental methods outlined below.

Methods

Corpus

The Children's Interactive Multimedia Project (ChIMP) database (Narayanan & Potamianos, 2002) was utilized for this work. ChIMP is a corpus of child-machine spoken dialogues in a Wizard-of-Oz game setting. Participants played "Where in the USA is Carmen Sandiego?" and located a cartoon criminal by communicating commands to game agents. Approximately 100 subjects, both male and female, between the ages of 7 and 14 participated (Table 1). Subjects formed three age groups: 7-9 years old (young), 10-11 y/o (middle), and 12-14 y/o (old).

Table 1: Distribution of subjects and number of utterances for each emotional class (neutral, polite, frustrated) for each gender-age group.

Group	N	Neutral	Polite	Frustrated	Total
7-9 y/o	38	3966	977	796	5739
10-11y/o	35	4004	1078	360	5442
12-14 y/o	30	3005	694	705	4404
Female	48	5035	1513	800	7438
Male	55	5940	1236	1061	8237
Total	103	10975	2749	1861	15585

The recorded spontaneous utterances were manually labeled with an emotional tag: polite, neutral, or frustrated (Table 1). The corpus contained over 15,000 labeled utterances, with approximately 700 unique words. The data set showed notable variation in age and gender behaviors. The middle group was more polite and less frustrated than the other two age groups during the game. The younger and older groups were nearly twice as frustrated as the middle age group. The frustration age trend was partially driven by subjects' exacerbation with the game's level of challenge or ease. Additionally, frustrated expressions occurred more often in losing games than in winning instances. By gender, girls were more polite and less frustrated overall in their interactions during the game than boys (Arunachalam, Gould, Andersen, Byrd, & Narayanan, 2001).

ChIMP was used previously to investigate polite and frustrated speech in children. Prior work utilized latent semantic analysis (LSA) for discourse topics and explored emotional salience in lexical features to predict polite and frustrated speech (Yildirim et al., 2011). Variations of this set have been used to improve uncertainty predictions and to hone ways to improve machine coding validity (Black, Chang, & Narayanan, 2008). This work expands on prior research by exploring LIWC categories, part-of-speech (POS), and word embeddings (WEs) as features to predict cognitive mechanisms.

Extracted Features

Feature extraction was motivated by the analysis of the child-machine interaction dialogues from ChIMP. Thus, acoustic, lexical, and syntactic features are proposed.

Acoustic 384 low-level descriptors (LLD) - such as such as pitch frequency, formant frequency, root mean square (RMS) energy, and zero-crossing-rate (ZCR) - were extracted using openSMILE (Eyben, Wöllmer, & Schuller, 2009). These extraction measures build upon the work of (Yildirim et al., 2011), and were combined with new lexical and syntactic features for analyses, described below.

Lexical Two separate features measure lexical variation - LIWC and word embeddings (WEs). LIWC version 2007 was used to generate the LIWC feature set. LIWC provides information about an utterance's psychological dimension and will measure semantic word choice variation. All LIWC categories were considered, and Pearson's correlation was calculated to determine which categories have more predictive power for determining polite and frustrated speech.

WEs were mappings of words in the vocabulary of the data set to vectors of real numbers. This feature captured meaning, semantic relationships, and context for words in ChIMP. Thus, vocabulary in frustrated utterances were represented in a feature space that were separate from polite utterances.

Syntactic Part of speech (POS) tags were generated by the Stanford Part of Speech Tagger (https://nlp.stanford.edu/software/tagger.shtml). We hypothesized that variation in POS should occur as a product of cognitive control. Hence, decreased complexity, represented as shallower trees, will align with the childs level of frustration.

Analysis

To determine how cognitive and developmental measures correspond to politeness and frustration expression by children, we conducted three analyses.

