Interspeech 2018
2-6 September 2018, Hyderabad

A Knowledge Driven Structural Segmentation Approach for Play-Talk Classification during Autism Assessment

Manoj Kumar[1], *Pooja Chebolu*[1], *So Hyun Kim*[2], *Kassandra Martinez*[2], *Catherine Lord*[2], *Shrikanth Narayanan*[1]

[1]University of Southern California, Los Angeles, United States
[2]Department of Psychiatry, Weill Cornell Medicine, New York, United States

Abstract

Automatically segmenting conversational audio into semantically relevant components has both computational and analytical significance. In this paper, we segment play activities and conversational portions interspersed during clinically-administered interactions between a psychologist and a child with autism spectrum disorder (ASD). We show that various acoustic-prosodic and turn-taking features commonly used in the literature differ between these segments, and hence can possibly influence further inference tasks. We adopt a two-step approach for the segmentation problem by taking advantage of the structural relation between the two segments. First, we use a supervised machine learning algorithm to estimate class posteriors at frame-level. Next, we use an explicit-duration hidden Markov model (EDHMM) to align the states using the posteriors from the previous step. The durational distributions for both play and talk regions are learnt from training data and modeled using the EDHMM. Our results show that speech features can be used to successfully discriminate between play and talk activities, each providing important insights into the child's condition.

Index Terms: Autism spectrum disorder, audio segmentation, explicit-duration hidden markov models

1. Introduction

Autism spectrum disorder (ASD) refers to a group of heterogeneous neuro-developmental disorders that are characterized by impairments in social communication and reciprocity. Estimates of ASD prevalence among children have been increasing steadily, from 1 in 150 (2002) to 1 in 68 (2014) [1].

Computational methodologies including objective speech/language feature analyses of conversational interactions during diagnostic sessions combined with machine learning [2, 3, 4] have helped validate hypotheses about behavioral markers and have provided insights into the diagnostic model. For instance, [5] associated subjective perception of awkward prosody with prosodic features extracted from the child's speech, and showed that the features were significant in classifying between ASD subjects and typically developing controls. Furthermore, studies [6, 7] have illustrated significant correlations between the interlocutor's prosody, language use, and discourse linguistic features and the subject's ASD severity.

Observational diagnostic sessions are designed to examine different socio-communicative behaviors [8, 9], and thus involve multiple segments with different objectives. For instance, the Autism Diagnostic Observation Schedule (ADOS) [10], which is considered to be the gold standard for autism diagnosis, contains 10-15 different activities based primarily on the individual's expressive language level and secondarily on chronological age. Many studies typically analyze acoustic/linguistic data from a subset of these subtasks [11, 4, 12], rather than the entire session as a whole. Therefore, manual annotations are necessary to segment a session before proceeding to feature extraction tasks.

In this paper, we look at data from the recommended administration for the Brief Observation of Social Communication Change (BOSCC). This is a 12-minute semistructured interaction that involves 8 minutes of play and 4 minutes of conversation between an individual and an examiner. We segment the play and talk regions by using the knowledge about the order of the segments within a session. We split the problem into two steps - using a supervised classifier built with training data, we first obtain a rough confidence score between play and talk at each time point. Next, we find the best possible state alignment using the confidence scores. We show that modeling the state duration using an Explicit-Duration Hidden Markov Model (EDHMM) provides segmentation with high accuracy and is robust to classifier errors, thus enabling us to analyze each of these segments individually.

2. Background

2.1. BOSCC

Brief Observation of Social Communication Change (BOSCC) [13] is a recently proposed treatment outcome measure to track changes in social-communication over the course of ASD treatment. The scheme is designed to be applicable in a variety of collection scenarios (clinics, homes, research labs) and captures a broad range of behavioral features of interest. In this work, we consider two modules which are applicable for verbally fluent children.

