Automatic Emotion and Behavior Recognition in Parent-Child Interaction

Bernd Huber, *Harvard Intelligent Interactive Systems* and David Ramsay, *MIT Media Lab*

Abstract—There are life-long ramifications for the 35% of children entering kindergarten already behind in language skills and social-emotional development. Research has shown that differences in the quality and style of parent-child interaction are the leading cause of this disparity. Unfortunately, few technical solutions exist to reduce and democratize the work of highly trained early intervention specialists. In this paper, we present a model to automatically classify parent-child interactions for both behavioral and emotional content using paralinguistic and linguistic features from speech. We examine which features are the most predictive, how emotion and behavior relate in these interactions, and introduce a framework for implementing a fully-automated system for real-world scenarios. We show 80% and 45% accuracy classifying behavior and emotion labels respectively, with some noise sensitivity. This model is a fundamental step towards an automated solution for empowering parents to expedite their child's social-emotional and language development. Our system enables a variety of powerful applications for parents, researchers, and practitioners.

Index Terms—PCI, PCIT, Speech Emotion Recognition, Behavioral Coding.

1 Introduction

School readiness is imperative to the long-term success of American children. [1], [2] At the age of five, school readiness predicts future success in elementary and high school, future earning potential, and mid-career social class. Unfortunately, 35% of children entering kindergarten are not school-ready; moreover, a significant disparity between the readiness of low- and high-income children is well documented. [1]

This issue first became prominent with Hart and Risley's 1995 paper revealing a 30 million-word gap by age 3 between low and high-income families. [3] In their paper, they demonstrated the word gap corresponds with a reduction in the child's IQ and vocabulary skills. Recent work has shown that major cognitive and language deficits at this age are not as related to the quantity of words spoken as to the quality of parent-child interactions in the home. [4] Practices such as conversational turn taking, 'parentese' inflection, expressiveness, engagement, and dialogic techniques (which include extension, repetition, completion prompts, and distancing) all correlate strongly with positive learning outcomes. [5], [6], [7]

The quality of parent-child interaction is also crucial for social-emotional development. Up to 14% of children under 5 have behavioral problems that impact their school performance. [8] Parenting style, as well as parental positive affect, warmth, and expressivity have been shown to mitigate the likelihood of social-emotional issues. [9], [10], [11], [12], [13]

There is an increasing body of research suggesting that language learning and social-emotional development are fundamentally linked, and must be studied together. [11], [14], [15], [16] Expressive language is a key tool for emotion regulation, and the use of language that identifies an internal state corresponds to both language ability and emotional maturity. Strategies that work well for language outcomes in a parent-child dyad are designed with similar scaffolding and engagement techniques as those for social-emotional competence. [17], [18], [19]

While it is clear that there are life-long ramifications for falling behind before the age of 5, and that the skill of the parent in a parent-child interaction is the strongest driver of success, there are no publicly available options to provide parents feedback. In general, problems are recognized only *after* a significant language deficit or emotional issue has become apparent in a school setting, and a trained expert is brought in to train the parent using some form of Parent Child Interaction Therapy (PCIT). [20] $5 billion in government funding has recently been funneled into these high-touch, early intervention programs nationwide. [2]

In this paper, we analyze the feasibility of a machine-learning system to automatically assess the behavioral and emotional aspects of parent-child interactions from speech recordings. Successful classification could provide invaluable feedback to parents about their parenting style and their child's development, empowering them to adopt the optimal techniques for positive outcomes. This technology could also have powerful applications as a research tool to assess the confluence of factors that influence early childhood learning.

2 Related Work

2.1 Machine Learning

Support Vector Machines (SVM) are linear discriminant classifiers frequently used in speech classification tasks. Support Vectors represent outlines in the feature space and make the model fit better to the training data. The tolerance of the optimization algorithm is controlled by the training complexity parameter C. For larger C, the algorithm tends

• *The authors are part of Dr. Roz Picard's Affective Computing course. e-mail: dramsay@mit.edu.*

Manuscript received April 19, 2015; revised September 17, 2015.

to generate more support vectors, which leads to overfitting. In the worse case, all training samples are represented by a support vector. However, because of their generalization properties, SVM are frequently used in speech and speech emotion classification.

Hidden Markov Models (HMM) are generative classifiers frequently used in speech recognition and discourse modeling [21]. This type of model will produce strong results for transition probabilities between states, and will handle uneven class distributions well. Unfortunately, HMM do typically not include a mechanism for unsupervised integration of long-term dependencies.

A nonlinear discriminative classifier often used when large numbers of training samples are available is based on neural networks (NN). The non-linear mapping capability, however, raises problems when applied on smaller sample-numbers, in which case overfitting may occur. In this work we use Long-Short Term Recurrent Neural Networks (LSTM-RNN) to model longer term dependencies of parent child interaction, like parent temperament or mood. A detailed description of LSTM can be found for example in [22].

2.2 Parent Child Interaction

There are many schools of thought around social-emotional and language development which inform the techniques that specialists use as they engage with delayed children. However, a few systematic approaches to therapy exist that warrant mention.

Parent Child Interaction Therapy (PCIT) is a system developed by the creator of the DPICS behavioral coding system we used in this study, which will be described in a later section of this paper. PCIT is informed heavily by initial observation and coding of the behaviors we are training our system to recognize. [20] It is a behavioral parent training program, in which parents are coached through their play interactions with children 2-7 years old. It has been shown to be effective, cost-effective, and generalizable to the home. It is practiced widely around the world. [23], [24], [25]

One of the other important initiatives in this space is the LENA project. LENA is a hardware recorder that monitors parent-child interactions, counting words, conversational turns, and monitoring the environment. Right now LENA is running small pilot studies in Rhode Island, where they are focused exclusively on literacy development. They currently do not offer insights or advice with their data, but supplement trained behavior specialists with the data who consult with the enrolled families. LENA represents the most sophisticated technical presence in this space today. [6]

Despite the large amount of research in this area, there are relatively few studies that attempt to model parent-child dyads. In [26], a state space model is presented for parent-child interactions in order to assess emotional flexibility. In [27], a structural equation model is used to show a mother's emotion talk is predictive of a child's emotion talk, with up to four layers of predictive relevance. These studies informed our thinking about the role of time-dependence in our model.

