Quantitative Analysis of Pitch in Speech of Children with Neurodevelopmental Disorders

Géza Kiss, Jan P.H. van Santen, Emily Prud’hommeaux, Lois M. Black

Center for Spoken Language Understanding
Oregon Health and Science University, Beaverton, Oregon
geza.kiss@cslu.ogi.edu, vansanten@cslu.ogi.edu, emilypx@gmail.com

Abstract

We analyzed the prosody of children with Autism Spectrum Disorder, Developmental Language Disorder, and typical development in conversational speech, using the CSLU ADOS speech corpus. We found several significant differences in the pitch characteristics of these diagnostic groups, and report automatic classification utilizing these features that are well above chance level. We show that the choice of pitch tracker, its parameters, and the pitch correction method can substantially affect the results, thus the scientific relevance of studies on prosody, and may be one of the reasons for conflicting findings.

Index Terms: prosody, pitch, F0, disorder, autism, language impairment, ASD, HFA, ALI, ALN, SLI, DLD

1. Introduction

The CDC (Center for Disease Controls and Prevention) estimates the prevalence of Autism Spectrum Disorder (ASD) to be 1 in 88 children in the US [1], using the current (DSM-IV [2]) diagnostic criteria. This is an extremely high number, given the difficulties associated with the condition, which include impaired social interaction, social communication, and restricted or repetitive behaviors. Since early therapeutic intervention is effective in helping children with this condition [3], early diagnosis is critical in improving their quality of life.

Neural processing in general and auditory processing in particular is atypical in ASD [4]. This manifests itself in atypical prosody, which is mentioned even in the earliest description of ASD [5]. Nevertheless, it is not a part of diagnostic procedures, at least in part because of reliability and validity issues with its subjective evaluation.

The goals of this research work are to characterize more accurately the speech prosody phenotype in ASD and to find features that can be used in automatic screening procedures. We focused on one aspect of prosody: intonation (pitch variation) in conversational speech of children with ASD, trying to find features that are reliable signs of the disorder. Here we do not deal with other aspects of prosody, such as rhythm, intensity, or pauses [6]. This work may contribute to the better understanding and sub-grouping of ASD, thus leveraging its research in other areas, too.

We discuss earlier work on the topic of prosody in ASD, and then delve into some of the challenges of using pitch information. We report statistical properties of pitch information that are significantly different between groups of children with Autism Spectrum Disorder (ASD), Specific Language Impairment (SLI), and typical development (TD). We also evaluate the features found to be significantly different in a classification framework.

2. Previous Work

A 2003 review paper by McCann and Peppé [7] concluded “that prosody in autism spectrum disorders is an under-researched area and that where research has been undertaken, findings often conflict”. Since then, several studies elaborated on the topic, but some findings still conflict, and the search for features with high sensitivity and specificity needs to continue. Two relatively recent papers analyze intonation quantitatively; we summarize them briefly, as we replicated their measurements with our data set.

Sharda et al. (2010) [8] initiated conversations with 15 children with ASD and 10 children with TD between the ages of 4 and 10 years, using pictures. They also acquired speech from mothers of typically developing infants, asking them to speak about the pictures as if they were talking to their infants. In this way, they acquired about 80 seconds of speech per subject, segmented it, and extracted pitch information using the cross-correlation algorithm in Praat [9]. They do not specify the parameters they used, so they presumably used the default ones, with a pitch ceiling of 600 Hz. They found significant differences between the ASD and TD groups in the means of the per-utterance mean, range, and pitch excursion measures. Moreover, they found that the children with ASD did not differ significantly on these measures from the motherese (i.e. child-oriented speech) samples.

Bonneh et al. (2011) [10] worked with 41 children with ASD and 42 aged-matched children with TD, who ranged in age from 4 to 6.5 and were native speakers of Hebrew. Their speech elicitation task was naming of pictures. They recorded 60 seconds or more of speech per child. They did not evaluate the pitch on a per-utterance level but used the entirety of the recording to calculate statistical properties of pitch, as well as spectral properties using Long Term Average Spectrum (LTAS). On the segmental level, they found fewer words per minute for the children with ASD as well as longer words (slower speaking rate). They extracted pitch information using the VoiceBox speech processing toolbox [11]. They do not specify the parameters used; based on their figures, the upper pitch limit may have been 400 Hz. They found a larger pitch range and pitch standard deviation (SD), as well as much shallower pitch histogram peak in the ASD group, but no significant difference between the means.