A typical BOSCC session involves the child taking part in play activities and engaging in conversation (henceforth referred to as 'play' and 'talk' respectively) with an interviewer (examiner). During play, the child is presented with a box of toys and is encouraged to choose one among them. The interviewer allows the child to take the lead during this activity, while also commenting on play and introducing their own ideas. This is followed by a semi-naturalistic conversation with the child, without the toys. The interviewer asks a few questions but also offers leads for the child to follow up on. The BOSCC typically involves 4 minutes of play followed by 2 minutes of conversation, and the sequence repeats one more time, resulting in a play-talk-play-talk sequence lasting about 12 minutes. We note that: (1) the segment boundaries are inherently subjective, since it is not possible to specify an exact time-instant where the segment changes from talk to play, or vice versa; and (2) the play segments may contain substantial amount of speech depending on how the session progresses.

Previous research has reported a close association between play skills at an early age and linguistic development in children with ASD [14, 15]. Further, toy play [16] was found to possess distinct information about rates of language development. More recently, children with ASD were shown [17] to exhibit different levels of eye contact with the interviewer between talk and play during BOSCC. Considering the unique insights that play

10.21437/Interspeech.2018-1516

activities can offer, and the difference in administration against conversation portions, we hypothesize that computationally extracted features differ significantly between them, and hence it may become necessary to segment them. In the following analysis, we demonstrate the same using audio-based features from the literature.

2.2. Acoustic-Prosodic and Turn-Taking feature analysis

We chose acoustic-prosodic and turn-taking features from the speech of both participants that have shown to be significantly correlated with, and predictive of, the child's diagnostic condition [18, 19]. We extract log-pitch and intensity contours using Praat [20] and normalize them per speaker, per session to remove individual variability. The contours are then parameterized using a second-order polynomial. The functionals (mean, std) of the coefficients are used as prosodic features. We also use the silence fraction, individual speaking fractions, and median latencies for both speakers as turn-taking features.

For our experiments, we consider audio from 30 BOSCC sessions that were collected across four different clinical centers. The segment boundaries between talk and play were manually annotated by one of the authors. A subset (14) of these sessions were also annotated for speaker boundaries. We compute each feature for each talk/play segment, resulting in 56 samples per feature. We compute significance between the features from play and talk segments using the Wilcoxon signed pair test and treat the feature distributions as non-parametric, since the subjects do not have the same autism severity score (based on ADOS Calibrated Severity Scores (ADOS-CSS) see [21]). The results are presented in Table 1.

While the turn-taking features suggest that both speakers speak longer during talk, the intensity features show inconsistent trends between the child and adult. The higher intensity variation from the child's speech may be attributed to more excitement while playing with the toys and increased presence of non-speech vocalizations like laughter, the adult's intensity variations do not show a consistent trend between the segments. Nevertheless, this warrants a closer look at the features from the two activities.

Table 1: *Significant features* ($p < 0.05$) *and their trends between 'play' and 'talk' activities. n.s denotes not significant*

Feature	**Trend during 'play' w.r.t 'talk'**	
	Child	**Adult**
Speaking Fraction (%)	Lower	Lower
Intensity σ (curvature, slope, intercept)	(Higher,Higher, n.s)	(Higher,Lower, Lower)

3. Methods

We adopt a two-step approach towards the segmentation problem - a supervised classifier at the frame level, followed by optimal state alignment of *play/talk* segments at the session level. The role of the classifier is to provide estimates of the class posteriors at each time instant, which are used as the state emission probabilities during the alignment process. An overview of the segmentation system is presented in Figure 1.

3.1. Supervised Classifiers Considered

3.1.1. Support Vector Machines

Support Vector Machines (SVMs) work by estimating a maximum margin hyperplane that separates features from different classes, possibly in a higher-dimensional space than the features themselves. Application of non-linear kernel functions combined with their discriminative nature have made SVMs one of the most popular choices for off-the-shelf supervised classifiers.

3.1.2. Logistic Regression

Logistic regression estimates the probability of a categorical dependent variable using one or more independent variables using the logistic function. Binary logistic regression is a natural choice for the two-class classification considered in this work and is considered robust to outliers.