3 Speech Dataset

In order to create a parent-child audio analysis system, an appropriate speech corpus is needed. A thorough search revealed two free options for child speech– the CMU CHILDES database, and Northwestern's OSCAAR database, which includes the kidLUCID dataset. [28], [29], [30] No options were available with preexisting behavioral and emotional labeled data. After extensive auditioning, we decided to use recordings from the Gleason 1988 experiment which are part of the CHILDES database. [31]

The Gleason audio samples include 24 children, 12 girls and 12 boys, ages 2 to 5. Each child is recorded twice– once interacting with the mother, and once with the father– in 50 minute play sessions in which they attempt three activities (a building task, a reading task, and a shopping play task). These varied tasks provided a variety of emotion and interaction style, as well as a representative variety of real world noise corruption.

Of these 48 recorded interactions, only fourteen included timestamped/easily separable utterances (and only one of those featured the mother). For this study, easily separable 3 hour-long father play sessions were used (Andy, Eddie, and David from the database). Each child was male and three years old, and each recording was annotated with a full transcript and morphosyntactic coding which we used in our analysis.

3.1 Coding Scheme: Behavior

The Dyadic Parent-Child Interaction Coding Scheme (DPICS) is an industry standard that has been in wide use since 1981. It has gone through three revisions, and has been validated in hundreds of publications as a meaningful, predictive tool for analysis of behavior and a predictor of social-emotional development in parent-child interaction. [32], [33], [34], [35]

The 2004 DPICS-II coding scheme is a 100-page document with complex coding rules for the numerous categories. A handful of physical codes were dropped from our audio-only project, leaving us with 21 labels. Nine of the labels are unique to parents and three are unique to children, as shown in the coding scheme below.

Parent-Only Labels	**Label #**
Direct Command	1
Indirect Command	2
Labeled Praise	3
Unlabeled Praise	4
Information Question	5
Descriptive/Reflective Question	6
Reflective Statement	7
Behavioral Description	8
Neutral Talk	9
Parent and Child Labels	**Label #**
Negative Talk	10
Playtalk	11
Answer	12
No Answer	13
Comply	14
No Comply	15
Yell	16
Whine	17
Laugh	18
Child-Only Labels	**Label #**
Command	19
Question	20
Prosocial Talk	21

Fig. 1: DPICS-II Labels, Adapted for Audio-Only

3.2 Coding Scheme: Emotion

A 25 label system for emotion classification as shown in [36] was used to label each utterance with discrete emotion.

3.3 Labeling

Both of the authors independently applied the DPICS standard to label the same conversation data from CHILDES. After labeling 2000 utterances, the results were compared. A label of zero implied either ambiguity or a mismatch between the audio and the transcript. There was substantial disagreement between the coders; in particular, there was one systematic coding difference between Neutral Talk (label 9) and either Prosocial Talk (label 21) or Play Talk (label 11). After consulting the manual in more detail, these labels were re-evaluated. All other disagreements were re-coded as label 0 and thrown out of further analysis.

Altogether, slightly over 25% of the labeled data was unused due to classification disagreement. The distribution of data is highly non-uniform as shown in Figure 2, which presents a particular challenge for robust detection and training for less frequent classes.

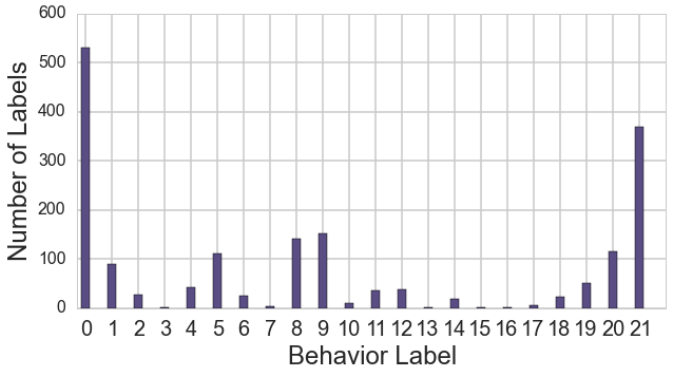

Fig. 2: Final DPICS Behavior Label Distribution.

For emotion data, the coding scheme is much simpler and less confusing. Currently, no cross-validation of emotion label agreement has been done, but in future revisions the authors will include some justification of coding agreement for emotion labeling as well. The coding distribution can be seen in Figure 3, with radius corresponding to frequency.

4 System

After finalizing our source audio, separating it into utterances, and coding it with both emotion and behavior labels, the next step was to create relevant paralinguistic and linguistic feature vectors to use as inputs into our machine learning models. With a cross-correlation evaluation metric, we could then rank and select the most relevant features for each model we constructed: the parent-emotion model, the parent-behavior model, the child-emotion model, and the child-behavior model. For each of these models, we trained and evaluated three variants: (1) stateless SVM, (2) memoryless but stateful HMM, and (3) stateful and long-memory capable deep learning LSTM RNNs. Finally, we built additional SVM, HMM, and LSTM RNN variants to analyze the relationship between behavior and emotion in the parent-child dyad.

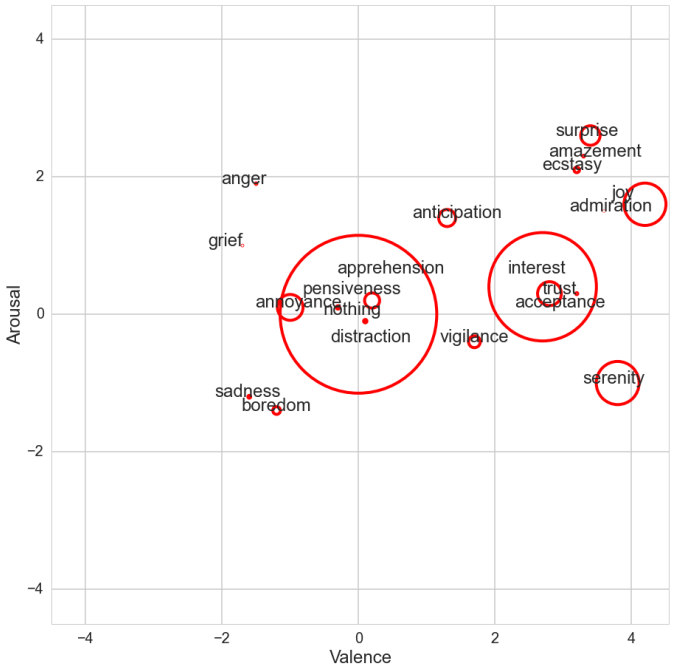

Fig. 3: Label Emotion Label Distribution.