As we can see, different studies use different kinds of speech materials (conversation, sequence of words) and subjects from different (sometimes quite broad) age ranges. In addition, the subject groups are not well characterized, as the authors provide the size of the diagnostic groups but do not specify the criteria they used to reach the diagnosis. Therefore, it is of interest to examine how well these findings generalize to other...
data sets. We also wanted to find other possibly more robust measures.

3. CSLU Corpus

We use the CSLU Autism Speech Corpus, which is a relatively large and very well characterized dataset. The group with ASD is divided into subgroups with and without language impairment, and there are two control groups: one with TD and one with SLI. This can help us to separate the differences that autism causes per se, and the effect of language impairment. In this way, we can address a major obstacle in autism research: the heterogeneity of the ASD population.

The CSLU Autism Speech Corpus was collected at CSLU between 2005 and 2012, in the course of a large NIH-supported project on expressive and receptive prosody in autism. It contains video and audio data, and manual transcriptions using the SALT notation [12] from 146 children with a firm diagnosis, aged 4 to 9, who are native monolingual speakers of American English. For this study, we analyzed the audio recordings of the Autism Diagnostic Observation Schedule (ADOS) [13], a diagnostic instrument consisting of a series of semi-structured activities, including conversation, play, a picture description, and narrating a wordless picture book. The speech is divided into utterances (so-called communication units: “an independent clause with its modifiers” [14]). We did not use utterances that overlapped with the speech of the clinician, those containing noise (e.g. toy sounds, as determined based on comments from the transcribers), and those shorter than 1 second or with very short voiced part (shorter than 0.5 seconds). See Table 1 for a numeric characterization of the subset of the data we used for this study, reflecting an earlier state of the corpus.

The diagnoses were determined in extensive clinical consensus meetings in accordance with the DSM-IV [2] criteria and the published cut-off scores on the ADOS [13] and either the ADI-R (Autism Diagnostic Interview, Revised [15]) or the SCQ (Social Communication Questionnaire [16]). The children’s abilities were also assessed on several neurocognitive measures, using language tests, IQ tests (both verbal and non-verbal IQ), and motor tests. The EpiSLI criteria [17] were used for diagnosing SLI.

The children in this study all have a firm diagnosis of ASD, SLI, or TD. Some of those with ASD also met the criteria for SLI; we denote this combined diagnosis as ALN (autism and normal language). We compared pairs of groups on appropriate measures; see Table 2. We used an iterative post-hoc matching algorithm. In each iteration, we drop the subject that is most different from the rest (using the p-value from MANOVA tests), until there is no longer a significant difference. Next, we turn to the details of how we analyzed the speech in terms of pitch.

Table 1: CSLU ADOS Corpus, the subset we used.

<table>
<thead>
<tr>
<th>diagnostic group</th>
<th>n</th>
<th>age of subjects (mean, range)</th>
<th>amount of speech (mean, range; in sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALN</td>
<td>14</td>
<td>6.4 (4.7 – 8.2)</td>
<td>419 (164 – 980)</td>
</tr>
<tr>
<td>ALI</td>
<td>25</td>
<td>6.4 (4.0 – 8.8)</td>
<td>405 ( 45 – 1191)</td>
</tr>
<tr>
<td>SLI</td>
<td>14</td>
<td>6.4 (4.2 – 8.2)</td>
<td>451 (111 – 1167)</td>
</tr>
<tr>
<td>TD</td>
<td>28</td>
<td>6.0 (4.0 – 8.5)</td>
<td>524 (228 – 938)</td>
</tr>
</tbody>
</table>

4. Statistical analysis of pitch

4.1. Pitch features

We extracted the pitch information using the Snack toolkit [18] ESPS method, a frame shift of 10 ms, and a frame size of 7.5 ms, with 100 to 600 Hz as the allowed pitch range. We analyzed the pitch information for each subject in two ways: first, by determining properties of the pitch distribution of all the available speech, and second, by determining the distribution of per-utterance pitch properties. We did the first by calculating statistics from all the available pitch data for the subject. In essence, these capture properties of the per-child pitch histogram. See Figure 1 for a pairwise comparison of these global pitch histograms. Apparently, the difference between the TD group and the other groups was much larger than between any pair of the latter. Our numeric features captured this difference, as we shall show later. We then calculated the per-utterance statistics, and quantified the distribution of these. We also calculated the features that Sharda [8] and Bonneh [10] showed to be significantly different between the ASD and TD groups.