3.1.3. Neural Networks

Neural networks have outperformed traditional learning paradigms in a large number of domains including speech recognition, computer vision, and natural language processing. Considering the limited availability of training data, we experiment with a simple architecture in this work.

3.2. Finding optimum state alignment

Hidden Markov models (HMMs) have been a popular modeling choice in speech processing applications, specifically for speech recognition. A Hidden Markov Model (HMM) is doubly stochastic - both the underlying state duration and the observation given state are modeled by probability distributions. In this work, the observation sequence is the feature representation across a session and the state sequence is represented by the *play/talk* label at every frame. Given an observation sequence $o_{1:T}$, the problem of computing the optimum underlying state sequence $s_{1:T}$ for an HMM can be efficiently solved using the Viterbi alogorithm.

However, the definition of HMMs implicitly assumes that the state durations follow a geometric distribution. This becomes a limiting factor when the durations of sound units (e.g phonemes) need to be modeled. Hidden semi-Markov models (HSMMs) were first proposed [22] as an alternative to HMMs in speech recognition. In its most generic form, an HSMM is defined [23] using λ=$\{a_{(i,d')(j,d)}, b_{j,d}(v_{k_1:k_d}), \pi_{i,d}\}$, where

$$\begin{aligned} a_{(i,d')(j,d)} &= P[S_{[t+1:t+d]} = j | S_{[t-d'+1,t]} = i] \\ b_{j,d}(v_{k_1:k_d}) &= P[o_{t+1:t+d} | S_{t+1:t+d} = j] \\ \pi_{i,d} &= P[S_{[t-d+1:t]} = j], \quad t \leq 0, d \in D \end{aligned} \tag{1}$$

$S_{[t+1:t+d]} = j$ denotes that state j starts at time $t+1$ and ends at time $t+d$. D is the set of integers representing possible state durations, and $v_{k_1:k_d}$ represents the set of observable sequences $\mathcal{V} \times \mathcal{V} \times ..\mathcal{V}$, with $\mathcal{V} = \{v_1, v_2, ...v_k\}$ being the set of observable values. An important property of HSMM is that a single state can last for multiple time units, emitting a sequence of observations, unlike the conventional HMM.

Under the assumptions that the transitions from state i to j are independent of the duration of state i, and the duration of state j is independent of the previous state, $a_{(i,d')(j,d)}$ reduces to $a_{ij}p_j(d)$ where $p_j(d)$ represents the durational probability distribution for state j; resulting in the EDHMM. Similar to the basic HMM, dynamic programming algorithms exist for finding the optimum state sequence in EDHMM. We define $\delta_t(j,d)$ as the maximum likelihood that the observed state sequence until t ends in state j with duration d.

$$\begin{aligned} \delta_t(j,d) &= \max_{s_{1:t-d}} P[s_{1:t-d}, S_{[t-d+1:t]} = j, o_{1:t} | \lambda] \\ &= \max_{\substack{i \in S \setminus j \\ d' \in D}} \delta_{t-d}(i,d') a_{ij} p_j(d) b_{j,d}(o_{t-d+1:t}) \end{aligned} \tag{2}$$

The previous state selected by $\delta_t(j,d)$ is recorded using $\Psi_t(j,d) = (t-d, i^*, d^*)$ where i^* is the previous state, d^*