4.1 Linguistic Features

The set of 1,918 linguistic features tested in this analysis can be divided into four main categories: vocabulary, syntax, sentiment, and repetition. Similar techniques were used for both vocabulary and syntax features. Firstly, common words and parts of speech were assigned weights for each class, based on their likelihood to appear within it. To compute a feature, each utterance received 22 best guess probabilities –one for each potential label– which were simply the sum of the probabilities that the utterance's individual words and parts of speech belonged with that label.

Normalization for these probabilities was computed using a variety of techniques– (1) the probability of a word or part of speech to belong to a given class versus others (# occurrences in this class / # occurrences in total), (2) the probability of the word to appear in a given utterance of this class (# occurrences in this class / # utterances of this class), (3) the probability that the word belongs to an important section of the class vocabulary (# occurrences in this class / # words in this class), and finally (4) the probably of being in the current class vocabulary weighted by the likelihood of class distribution. These are all valid ways of thinking about probabilistic weighting of a given word or part of speech, so instead of making a best guess, we decided to test all of them and reduce the features down at a later stage.

Beyond individual words and parts of speech, common bi-grams and tri-grams (two and three word phrases occurring together) were also used in the context of their relative co-occurrence with different class labels. In these cases, the beginning of the sentence and the end of the sentence were also included as 'words'. Each utterance was again given a likelihood score of coming from any individual class by combining the probabilities of all the common ngrams in that phrase. In this way, common syntactic structures and

common phrases could be found.

Additionally, appearances of extremely common or class-unique words, bi-grams, and tri-grams were treated each as binary features.

While these probabilities were computed over the three conversations that we labeled, we also generated lists of common elements from all 48 transcripts in the CHILDES database. We compared Fathers and Mothers vocabulary, syntax, and ngrams separately. Based on this analysis, we concluded that Fathers and Mothers use a nearly identical distributions of words and syntax. We also confirmed that our feature set was representative of the entire population, and therefore generalizable.

Beyond analyzing vocabulary and syntax, sentiwordnet was used to compute a positive and negative sentiment score for each phrase. Sentiwordnet is an open library organized into 'synsets,' or groups of synonyms. Each group has a positive or negative score associated with it. The AFINN database was used to create an additional, independent sentiment feature. This database ranks individual words on a +-5 point valence scale. Both AFINN and sentiwordnet are common tools for opinion mining. [37]

Sentiment scores were computed with no normalization per utterance or with a SQRT(length) normalization. Additional sentiment-based features included the previous utterance's score, the sentiment of the previous two or three utterances, and the cumulative sentiment of the entire conversation until this point. These curated features attempt to introduce 'mood' and emotional parent-child synchrony into the feature vector.

Finally, repetition features were introduced. Repetition was computed using repeated words and parts of speech between the current and previous utterance, both occuring at random as well as in the proper order (which is particularly important for syntax). Ordered repetition did not penalize additional intervening words in the original order– this preserves instances of parental extension (i.e. child: "I have a car!" parent: "You have a big, blue car!"). Normalization was computed using the length of the shorter phrase.

Repetition was also calculated against the penultimate and antepenultimate utterances, to allow for normal interjections that may separate reflective statements in conversation. Sentiment repetition/synchrony (i.e. two positively scored or negatively scored utterances in a row) was also captured.

4.2 Paralinguistic features

The Geneva Acoustic Parameter Set (GeMAPS) featureset contains a total number of 88 parameters. This featureset combines promising features from large brute-force featuresets and hand-crafted, psychologically motivated features. It contains features related to frequency, energy and spectrum. The set also contains cepstral parameters and dynamic parameters (delta regression coefficients and difference features, slopes of rising and falling F0 and loudness segments encapsulate some dynamic information). Furthermore, functionals are applied to the low-level descriptors (LLD). The paralinguistic features were extracted using *openSMILE* [38], an audio analysis toolbox frequently used in speech emotion analysis.

Frequency related LLD	**Group**
Pitch: log. F_0	prosodic
Jitter: F_0 = period length deviations	voic.q.
Formants 1,2,3: Centre frequency	spectral
Formant 1: Bandwidth	spectral
Energy related LLD	**Group**
Shimmer (local)	voic.q.
Sum of auditory spectrum (loudness)	prosodic
Harmonics-to-Noise Ratio(HNR)	voic.q.
Spectral related LLD	**Group**
Alpha ratio (at 50-1000Hz,1-5kHz)	spectral
Hammarberg index	spectral
Spectral slope at 0-500,500-1500Hz	spectral
Formants 1,2,3: Energy ratio, ratio energy center frequency to F_0 energy	spectral
Harmonic difference H1,H2: Ratio energy of first and second F_0 harmonic	spectral
Harmonic difference H1,A3: Ratio energy of first F_0 harmonic and highest formant 3 harmonic	spectral
Functionals applied to all LLD	**Group**
Arithmetic mean, normalized std.dev.	moments
Functionals applied to loudness,pitch	**Group**
Percentile 20/50/80th value,range	percentiles
Rate of loudness peaks	temporal
Mean length/std.dev. of $F_0 > 0$ and $F_0 = 0$	temporal
Continuous $F_0 > 0$ rate	temporal
Mean / std.dev. of rising / falling slopes	peaks
Functionals applied to alpha ratio, Hammarberg index,spectral slopes	**Group**
Arithmetic mean over unvoiced segments	moments

Fig. 4: Features in minimal *GeMAPS* featureset, derived from LLD and functionals.