Analysis 1 We conducted five-fold cross validation experiments to compare feature sets. Two machine learning techniques were used: feed forward neural network (FFNN) for utterance level feature sets (acoustic and LIWC features); long short term memory (LSTM) for sequence based features (WEs and syntactic features). As a previous study has previously explored a Bayesian classifier to predict polite and frustrated speech (Yildirim et al., 2011), we expand on those findings by implementing two new state-of-the-art models in our study. Five speaker-independent cross validation folds, approximately balanced across age and gender, were created.

Each neural network was implemented and trained using Keras (Chollet et al., 2015) with Theano (Theano Development Team, 2016) as the back-end. All the systems were trained using categorical cross-entropy loss and optimized using the Adam algorithm (Kingma & Ba, 2014). The loss function was weighted by class according to the inverse frequency of each class (to account for class imbalance). Within each fold the utterances of approximately 10% of the speakers were separated as a validation set. They were trained for a maximum of 30 epochs with an early stopping strategy to terminate if the validation loss did not decrease after three consecutive epochs, and only the model with the lowest validation loss was retained. Each model was trained in 10 trials using different random initializations and the reported results were averaged across these trials.

The feed forward networks consisted of two hidden layers where the first layer was of equal dimension to the input feature vector and the second hidden layer dimension was 10% of the first layers'. Both hidden layers had sigmoid activation. The LSTM networks consisted of an embedding layer at the input followed by a bidirectional layer of dimension 50 for the word embedding features and 26 for the POS sequences (the number of unique POS tags). All the networks had a softmax activation at the output layer.

Analysis 2 We used decision level fusion to combine the feature sets to compare the power of multimodality in predicting emotional states. We used the average fusion algorithm (Yildirim et al., 2011) to combine the computed posterior probabilities of each single feature set classifier in order to estimate the posterior probability of a combined classifier. We hypothesized that the fusion of acoustic, lexical, and syntactic would provide more predictive power than single-feature models alone.

Analysis 3 We executed training on one age group and testing as well as experiments on the others for age specific performance on politeness and frustration. We expected that the expression of politeness was age independent, whereas frustration was age dependent.

Results

We presented results to detect frustrated and polite attitudes in childrens speech using the selected features for three-way classification tasks: single feature evaluation, fused features evaluation, and age specific performance.

Analysis 1

The results for this analysis are shown in the solid bars to the left in Figure 1. For single-feature classifications, WEs provided the best predictive power, while acoustic features displayed the worst performance. To understand how well each feature predicted the three emotional classes, F1 scores were determined for each emotion class (Figure 2). It is clear that acoustic features performed poorly overall due to the poor predictive power for neutral and polite classes, despite being the best at classifying frustrated utterances. Both LIWC and POS exhibited good predictive power for neutral and polite speech, but performed poorly for frustrated. The models trained on WEs features were the most successful, but performed the best at correctly classifying polite utterances. Overall, the features tended to be more predictive of polite speech.

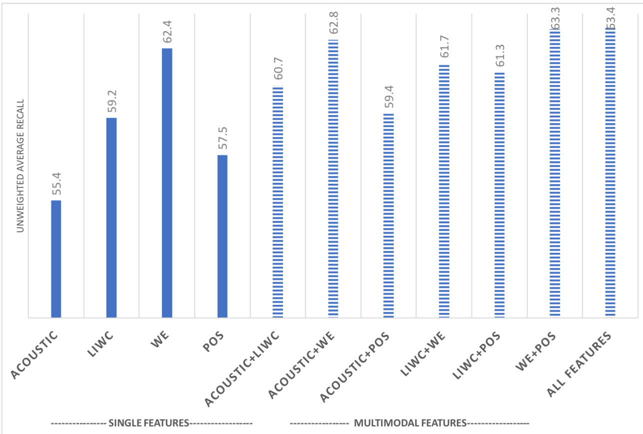

Figure 1: Results from Experiments 1 and 2. Word embeddings (WE) performed the best as a single feature (experiment 1) while the fusion of all features produced the highest prediction rate overall (experiment 2).