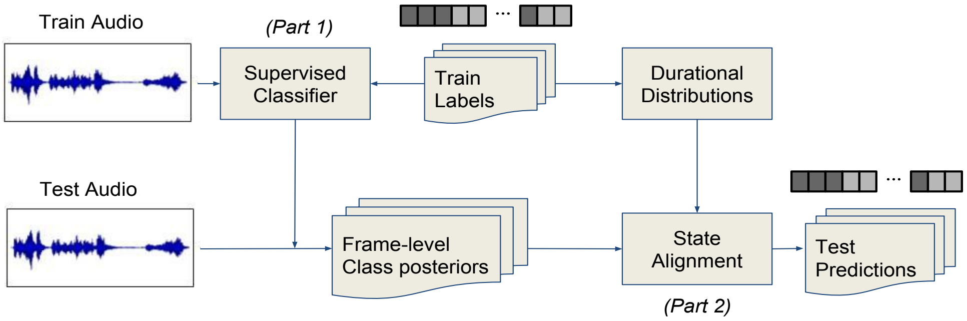


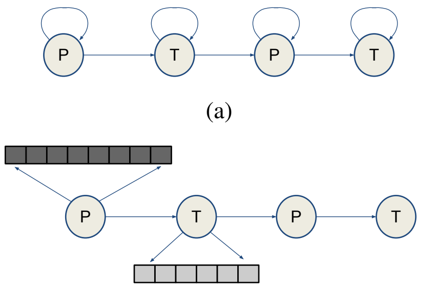
Figure 1: *Illustrating the two-step approach involving a supervised classifier and a state alignment system*

is its duration and $(t-d)$ its ending time. Note that i^* and d^* are obtained as the solutions for (2).

Algorithm 1: Estimating optimum State Sequence for EDHMM

```
Inputs:
$a_{ij}\ \forall i, j \in S$   ⇒ transition probability
$b_j(o_t)\ \forall j \in S, \forall t \in [1, T]$ ⇒ emission probability
$p_j(d)\ \forall j \in S, \forall d \in [1, T]$ ⇒ durational distribution
Output:
$x$   ⇒ state sequence
Initialize:
$\delta_t(j, d) = -\infty\ \forall t, d \in [1, T], \forall j \in S$
$\Psi_t(j, d) = (0, 0, 0)\ \forall t, d \in [1, T], \forall j \in S$
for $t = 2 : T$ do
    for $d = 1 : t$ do
        for $j = 2 : |S|$ do
            $Q_{i,d'} = -\infty\ \forall i \in S, \forall d' \in [1, T]$
            for d' = 1:T do
                $i \leftarrow j - 1$
                $Q_{i,d'} \leftarrow \delta_{t-d}(i, d') + \log a_{ij} +$
                  $\log p_j(d) + \Sigma_{k=t-d:t}^{t} \log b_j(o_k)$ ;
                // Viterbi update
            end for
            $i^*, d^* \leftarrow \text{argmax}_{i,d'}(Q_{i,d'})$
            $\delta_t(j, d) \leftarrow Q_{i^*,d^*}$
            $\Psi_t(j, d) \leftarrow (t - d, i^*, d^*)$
        end for
    end for
end for
$j^*, d^* = \text{argmax}_{j,d}(\delta_T(j, d))$ ;   // Backtracking
$t_{prev}, j_{prev}, d_{prev} \leftarrow (T, j^*, d^*)$
$x \leftarrow (t_{prev}, j_{prev}, d_{prev})$
while $t_{prev} > 0$ do
    $x = x \bigcup \Psi_{t_{prev}}(j_{prev}, d_{prev})$
    $(t_{prev}, j_{prev}, d_{prev}) \leftarrow \Psi_{t_{prev}}(j_{prev}, d_{prev})$
end while
```

We illustrate the different ways of modeling play and talk activities in a BOSCC session in Figure 2. Under the assumption of a left-right HMM (as is the case of this work), we can improve the computational efficiency at (2) by constraining i to be the previous state of j [24, 25]. We provide the algorithm for finding the optimum state sequence using EDHMM in Algorithm 1.



(a)

(b)

Figure 2: *Modeling a typical BOSCC session using a HMM (a) and EDHMM (b). Each state in HMM corresponds to a single observation frame, while EDHMM allows for a sequence of observations per state. $P \rightarrow$ Play, $T \rightarrow$ Talk*

4. Experiment

As mentioned in Section 2, 30 BOSCC sessions were manually annotated for play and talk segment boundaries. 13 dimensional Mel frequency cepstral coefficients (MFCCs) were extracted using short-time windows of length 100ms and shifted every 50ms. The features were normalized to have zero mean and unit variance per BOSCC session to remove any session-related variabilities. In order to capture information from a large enough time frame, we compute the mean and standard deviation of the coefficients every 2 seconds, resulting in a 26 dimensional vector. This feature representation is used to train the supervised classifiers and estimate class posteriors for aligning HMM states.