4.3 SVM/HMM/LSTM-RNN

Initially, we created four models, looking at each pair of child or parent with emotion or behavior. For each of these four models, we trained and evaluated three variants: (1) stateless SVM, (2) memoryless but stateful HMM, and (3) stateful and long-memory capable deep learning LSTM RNNs. We used Weka 3 [39] for SVM analysis, and Python for building the HMMs and LSTM RNNs.

We built three variants to evaluate different models of time-dependent state information. In particular, we hypothesize that temperament, mood, and past behavior are very relevant to the accurate prediction of current behavior and emotion. We also hypothesize that immediate interaction state information is highly relevant– for instance, the previous statement (i.e. a question) and mood (i.e. happy) of the parent is likely to prompt the child response (i.e. a happy answer, assuming past behavior indicates a healthy, emotionally synchronous relationship).

SVMs are static, and thus any state information must be designed into the feature vector. Generally, SVMs are not used for adaptive, time-variant systems like this one. HMMs provide a better alternative, and can estimate state in an unsupervised way, however they are not optimized to handle long-term dependencies as well as short term ones. LSTM-RNNs provide the framework for accurately counterbalancing long-term and short-term dependencies in an unsupervised way.

On the other hand, more unsupervised complexity requires more data to train accurately. This can be a serious limitation for advanced models like the LSTM-RNN, which should be trained using tens of thousands of interactions to begin recognizing latent connections between long-term behavior patterns.

After building and evaluating these three techniques for predicting emotion and behavior, we used the same machine

learning models to examine basic relationships between the two. Instead of our normal paralinguistic and linguistic input vector, we used previous emotion, previous behavior, and current emotion to predict current behavior and vice versa.

5 Results

In this section we present a preliminary evaluation of the system. All of the following analysis was done using 'leave one conversation out' cross-validation– so one father's behavior is predicted based on the other two.

5.1 Feature selection

5.1.1 Pearson Product Moment Correlation (PPMC)

With this cross correlation evaluation metric, relevant features and their linear relationship between the classes [40] are captured. Given a number of n pairs of samples $(x_1, y_1), ..., (x_n, y_n)$, the PPMC r_{xy} can be calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \quad (1)$$

We used this metric as a measure for ranking and selecting a subset of our previously defined generic set of linguistic and paralinguistic features.

5.1.2 Feature Analysis

Using the PPMC, each of the four models was evaluated to select and understand the most relevant features. Each feature was grouped into a larger category– paralinguistic, vocabulary ngrams, syntax ngrams, repetition, and sentiment. The charts below show how many of the top 100 most predictive features fell into each category.

Additionally, we corrupted the transcript to simulate real world conditions, where no canonical transcript is available and speech-to-text is likely to be unreliable. In each graph, increasing rates of transcript corruption are introduced by removing words and syntax information from each utterance.

General trends show that linguistic features are more predictive of behavior, and paralinguistic features are more predictive of emotion. Parents and children are similar, with the caveat that linguistics matter more in the father emotion model. This suggests adults use more emotionally descriptive language.

Sentiment analysis from the transcript appeared to be an ineffective feature. Repetition, on the other hand, proved to be very valuable predictor for the few classes in which it matters (i.e. Reflective Statements and Questions).

5.2 SVM, HMM, and RNN Performance

Figure 9 shows the performance of our three child models for both behavior and emotion, while Figure 10 shows the parent model. For emotion, performance is relatively stable with increasing corruption of the text transcript (along the x axis) because paralinguistic features dominate the emotion model accuracy. For behavior, linguistic features matter more, and thus performance degrades roughly linearly with increasing corruption. LSTM-RNNs were chronically undertrained, and all versions performed with a 1% average recognition rate. For behavior the HMM was the best with an average recognition rate around 80%, followed very closely by the SVM approach. For emotion, the SVM gave the best performance, with a 50% average recognition rate for the parent, and a 40% rate for the child. The HMM version performed markedly worse at 20 and 2% recognition rates, respectively.

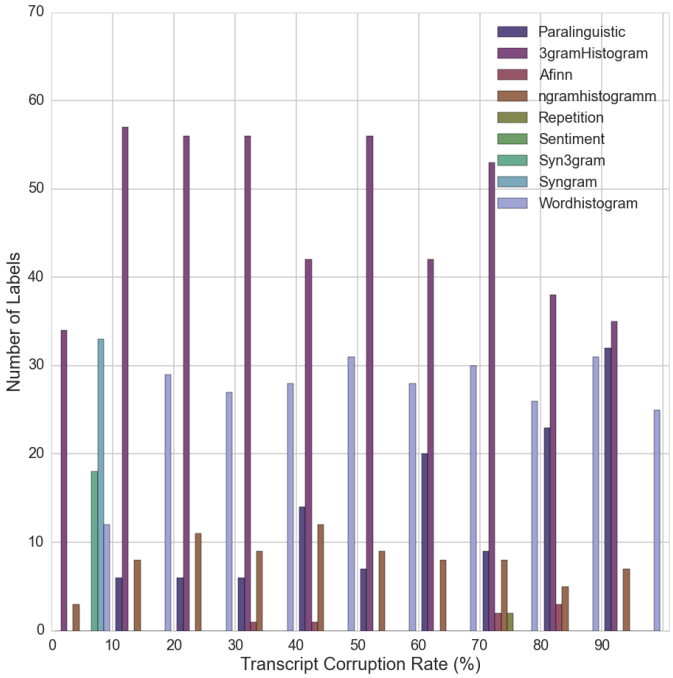

Fig. 5: Categories of the 100 Most Predictive Features for the Child Behavior Model.