Figure 1: Global pitch histograms for each comparison

4.2. Identifying group differences

We compared the results of several statistical testing procedures to ensure that the differences we found between the diagnostic groups are in fact significant. We used two-sample t-tests on the features one by one, and then used FDR (False Discover-
ery Rate correction) for a feature set to eliminate discoveries due to some features being different by chance. Consequently, we did not correct the values when replicating earlier findings, as we use a different dataset to evaluate those. We also used 50-50 MANOVA [19], which has several advantages compared to MANOVA. In particular, it does not break down when the features are highly correlated, and the implementation we used (the ffmanova R package) contains Monte Carlo simulations for deriving adjusted p-values for the features that it found to be significantly different. Third, we check our results by using a Monte Carlo simulation. We calculated the ratio of the difference between the group means and the difference between the mean within group differences, both for the actual group memberships and for 1000 randomly permuted group memberships. If the value for the actual group memberships was lower than the randomly permuted ones less than 5% of the time, then we can consider the feature to be significantly different at the \( p = 0.05 \) level. For all our tests, we used the \( p = 0.05 \) level. We report here the results that each of the tests confirmed. (We note that the tests were often in complete agreement with each other regarding which features are different.)

4.3. Significant pitch differences between the groups

First we calculated several statistical features related to different moments, and calculated their means to get four features: mean and median for 1st moment (location); SD (standard deviation), MAD (median absolute deviation from the median), and IQR (interquartile range) for the 2nd moment (spread), skewness for the 3rd moment (asymmetry), and kurtosis for the 4th moment (peakedness). We found that all four differed significantly for the ALN – TD comparison, whereas spread and asymmetry were significantly different (but asymmetry much less) for the SLI – TD comparison.

Second, we compared a reduced set of the original features: either the non-robust ones (mean, SD, CoV, i.e. Coefficient of Variation), together with skewness and kurtosis, or the robust versions of some of those (median, MAD, IQR). For the ALN – TD comparison, mean, skewness, and kurtosis, as well as median and MAD were significantly different. For the SLI – TD comparison, SD and MAD were significantly different, as was median but to a much lesser degree. Note that the robust versions of statistical measures (median, MAD) often distinguish the groups even when the other measures (mean, SD) do not.

Third, we compared statistics of the per-utterance features, either the basic or the robust versions. For ALN – TD, CoV of SD and of MAD, median of median and of MAD were significantly different. For SLI – TD, MAD of mean, of median, and of MAD, as well as median of MAD were significantly different. These are interpretable as a tendency of the global pitch statistics (shown earlier, including larger mean and spread) to emerge even on the sentence level, as well as a larger variation in the per-utterance spread (MAD of MAD). Although the difference in these variables was large, noise and variance in the data can obviously affect such combined statistics, from a numeros set. Therefore, we need further research to confirm if these features reflect robust differences between the diagnostic groups, and how sentence-curve properties that are easier to interpret affect these measures. Note that since we used a conservative significance test, when a feature is significantly different for only one of the ALN – TD and SLI – TD comparisons, it does not mean that ALN and SLI are significantly different on that measure. Moreover, we did not find any significant differences on any of the pitch feature sets for the SLI – ALI and ALN – ALI group comparisons.

Sharda [8] reported significant pitch differences on their data on the mean of per-utterance mean, range, and pitch excursion. We found two of the three (mean of mean and range) differences for the ALN – TD comparison, and all three for the SLI – TD comparison.

Bonneh [10] reported significant pitch differences on their data on SD, range, and the logarithm of the highest peak of the pitch histogram (log pitch peak), but no significant difference in mean. On our data, we found the opposite for all but one measure: log pitch peak, which we also found to be significantly different. However, we did find MAD, a robust measure of spread, to be significantly different. We shall discuss potential reasons for this difference in the Discussion section.