Among the supervised classifiers (Part 1), we use an SVM with RBF kernel for the nonlinear feature transformation. Posterior probabilities are estimated using the Platt scaling method [26]. In the case of the neural network, we use 2 hidden layers with 32 neurons each and a rectified linear function for the activation. During the training phase, network connections are randomly dropped with a probability of 0.2 in order to reduce effects of overfitting. The network is optimized with Adam [27] to minimize the binary cross entropy loss. Training is performed for 30 epochs using a batch size of 128. Since the duration of play activities is more than talk, we randomly resample features from the latter during the training phase for all supervised classifiers to account for class imbalance.

We experiment with both conventional HMM and EDHMM for finding the optimal state sequence, using the models presented in Figure 2. The state transition probabilities are estimated using the labels from training data for both models. However, we note that the self transition probability values are irrelevant in the case of EDHMM since one state emits a sequence of observation vectors instead of self-transition. We use the class posteriors estimated from Part 1 for the emission probabilities. Further, we estimate the durational distributions for both activities by normalizing and smoothing the histograms of durations (Figure 3) obtained from manual annotations.

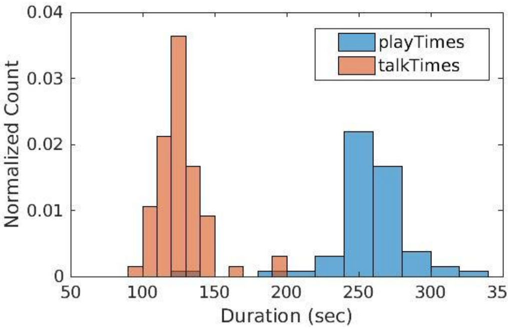


Figure 3: *Durational densities for play and talk activities collected from all 30 sessions represented using probability mass functions. Similar distributions were estimated at every fold using the training data.*

The sessions are split into 6 folds, with the first five folds treated as training data, and the sixth fold treated as test data. The test fold is switched so that every session is considered as test data once during the entire experiment. We report the mean frame-level accuracy to evaluate segmentation performance at different stages of experimentation. For the baseline system, we smooth the decisions from the supervised classifier using a median filter. At each fold, we treat a subset (5 sessions) of the training set as the development set. 20 sessions are used for training the classifier, the filter window size that maximizes accuracy for the development set is chosen as the optimal window size, and is used to smooth the predictions for test set.

4.1. Results and Discussion

From Table 2, we notice that the proposed EDHMM based approach provides the best results overall. All supervised classifiers (Part1) are only able to achieve a moderate improvement in classification accuracy over majority (67.30%). The primary reason is perhaps due to the presence of significant talking regions during the play activities and background noises from furniture during talk activities which might resemble the toy noises during play. We do not aim for perfect classification and depend on the state alignment to correct the errors made at this stage.

Smoothing the predicted labels (the baseline system for segmentation) improves the accuracy consistently across classifiers. Although it provides a better estimate for the segmentation when *play/talk* order is unknown, the performance is still not satisfactory and necessitates an alignment system.

Table 2: *Mean frame-level accuracy (%) for different choices of supervised classifiers, and at different stages of segmentation. 'Part1' denotes the classifier, possibly one among SVM, logReg and Neural Network.*

System	SVM	Neural Net	LogReg
Part1	74.04	74.82	73.36
Baseline (Part1+Smooth)	79.60	78.51	82.02
Part1 + HMM	70.08	69.39	69.60
Part1 + EDHMM	87.97	87.95	87.04
Part1 + EDHMM (2σ)	**91.26**	**89.58**	**91.03**

The HMM (Part1 + HMM) system performed worse than using only the supervised classifier. Upon closer inspection, we found that the HMM was unable to align most of the sessions, and predicted play activity (the first state) for the entire duration. We suspect that the low values for state transition probabilities ($P_{talk \rightarrow play} \approx 0.015, P_{play \rightarrow talk} \approx 0.007$) proved insufficient to correct for errors during frame classification from Part 1. In contrast, the EDHMM constrains the state durations using the duration densities learnt from the training data. It was able to segment the sessions significantly better than using only the supervised classifier and the baseline. Removing outliers ($\mu \pm 2\sigma$) from the durational densities for both play and talk from the training data further improved the accuracy consistently for all supervised classifiers.