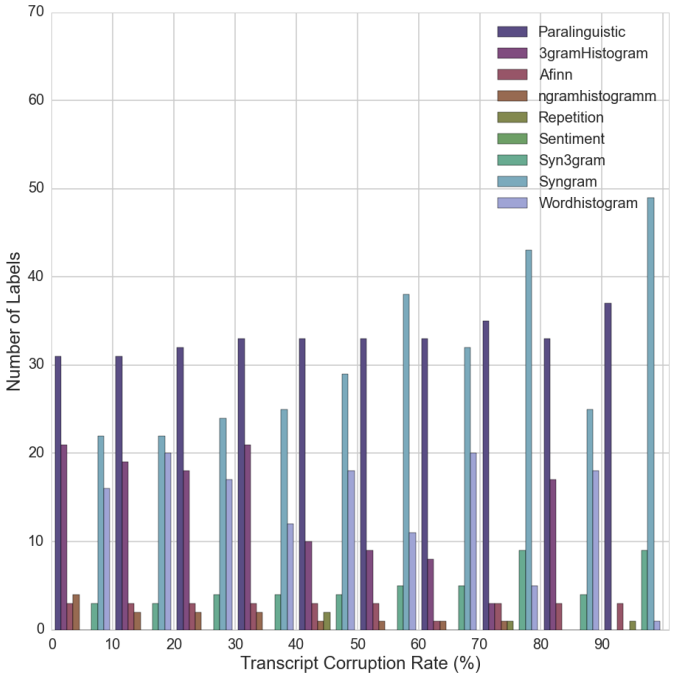

Fig. 6: Categories of the 100 Most Predictive Features for the Child Emotion Model.

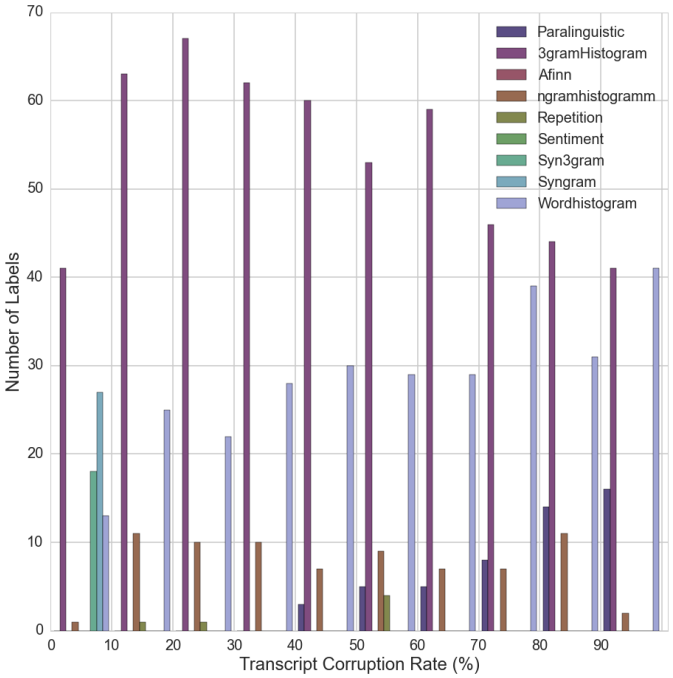

Fig. 7: Categories of the 100 Most Predictive Features for the Father Behavior Model.

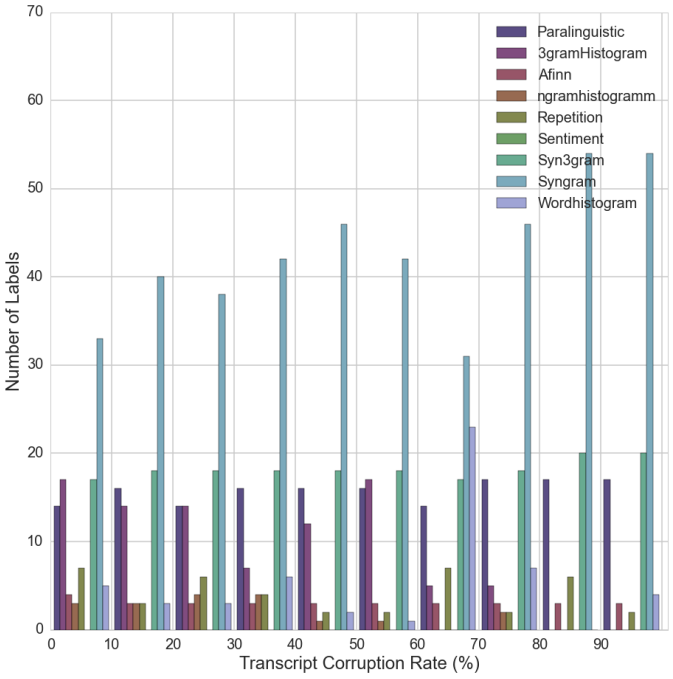

Fig. 8: Categories of the 100 Most Predictive Features for the Father Emotion Model.

5.3 Emotion and Behavior: Relationships

After analyzing behavior and emotion models separately, we turned our attention to analyzing the relationship between behavior and emotion during parent-child interaction. First we examined frequent pairs of emotion and behavior labels. Some notable common pairs were 'Information Question' and 'Interest', as well as 'Neutral Talk' and 'Suprise'.

To examine this relationship further, we used past behavior, past emotion, and current emotion to predict current behavior. Using SVMs, we found that for these parents, emotion is more predictive of behavior than vice versa (Figure 11). Looking at the relevant features in Figure 12, we see that current emotion is best predicted by previous emotion, while current behavior is best predicted by current emotion. Emotion is not only the better predictor in all cases, it is more reliably predictable based on the previous state than behavior.

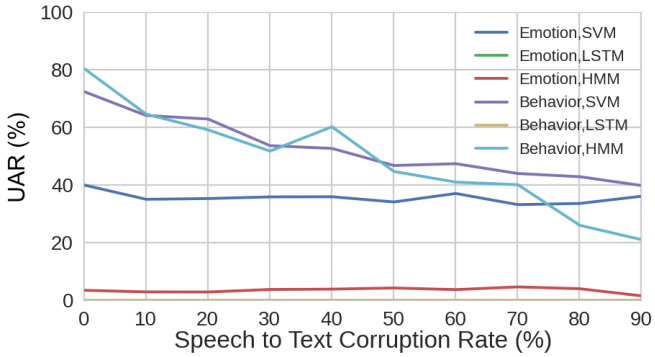

Fig. 9: Unweighted Average Recognition Rate of Child Models.