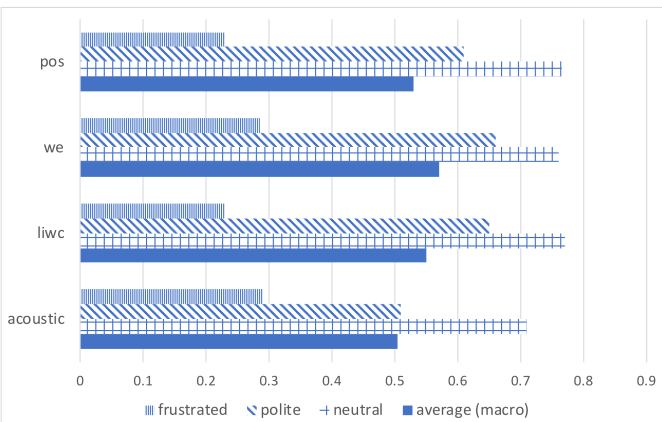

Figure 2: F1 scores for emotional states for each features from Experiment 1.

Analysis 2

The results for fusing the selected features are shown in the striped bars to the right in Figure 1. The majority of the fused features were more successful at predicting than the single-feature classification, with the combination of all the features performing the best (Unweighted average recall of 63.4%) of all the models. Both the WEs+POS and Acoustic+LIWC+WEs+POS fused feature models showed a significant improvement with respect to the best single system result in Analysis 1 ($p<0.01$).

Analysis 3

We conducted experiments where each age-group was used for training and the system was tested on the remaining age groups to determine age specific prediction for each feature, summarized in Table 2. In general, systems trained on the middle group were the least successful at predicting frustration, but the inverse trend at predicting politeness. The results indicated that systems trained on WEs for the middle group produce the best scores for politeness, while systems trained on either the acoustic or WEs features for the young or old groups are able to predict frustration the best. This seemed to be reflective of the large difference in the politeness and frustration distributions for the middle group versus the other two groups rather than reflecting age-related differences.

Table 2: Leave-one-out train versus test F1 score by age group for each feature for politeness and frustration emotional states.

Emotion	Age group	Young	Middle	Old
	Acoustic			
Politeness	Young	-	51.2	37.6
	Middle	51.0	-	37.3
	Old	36.5	40.2	-
Frustration	Young	-	19.0	34.2
	Middle	31.0	-	32.1
	Old	32.6	16.3	-
	LIWC			
Politeness	Young	-	66.8	61.2
	Middle	64.2	-	57.4
	Old	61.1	70.0	-
Frustration	Young	-	11.6	31.5
	Middle	28.7	-	28.6
	Old	28.3	14.9	-
	WE			
Politeness	Young	-	75.8	62.7
	Middle	66.6	-	59.9
	Old	65.6	71.5	-
Frustration	Young	-	15.7	32.9
	Middle	30.0	-	29.1
	Old	34.6	15.5	-
	Part of Speech			
Politeness	Young	-	56.4	52.2
	Middle	60.8	-	52.9
	Old	53.3	53.0	-
Frustration	Young	-	9.9	32.0
	Middle	10.6	-	7.3
	Old	29.7	12.3	-

Feature analysis

Finally, we analyzed acoustic and LIWC features. First, we calculated Pearson correlations to determine the top features with predictive power with respect to politeness and frustration (Table 3).

The top LLD values for the acoustic feature set showed positive correlations with frustration, and negative correlations with politeness. The features positively correlated with frustration deal with signal frame energy, while the features most correlated with politeness relate to zero-crossing rate of time signal.

Table 3: Pearson's correlation (rho) for acoustic and LIWC top features.

	Feature	rho
Acoustic politeness	zcr_linregerrQ	-0.204
	zcr_stddev	-0.199
	RMSenergy_de_minPos	-0.198
Acoustic frustration	RMSenergy_amean	0.165
	RMSenergy_stddev	0.157
	RMSenergy_range	0.147
LIWC politeness	you	0.432
	posemo	0.402
	affect	0.393
LIWC frustration	inhib	0.123
	i	0.118
	verb	-0.078

For LIWC features, politeness showed negative correlation with affective processes (e.g. affect). The correlations illustrated that the LIWC category inhibition (e.g. inhib) was positively correlated with frustration.