However, we note that the EDHMM aligned only 3 activities for 6/30 sessions. Further analysis revealed that the supervised classifier had significant errors while making decisions at frame level. Figure 4 (*bottom*) shows an example of one such session, as opposed to a session with near-perfect alignment (*top*). The neural network is used as the supervised classifier in both cases. The network predicts noisy labels (during Part1) for a prolonged duration within the second play activity, which resulted in highly incorrect segment boundaries.

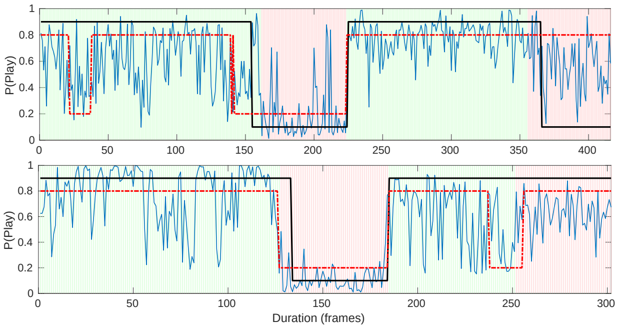


Figure 4: *State sequence predicted by the EDHMM (solid, black line) is affected by mistakes in class posteriors (continuous, blue curve). Ground truth labels are indicated using the background colors and the baseline is represented with broken red line*

5. Conclusions

We explored the task of segmenting a semi-structured, naturalistic interaction between a psychologist and a child with ASD. The play and talk activities are designed to create opportunities to elicit varying socio-communicative behaviors, and hence the patterns of interactions between the dyads may vary across these activities. We first showed that audio-based features used in the literature were significantly different between play and talk activities. Using the knowledge of *play/talk* order, we modeled the session using an explicit duration hidden markov model. We show it is possible to reliably segment using a two-step methodology.

We observe that although the EDHMM is robust to errors, there is room for improvement in terms of classifier accuracy. Further work will consider feature representations and algorithms robust to noise conditions, including any discriminative information from the lexical modality between the two activities. We also aim to investigate the relation between the automatically measured speech dynamics of interaction within each activity and the childs specific and overall social communication skills as evaluated by trained human coders.

6. Acknowledgements

This work was supported by funds from the National Institute of Health (NIH) and Simons Foundation (SFARI).

7. References

[1] J. Baio, "Prevalence of autism spectrum disorder among children aged 8 yearsautism and developmental disabilities monitoring network, 11 sites, united states, 2010," *Morbidity and Mortality Weekly Report: Surveillance Summaries*, vol. 63, no. 2, pp. 1–21, 2014.

[2] T. J. Goh, J. Diederich, I. Song, and M. Sung, "Using diagnostic information to develop a machine learning application for the effective screening of autism spectrum disorders," in *Mental health informatics*. Springer, 2014, pp. 229–245.

[3] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, 2015.

[4] M. Kumar, R. Gupta, D. Bone, N. Malandrakis, S. Bishop, and S. S. Narayanan, "Objective language feature analysis in children with neurodevelopmental disorders during autism assessment." in *INTERSPEECH*, 2016, pp. 2721–2725.

[5] D. Bone, M. P. Black, A. Ramakrishna, R. B. Grossman, and S. S. Narayanan, "Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism." in *INTERSPEECH*. ISCA, 2015, pp. 1616–1620.

[6] R. Paul, L. D. Shriberg, J. McSweeny, D. Cicchetti, A. Klin, and F. Volkmar, "Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 35, no. 6, p. 861, 2005.