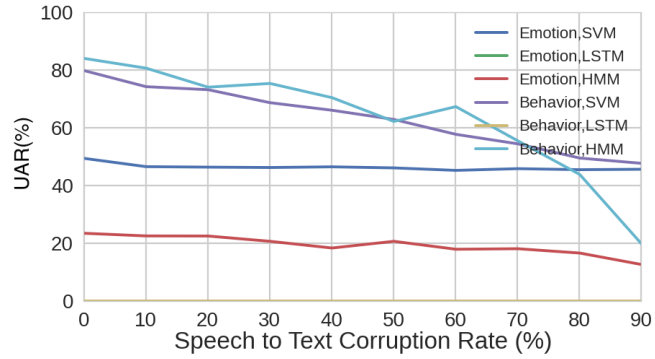

Fig. 10: Unweighted Average Recognition Rate of Parent Models.

There are more interesting insights to draw from this type of emotion/behavior analysis on an individual basis. With enough individual data, this type of analysis could provide answers to questions like: How does your mood affect your interaction with your child? How does your behavior affect your child's mood? How synchronous are your emotions with your child, and are you able to improve?

6 Discussion and Future Work

Our results provide interesting insight into the relative value of linguistic and paralinguistic features for predicting emotion and behavior, and offers interesting insights into the relationship of emotion and behavior itself.

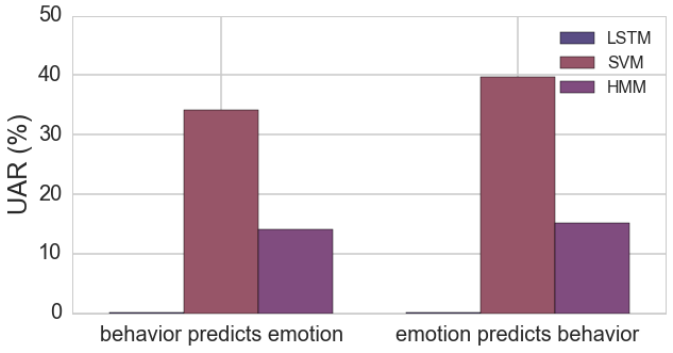

Fig. 11: Emotion and Behavior Models.

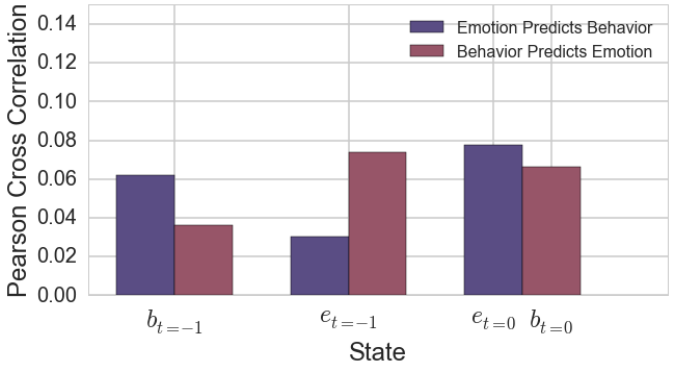

Fig. 12: Relative Feature Relevance for State Prediction.

The strong behavior recognition rate is encouraging, though a more thorough analysis of the effect of non-uniform class distribution might reveal some interesting shortcomings for the less common classes.

The recognition rate for emotion is much lower, but also encouraging. This real-world audio was full of very loud 'play sounds', drastic changes in distance and clarity of the speakers, concurrent talking, and abrupt, inaccurate separation of utterances. Pre-processing the audio could dramatically improve this low recognition rate.

It is interesting to see that HMM works well for behavior, which have specific call-and-response patterns, but does much more poorly for emotion estimation. Emotion in parent-child interactions tend to be purposely over-emphasized and rapidly changing, so it is not surprising that pure paralinguistic analysis of an utterance provides the best results.

Our LSTM RNNs were terribly under-seeded. However, with enough new data, it is still our belief that this deep learning approach can reveal and account for more interesting latent time-dependencies in the data. Eventually we hope to code enough data to train such a model.

6.1 Full System Architecture

At the outset, we had a vision of a mobile application that could record speech in parent-child interaction and provide meaningful, immediate insight and analysis to the parent. To understand the feasibility of such a system (outlined in Fig 13), we built a native application that records and uploads audio to a server, as well as scripts that (1) automatically separate child and parent speakers, and (2) generate

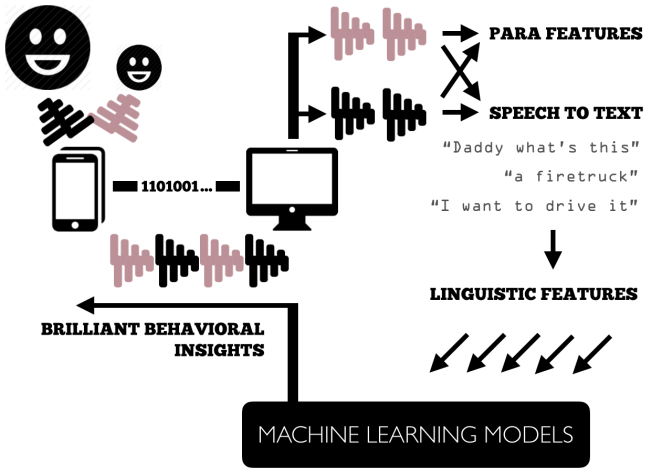

Fig. 13: Schematic of the Full System.

transcripts from raw audio using the IBM Watson API. The separation script uses the amplitude envelope of the incoming audio, as well as F_0 estimation based on running auto-correlations, to pull out the distribution of F_0 for each 20 ms speech section of the entire interaction. It then fits two Gaussians to the F_0 distribution to make a decision boundary for whether an utterance belongs to the child or a parent. This data is incredibly noisy, and to work well, it will likely be necessary to include more advanced logic (even potentially a machine learning model for speaker separation).