Correlations for age groups were next reviewed for both features, which can be seen in Table 4. For acoustic features, the top acoustic features were consistent across age for frustration but not politeness. So, frustration was more consistent in its acoustic expression across age groups whereas politeness was not. The top LIWC features were consistent across age for politeness but not frustration. So, expression of politeness was more uniform across age with respect to language whereas expression of frustration through language varies more across age groups.

Table 4: Correlations for utterance level features with politeness and frustration for each age group.

	Politeness
Acoustic	
Young	RMSenergy_de_minPos (-0.212),
Middle	zcr_stddev (-0.222),
Old	fftMag_mfcc[1]_linregerrQ (-0.203)
LIWC	
Young	you (0.478),
Middle	you (0.462),
Old	you (0.328),
	Frustration
Acoustic	
Young	RMSenergy_amean (0.200),
Middle	RMSenergy_amean (0.113),
Old	RMSenergy_amean (0.178),
LIWC	
Young	inhib (0.202),
Middle	social (0.097),
Old	i (0.149),

Discussion

The experiments in this work produced several interesting findings. First, WEs was the best predictor when looking at the single-feature systems. Second, the best single feature predictor differed across types of speech, as LIWC and WEs were more successful when predicting polite speech, while acoustic features were more successful with predicting frustrated speech. Third, training the models by age group yielded differing levels of success. Fourth, several features correlated well with polite or frustrated speech. Overall, it appeared that several factors influenced predictiveness the most, mainly semantic features and age groups.

The overall contribution of the WEs could be attributed to its success in predicting polite speech and the number of polite speech in the corpus, as these represented double the number of frustrated utterances. It is possible that WEs and LIWC categories were the most successful as a result of the formulaic nature of polite speech (e.g. "thank you"). The result that "you" was the LIWC category most correlated with polite speech supported this possibility. Additionally, there were negative correlations with cognitive mechanism words, which could be attributed to the scripted rather than engaged and thoughtful speech.

In contrast, acoustic cues were most predictive of frustrated speech. It may be that semantic cues were not reliable and that there was a lot of variability within the words used. Previous research found that poor cognitive control was associated with different levels of sensitivity to lexical associations (Boudewyn et al., 2012). Participants who performed poorly on cognitive tests, particularly those with difficulty on the suppression tasks, in the study were more sensitive to lexical associations. It was possible that sensitivity to lexical associations produced speech that was less characteristic with some of the categories within the LIWC dictionary. An alternative hypothesis is that frustrated speech was more variable, irrespective of the lexical associations that children might make when frustrated. Nozari, Freund, Breining, Rapp, and Gordon (2016) described the use of cognitive control on different stages of language production, one of which required monitoring and revising word choice errors.

Across age groups and speech categories, there were some differences in the predictive strength for certain features. In polite speech, LIWC and WEs were very good in training classification models across time groups. These features did not change across time. It may be the case that polite speech in this corpus did vary across age groups. Studies by Gleason, Perlmann, and Greif (1984) and Greif and Gleason (1980) found this trend in their studies and attributed it to the scripted nature and acquisition of the speech.

In frustrated speech, the F-1 scores of the middle age group in all linguistic feature categories were lower than the other two age groups. This might result from the smaller number of frustrated utterances in this age group.

While these findings conform to the trends reported by previous literature, the age groups investigated in the current study was different. Previous studies have generally investigated children in a younger age group (e.g. two to five years old). There may be additional features that have not been captured by the literature and factors that were not considered by the classification experiments conducted in this study. It will be important to further consider the developmental norms of older age groups, especially if such classification models are used in clinical and practical settings.

