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

Manoj Kumar¹, Pooja Chebolu¹, So Hyun Kim², Kassandra Martinez², Catherine Lord³, Shrikanth Narayanan¹

¹University of Southern California, Los Angeles, United States
²Department 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
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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Trend during ‘play’ w.r.t ‘talk’</th>
<th>Child</th>
<th>Adult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking Fraction (%)</td>
<td>Lower</td>
<td>Lower</td>
<td></td>
</tr>
<tr>
<td>Intensity ($\sigma$)</td>
<td>(Higher, Higher, n.s)</td>
<td>(Higher, Lower)</td>
<td></td>
</tr>
</tbody>
</table>

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 posterior 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 $s_1:T$, the problem of computing the optimum underlying state sequence $s_1:T$ for an HMM can be efficiently solved using the Viterbi algorithm.

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 $\lambda = (\pi, a, b, \delta)$, where

$$ a_{ij}(t) = P[s_{t+1} = j|s_{t} = i, \sigma_t] $$

$$ b_{ij}(v_k) = P[d_k|s_{t} = i, \sigma_t] $$

$$ \delta_{i,j}(t) = \max_{s_{t-d}^{1:t-1}} \{ \pi_{i,j}(t) + b_{ij}(d_k) \} $$

The previous state selected by $\delta_{i,j}(t)$ is recorded using $\Psi_t(j, d) = (t - d, i^*, d^*)$ where $i^*$ is the previous state, $d^*$ is the previous duration.
is its duration and \(t - d\) its ending time. Note that \(i^*\) and \(d^*\) are obtained as the solutions for (2).

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.

**Algorithm 1: Estimating optimum State Sequence for EDHMM**

**Inputs:**
- \(a_{ij} \forall i,j \in S\) \(\Rightarrow\) transition probability
- \(b_j(\cdot) \forall j \in S, \forall t \in [1, T]\) \(\Rightarrow\) emission probability
- \(p_j(d) \forall j \in S, \forall d \in [1, T]\) \(\Rightarrow\) durational distribution

**Output:**
- \(\Psi_t(\cdot)\) \(\Rightarrow\) state sequence

**Initialize:**
1. \(\delta_1(j, d) = -\infty \forall t, d \in [1, T], \forall j \in S\)
2. \(\Psi_1(j, d) = (0, 0) \forall t, d \in [1, T], \forall j \in S\)

**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**

\(i' \leftarrow j - 1\)

\(Q_{i',d'} \leftarrow \delta_{i'-d'}(i, d') + \log a_{ij} + \log p_j(d) + \sum_{k=1}^{t-1} \log b_j(o_k)\)

// 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

17. \(j^*, d^* = \text{argmax}_{j,d}(\delta_T(j, d))\); // Backtracking
18. \(t_{prev}, j_{prev}, d_{prev} \leftarrow (T, j^*, d^*)\)
19. \(x \leftarrow (t_{prev}, j_{prev}, d_{prev})\)
20. **while** \(t_{prev} > 0\) **do**
21. \(x = x \cup \Psi_{t_{prev}}(j_{prev}, d_{prev})\)
22. \((t_{prev}, j_{prev}, d_{prev}) \leftarrow \Psi_{t_{prev}}(j_{prev}, d_{prev})\)
23. **end while**

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 irrele-
vant 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.](image)

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.

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

### 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 child’s 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