We also characterized the IBM Watson speech-to-text service against 500 utterances from the CHILDES database, and found a 14.6% accuracy. This underrepresents the adult accuracy (closer to 25%) and overrepresents child accuracy (closer to 0%). This is not surprising– the audio quality is poor and child speech models are a niche technology that requires specific forms of pre-processing. We expect much higher accuracy can be rectified by training our own child speech model, and pre-processing the audio.

Currently, the pieces of this system are not robust enough to support a working implementation of our model. However, there are concrete and clear steps forward to solve many of the most pressing issues that stand in the way of a real-world implementation.

6.2 Applications

The first application we're pursuing is an iPhone application that can provide useful analysis and advice, democratizing the early intervention techniques to a wide audience. This application will eventually be able to adapt to individual behavioral and emotional patterns, and give very specific advice based on parenting style that is relevant to the child's learning level and personality.

Another application we foresee is an automatic TV cartoon grader for parents. [41] using DPICS to analyze cartoons, and cautioned that 'family-friendly' programming does not always model prosocial behavior. This tool would be useful to inform both personal and policy decisions around family programming.

We also see this tool having an impact on a few aspects of research. Current PCIT therapists manually code using DPICS on a regular basis. With further refinement, this system could automate that work.

We believe, with enough adoption, this model could lead to more nuanced and specific advice for achieving specific learning outcomes. Should parenting style adapt to the child's personality, behavior, or mood? What are the appropriate dynamics of parenting strategy to promote behavioral and language learning?

Finally, the estimation of emotion and behavior also makes this a very useful tool for HCI research around any parent-child intervention. With this technology, we can quantify the efficacy of an intervention to drive usage, alter mood, and change behavior.

6.3 Concerns and Future Work

The first issue we will address is the lacking quantity and accuracy of our DPICS coding. We have several ideas for crowd-sourcing reliable recommendations, however our primary strategy will be to contact DPICS practitioners and ask for access to their professionally coded audio. Other options include paying a professional to code our data, or receiving DPICS training. With a large, accurate corpus of data, we can draw stronger, more interesting, and more predictable conclusions.

Our other main challenges center around audio quality, which affects paralinguistic analysis, speaker utterance separation, and speech-to-text accuracy. We are planning to incorporate noise reduction, transient detection (for blocks hitting and clapping), and voice activity detection to help mitigate these concerns. Spatialized audio recording– leveraging stereo microphones– could significantly improve speaker separation and noise reduction at the expense of additional hardware.

We will also need to train our own speech-to-text model using CMU Sphinx or another open service for accurate child speech recognition instead of using an off-the-shelf solution.

7 Conclusion

There are serious, life-long implications for the 35% of children entering kindergarten already behind in language skills and social-emotional development. Research has shown that differences in the quality and style of parent-child interaction are the major cause of this disparity. In this paper, we presented a speech analysis system to automatically assess emotion and behavior in the parent-child dyad, with the goal of enabling technology to address this problem.

We showed that both paralinguistic and linguistic speech features are important for these models to succeed. We demonstrated that paralinguistic features are important for predicting emotion, particularly in the child model. For behavioral coding, linguistic features were the most important. Repetition, though representing a small fraction of labels, was also a great indicator. Sentiment analysis of a conversation transcript was not.

We demonstrated a high success rate of 80% for automatically predicting behavior in the parent-child dyad using HMMs. This success is sensitive to transcription errors, so an accurate speech-to-text engine is important. However, even with high levels of corruption the recognition rate maintains around 50%.

We demonstrated a success rate of 45% for emotion prediction using SVMs. This success rate is encouraging, though it speaks to the difficulties of working with noisy, real-word data. We believe we can improve this substantially with further audio pre-processing and noise reduction techniques.

Beyond HMMs and SVMs, we attempted to create a deep learning model that accounts for temperament, mood, and short-term dyadic influence on the current predictions. Unfortunately, our relatively small dataset and our large, non-uniform classification space rendered our efforts to train a neural network futile. However, we now have an infrastructure in place to quickly train the model once we have more data, and we're still optimistic that unsupervised deep learning techniques will eventually overtake our HMMs in accuracy.

After analyzing behavior and emotion models separately, we turned our attention to analyzing the relationship between behavior and emotion during parent-child interaction. Again using SVMs, we found that that mood is more predictive of behavior than vice versa. We believe there are interesting insights to draw from this type of analysis on an individual basis to show powerful links between behavior and mood.

Finally, we built out (1) an algorithm for speaker/utterance separation, (2) a UI for speech data collection, and (3) an interface for automatic speech-to-text translation, in order to estimate how difficult it would be to build a real world, automated system around this machine-learning model. While there are several technical hurdles still to overcome, we believe they are all manageable.

This work represents a thoughtful analysis of a comprehensive model for automatic classification of parent-child interaction. It is intentionally geared towards real-world applications using real-world data. We are hopeful that this paper lays the groundwork for the development of a useful and robust automatic system for parents to use with their children, psychologists to use with their patients, and researchers to use with their subjects.

References

[1] J. B. Isaacs, "Starting school at a disadvantage: The school readiness of poor children. the social genome project.," *Center on Children and Families at Brookings*, 2012.

[2] S. Daily, M. Burkhauser, and T. Halle, "A review of school readiness practices in the states: Early learning guidelines and assessments. early childhood highlights. volume 1, issue 3.," *Child Trends*, 2010.

[3] B. Hart and T. R. Risley, "The early catastrophe: The 30 million word gap by age 3," *American educator*, vol. 27, no. 1, pp. 4–9, 2003.

[4] J. L. Malin, N. J. Cabrera, and M. L. Rowe, "Low-income minority mothers' and fathers' reading and children's interest: Longitudinal contributions to children's receptive vocabulary skills," *Early childhood research quarterly*, vol. 29, no. 4, pp. 425–432, 2014.

[5] D. S. Arnold and G. J. Whitehurst, "Accelerating language development through picture book reading: A summary of dialogic reading and its effect.," 1994.