[7] S. Peppé, J. McCann, F. Gibbon, A. OHare, and M. Rutherford, "Receptive and expressive prosodic ability in children with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 4, pp. 1015–1028, 2007.

[8] N. Akshoomoff, C. Corsello, and H. Schmidt, "The role of the autism diagnostic observation schedule in the assessment of autism spectrum disorders in school and community settings," *The California School Psychologist*, vol. 11, no. 1, pp. 7–19, 2006.

[9] C. A. Mazefsky and D. P. Oswald, "The discriminative ability and diagnostic utility of the ados-g, adi-r, and gars for children in a clinical setting," *Autism*, vol. 10, no. 6, pp. 533–549, 2006.

[10] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler, "Austism diagnostic observation schedule: A standardized observation of communicative and social behavior," *Journal of autism and developmental disorders*, vol. 19, no. 2, pp. 185–212, 1989.

[11] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[12] G. Celani, M. W. Battacchi, and L. Arcidiacono, "The understanding of the emotional meaning of facial expressions in people with autism," *Journal of autism and developmental disorders*, vol. 29, no. 1, pp. 57–66, 1999.

[13] R. Grzadzinski, T. Carr, C. Colombi, K. McGuire, S. Dufek, A. Pickles, and C. Lord, "Measuring changes in social communication behaviors: Preliminary development of the brief observation of social communication change (boscc)," *Journal of Autism and Developmental Disorders*, vol. 46, no. 7, pp. 2464–2479, Jul 2016. [Online]. Available: https://doi.org/10.1007/s10803-016-2782-9

[14] J. A. Ungerer and M. Sigman, "Symbolic play and language comprehension in autistic children," *Journal of the American Academy of Child Psychiatry*, vol. 20, no. 2, pp. 318–337, 1981.

[15] J. Amato Jr, M. Barrow, and R. Domingo, "Symbolic play behavior in very young verbal and nonverbal children with autism." *Infant-Toddler Intervention: The Transdisciplinary Journal*, vol. 9, no. 2, pp. 185–94, 1999.

[16] K. Toth, J. Munson, A. N. Meltzoff, and G. Dawson, "Early predictors of communication development in young children with autism spectrum disorder: Joint attention, imitation, and toy play," *Journal of autism and developmental disorders*, vol. 36, no. 8, pp. 993–1005, 2006.

[17] R. M. Jones, A. Southerland, A. Hamo, C. Carberry, C. Bridges, S. Nay, E. Stubbs, E. Komarow, C. Washington, J. M. Rehg, C. Lord, and A. Rozga, "Increased eye contact during conversation compared to play in children with autism," *Journal of Autism and Developmental Disorders*, vol. 47, no. 3, pp. 607–614, Mar 2017.

[18] D. Bone, C. Lee, T. Chaspari, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand," in *ProcInterspeech 2013, Lyon, France, August 25-29, 2013*, 2013, pp. 2400–2404.

[19] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. S. Narayanan, "Acoustic-prosodic and turn-taking features in interactions with children with neurodevelopmental disorders," in *ProcInterspeech 2016, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 1185–1189.

[20] P. Boersma. Praat: doing phonetics by computer. [Online]. Available: http://www.praat.org/

[21] K. Gotham, A. Pickles, and C. Lord, "Standardizing ados scores for a measure of severity in autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 39, no. 5, pp. 693–705, 2009.

[22] J. Freguson, "Variable duration models for speech," in *Proc. Symposium on the Application of Hidden Markov Models to Text and Speech, 1980*, 1980.

[23] S.-Z. Yu, "Hidden semi-markov models," *Artificial intelligence*, vol. 174, no. 2, pp. 215–243, 2010.

[24] S. E. Levinson, M. Y. Liberman, A. Ljolje, and L. Miller, "Speaker independent phonetic transcription of fluent speech for large vocabulary speech recognition," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 441–444.

[25] S. Levinson, A. Ljolje, and L. Miller, "Large vocabulary speech recognition using a hidden markov model for acoustic/phonetic classification," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 505–508.

[26] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 61–74.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.