References

Arunachalam, S., Gould, D., Andersen, E., Byrd, D., & Narayanan, S. (2001). Politeness and frustration language in child-machine interactions. In *Seventh european conference on speech communication and technology.*

Black, M., Chang, J., & Narayanan, S. (2008). An empirical analysis of user uncertainty in problem-solving child-machine interactions. In *First workshop on child, computer and interaction.*

Boril, H., Omid Sadjadi, S., Kleinschmidt, T., & Hansen, J. H. (2010). Analysis and detection of cognitive load and frustration in drivers' speech. *Proceedings of INTERSPEECH 2010*, 502–505.

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2012). Cognitive control influences the use of meaning relations during spoken sentence comprehension. *Neuropsychologia*, *50*(11), 2659–2668.

Chollet, F., et al. (2015). *Keras.* `https://github.com/fchollet/keras`. GitHub.

Eyben, F., Wöllmer, M., & Schuller, B. (2009). Openearintroducing the munich open-source emotion and affect recognition toolkit. In *Affective computing and intelligent interaction and workshops, 2009. acii 2009. 3rd international conference on* (pp. 1–6).

Forbes, E. E., & Dahl, R. E. (2005). Neural systems of positive affect: relevance to understanding child and adolescent depression? *Development and psychopathology*, *17*(3), 827–850.

Gaer, E. P. (1969). Children's understanding and production of sentences. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 289–294.

Gleason, J. B., Perlmann, R. Y., & Greif, E. B. (1984). What's the magic word: Learning language through politeness routines. *Discourse Processes*, *7*(4), 493–502.

Greif, E. B., & Gleason, J. B. (1980). Hi, thanks, and goodbye: More routine information. *Language in Society*, *9*(2), 159–166.

Hinshaw, S. P. (2002). Process, mechanism, and explanation related to externalizing behavior in developmental psychopathology. *Journal of Abnormal Child Psychology*, *30*(5), 431–446.

Hsu, N. S., & Jaeggi, S. M. (2013). The emergence of cognitive control abilities in childhood. In *The neurobiology of childhood* (pp. 149–166). Springer.

Jellinek, M. S., Murphy, J. M., Little, M., Pagano, M. E., Comer, D. M., & Kelleher, K. J. (1999). Use of the pediatric symptom checklist to screen for psychosocial problems in pediatric primary care: a national feasibility study. *Archives of Pediatrics & Adolescent Medicine*, *153*(3), 254–260.

Kingma, D. P., & Ba, J. (2014, December). Adam: A Method for Stochastic Optimization. *ArXiv e-prints*.

Kuntsche, E., Knibbe, R., Engels, R., & Gmel, G. (2007). Drinking motives as mediators of the link between alcohol expectancies and alcohol use among adolescents. *Journal of Studies on Alcohol and Drugs*, *68*(1), 76–85.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive psychology*, *24*(1), 56–98.

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control: Three key transitions. *Current directions in psychological science*, *21*(2), 71–77.

Narayanan, S., & Potamianos, A. (2002). Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, *10*(2), 65–78.

Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, cognition and neuroscience*, *31*(7), 886–903.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Tech. Rep.).

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge.

Semansky, R. M., Koyanagi, C., & Vandivort-Warren, R. (2003). Behavioral health screening policies in medicaid programs nationwide. *Psychiatric Services*, *54*(5), 736–739.

Theano Development Team. (2016, May). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, *abs/1605.02688*. Retrieved from `http://arxiv.org/abs/1605.02688`

Wyman, P. A., Cross, W., Brown, C. H., Yu, Q., Tu, X., & Eberly, S. (2010). Intervention to strengthen emotional self-regulation in children with emerging mental health problems: Proximal impact on school behavior. *Journal of abnormal child psychology*, *38*(5), 707–720.

Wyman, P. A., Gaudieri, P. A., Schmeelk-Cone, K., Cross, W., Brown, C. H., Sworts, L., ... Nathan, J. (2009). Emotional triggers and psychopathology associated with suicidal ideation in urban children with elevated aggressive-disruptive behavior. *Journal of abnormal child psychology*, *37*(7), 917–928.

Yildirim, S., Narayanan, S., & Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language*, *25*(1), 29–44.