[6] F. J. Zimmerman, J. Gilkerson, J. A. Richards, D. A. Christakis, D. Xu, S. Gray, and U. Yapanel, "Teaching by listening: The importance of adult-child conversations to language development," *Pediatrics*, vol. 124, no. 1, pp. 342–349, 2009.

[7] D. G. K. Nelson, K. Hirsh-Pasek, P. W. Jusczyk, and K. W. Cassidy, "How the prosodic cues in motherese might assist language learning," *Journal of child Language*, vol. 16, no. 01, pp. 55–68, 1989.

[8] J. L. Cooper, R. Masi, and J. Vick, "Social-emotional development in early childhood: What every policymaker should know," 2009.

[9] D. Benoit, "Infant-parent attachment: Definition, types, antecedents, measurement and outcome," *Paediatrics & child health*, vol. 9, no. 8, p. 541, 2004.

[10] A. M. Conway, "The development of emotion regulation: The role of effortful attentional control and positive affect," 2005.

[11] A. Adger-Antonikowski, "A functionalist perspective of language ability and behavioral synchrony in the development of emotion regulation," 2008.

[12] S. A. Denham, S. M. Renwick, and R. W. Holt, "Working and playing together: Prediction of preschool social-emotional competence from mother-child interaction," *Child Development*, vol. 62, no. 2, pp. 242–249, 1991.

[13] J. L. Carson and R. D. Parke, "Reciprocal negative affect in parent-child interactions and children's peer competency," *Child development*, vol. 67, no. 5, pp. 2217–2226, 1996.

[14] B. H. Ellis, "Relations between emotion language and emotion regulation in maltreated preschoolers," 2000.

[15] P. M. Cole, L. M. Armstrong, and C. K. Pemberton, *The role of language in the development of emotion regulation*, pp. 59–77. Child development at the intersection of emotion and cognition., American Psychological Association, Washington, DC, 2010.

[16] W. S. Gilliam and P. B. de Mesquita, "The relationship between language and cognitive development and emotional-behavioral problems in financially-disadvantaged preschoolers: A longitudinal investigation," *Early Child Development and Care*, vol. 162, no. 1, pp. 9–24, 2000.

[17] W. B. Brooke Graham Doyle, "Promoting emergent literacy and social-emotional learning through dialogic reading," *The Reading Teacher*, vol. 59, no. 6, pp. 554–564, 2006.

[18] D. M. Almeida, E. Wethington, and A. L. Chandler, "Daily transmission of tensions between marital dyads and parent-child dyads," *Journal of Marriage and the Family*, pp. 49–61, 1999.

[19] E. K. Adam, M. R. Gunnar, and A. Tanaka, "Adult attachment, parent emotion, and observed parenting behavior: Mediator and moderator models," *Child development*, vol. 75, no. 1, pp. 110–122, 2004.

[20] H. Switzer, *The use of parent-child interaction therapy with parents and children referred by a child protective service*. 1997.

[21] B. J. Grosz, S. Weinstein, and A. K. Joshi, "Centering: A framework for modeling the local coherence of discourse," *Computational linguistics*, vol. 21, no. 2, pp. 203–225, 1995.

[22] A. Karpathy, J. Johnson, and F.-F. Li, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[23] A. B. Tempel, S. M. Wagner, and C. B. McNeil, "Parent-child interaction therapy and language facilitation: The role of parent-training on language development.," *The Journal of Speech and Language Pathology–Applied Behavior Analysis*, vol. 3, no. 2-3, p. 216, 2009.

[24] A. T. Naik-Polan and K. S. Budd, "Stimulus generalization of parenting skills during parent-child interaction therapy.," *Journal of Early and Intensive Behavior Intervention*, vol. 5, no. 3, p. 71, 2008.

[25] M. E. Goldfine, S. M. Wagner, S. A. Branstetter, and C. B. Mcneil, "Parent-child interaction therapy: An examination of cost-effectiveness.," *Journal of Early and Intensive Behavior Intervention*, vol. 5, no. 1, p. 119, 2008.

[26] T. Hollenstein and M. D. Lewis, "A state space analysis of emotion and flexibility in parent-child interactions.," *Emotion*, vol. 6, no. 4, p. 656, 2006.

[27] D. Ridgeway, E. Waters, and S. A. Kuczaj, "Acquisition of emotion-descriptive language: Receptive and productive vocabulary norms for ages 18 months to 6 years," *Developmental psychology*, vol. 21, no. 5, pp. 901–908, 1985.

[28] B. MacWhinney, *The CHILDES project: The database*, vol. 2. Psychology Press, 2000.

[29] Kendall, "Oscaar, http://oscaar.ling.northwestern.edu," 2010.

[30] V. Hazan, M. Pettinato, and O. Tuomainen, "kidlucid: London ucl children's clear speech in interaction database.,"

[31] E. F. Masur and J. B. Gleason, "Parent–child interaction and the acquisition of lexical information during play.," *Developmental Psychology*, vol. 16, no. 5, p. 404, 1980.

[32] S. Eyberg, M. Nelson, M. Duke, and S. Boggs, "Manual for the dyadic parent-child interaction coding system," *Retrieved July*, vol. 28, p. 2006, 2005.

[33] J. L. Bessmer, *The Dyadic Parent-Child Interaction Coding System II (DPICS II): Reliability and validity.* PhD thesis, 1998.

[34] M. M. Deskins, *The Dyadic Parent-Child Interaction Coding System II (DPICS II): Reliability and validity with school aged children.* PhD thesis, 2005.

[35] R. C. Foote, *The Dyadic Parent-Child Interaction Coding System II (DPICS II): Reliability and validity with father-child dyads.* PhD thesis, 2000.

[36] B. Cardoso, O. Santos, and T. Romão, "On sounder ground: Caat, a viable widget for affective reaction assessment," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pp. 501–510, ACM, 2015.

[37] F. Å. Nielsen, "Afinn," 2011.

[38] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[40] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions-a new perspective on feature selection.," in *Proc. of INTERSPEECH*, pp. 2044–2048, ISCA, 2013.

[41] L. Klinger, "What are your children watching? a dpics-ii analysis of parent-child interactions in television cartoons," 2006.