5. Classification results

We compared the performance of the features that we identified as significantly different in a simple classification scheme. We used all the features with a Naïve Bayes classifier, using the Weka Machine Learning Toolkit [20] with the default settings. We assessed the performance using leave-one-out cross-validation. We got accuracy measures that are significantly better than chance and at least as good using the per-utterance features as the global features; see the results in Table 3. The setup could be improved, for example, by using feature selection to identify the best set of features, using more sophisticated machine learning algorithms, tuning their parameters, and evaluating them using leave-pair-out cross-validation; this was, however, not the focus of this research. We can therefore regard these as a lower bound on the performance of these features.

Table 3: Diagnostic classification results.

<table>
<thead>
<tr>
<th>DX group</th>
<th>Features</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALN – TD</td>
<td>median, MAD, mean, kurtosis, skewness</td>
<td>accuracy: 74%, precision: 57%, recall: 86%</td>
</tr>
<tr>
<td></td>
<td>CoV-SD, median-MAD, median-median, CoV-MAD</td>
<td>accuracy: 74%, precision: 60%, recall: 64%</td>
</tr>
<tr>
<td>SLI – TD</td>
<td>MAD, SD, median</td>
<td>accuracy: 69%, precision: 54%, recall: 50%</td>
</tr>
<tr>
<td></td>
<td>MAD-MAD, MAD-mean, median-MAD</td>
<td>accuracy: 71%, precision: 56%, recall: 64%</td>
</tr>
</tbody>
</table>

6. Discussion

We found very similar tendencies to those reported by Sharda [8] and Bonneh [10], but could not replicate all of their findings at the same significance levels, especially those of Bonneh. The greatest difference between our findings is that we found a significant mean difference despite having fewer subjects. The language difference is unlikely to be the cause, since we are assessing between-group differences. Part of the reason may be that our dataset is different from Bonneh’s in that ours contains conversational speech, while theirs contains word lists. The lack of information on the co-morbidity of SLI for their subjects may...
also explain the differences to some extent, as we did not find a significant mean difference for the SLI – TD comparison. However, we believe that one cause of the difference may be the use of a different pitch analysis method.

When we started this work, we used a lower pitch ceiling, and then found out that it often resulted in pitch segments with halved values. For a pitch ceiling of 450 Hz, we did not find significant differences in mean pitch or skewness, but found larger differences in the peakedness (similarly to the result of Bonneh et al. [10]), manifested in highly significant differences in kurtosis values. Afterwards, we raised the pitch ceiling to 600 Hz, used an automatic method to correct the tracker errors, and verified through listening experiments that the changes improved the reliability of the pitch curves; we report results for this setup. We conversely found significant mean and skewness difference, and less significant kurtosis difference.

We believe that for any findings to be reproducible, at least on the same dataset, studies making use of pitch information need to report the pitch tracker algorithm and its settings. Moreover, whenever possible, researchers should make efforts to ensure that the automatically extracted pitch curves reflect the perceived pitch, so that what we quantify as differences in pitch do not later turn out to be differences in vocal quality, in noise conditions, or trivial consequences of the pitch tracker settings.

We did not find significant differences between the ASD and SLI subgroups. Although the number of subjects in our study was limited, this shows that we must be careful when interpreting findings where only a TD control group is used. Some of the atypical features may not be specific to autism but instead may be part of the phenotype of several disorders, may indicate comorbidity, or may be a characteristic of a subtype of ASD.

7. Conclusions

We presented a study on the atypicality of pitch in neurodevelopmental disorders, namely ASD and SLI, on the CSLU ADOS Corpus, a relatively large and very well characterized database. We evaluated features that earlier studies used, as well as other features, namely robust measures of location and spread, and higher-order statistics. We found significant differences compared to the TD control group in the global pitch distribution, as well as statistics of sentence-level features, but did not find any significant differences between the ASD and SLI subgroups. These features allow for automatic classification well above chance level, even with a simple classification scheme. We discussed evidence that the way pitch information is extracted from the speech can have a considerable impact on the findings, which calls for care and detailed documentation of this step in research work dealing with pitch information.

We plan to relate these features to the available neurocognitive measures, to include other aspect of prosody such as loudness and rhythm, and to identify properties of utterance curves that distinguish the diagnostic group, including identifying the relationship between per-utterance statistical features and utterance curve forms.

8. Acknowledgements

This research was supported by grants NIH-1R01DC007129 (van Santen, PI), NSF IIS-0905095 (van Santen, PI) and NIH-21DC010239 (Black, PI) The views herein are those of the authors and do not necessarily reflect the views of the above funding agencies. Lois Black passed away in October 2011.

9. References


