34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Objective Child Behavior Measurement with Naturalistic Daylong Audio Recording and its Application to Autism Identification*

Dongxin Xu, Jill Gilkerson, Jeffrey A. Richards *

Abstract— **Child behavior in the natural environment is a subject that is relevant for many areas of social science and bio-behavioral research. However, its measurement is currently based mainly on subjective approaches such as parent questionnaires or clinical observation. This study demonstrates an objective and unobtrusive child vocal behavior measurement and monitoring approach using daylong audio recordings of children in the natural home environment. Our previous research has shown significant performance in childhood autism identification. However, there remains the question of why it works. In the previous study, the focus was more on the overall performance and data-driven modeling without regard to the meaning of underlying features. Even if a high risk of autism is predicted, specific information about child behavior that could contribute to the automated categorization was not further explored. This study attempts to clarify this issue by exploring the details of underlying features and uncovering additional behavioral information buried within the audio streams. It was found that much child vocal behavior can be measured automatically by applying signal processing and pattern recognition technologies to daylong audio recordings. By combining many such features, the model achieves an overall autism identification accuracy of 94% (N=226). Similar to many emerging non-invasive and telemonitoring technologies in health care, this approach is believed to have great potential in child development research, clinical practice and parenting.**

I. Introduction

Objective data about child's behavior, their environment and how they interact with the environment are critical for many areas related to child development, e.g., child development research, early identification of developmental disorders, treatment monitoring, home based monitoring, etc. Audio recordings contain rich information about a child's vocal behavior, social-emotional interaction, verbal communication, articulatory motor patterns and his/her environment. There are advantages of audio signals over other signals in terms of convenience, required physical conditions (such as light for visual signals) and so on. This study and the proposed hardware-software framework uses automated analysis of unobtrusive daylong audio recordings obtained from a child's natural home environment. This method allows for a novel type of information extraction for the measurement of child behavior, targeting a new audio-based approach which is objective, naturalistic, scalable, automatic and convenient. A large number of audio samples can be collected relatively easily, resulting in stable, reliable and accurate macro-statistics for characterizing children's behavior and their environments, ultimately informing research, clinical practice and parenting.

A lightweight digital recorder is worn by a child for a whole day to collect his/her vocal output and the sounds in the environment. Signal processing and pattern recognition technologies are used to automatically detect different sound segments, including key-child (who wears the recorder), other-child, adult-male, adult-female, overlapped-sounds, noise, TV and electronic media sounds, and silence, producing a sequence of segment labels. Key-child segments can be further processed (e.g., with a phone recognizer) to produce information about a child's phonetic behavior at a macro-level using daylong recordings [1,2,3,4]. Unsupervised approaches can also be explored. For instance, all key-child vocalizations can be clustered into self-organized data categories, providing certain unsupervised categorization information about child vocalizations.

Our previous research has focused on data-driven modeling and the overall performance of childhood autism probability estimation using the frequency features of phone-level sound categories of child vocalizations (child vocalization composition), and achieved good performance overall [1,2,3,4]. However, there remains the question of why it works. Indeed, this research was intended as a simple proof of concept, so the meaning and details of each underlying feature of child vocal behavior could not be addressed individually. The modern machine learning and data-driven approaches can sometimes work as a black box, whereby it is often unnecessary to know the details of the input data. For this particular case, this can result in a situation in which a child is reported to be at high risk for autism, but without elucidation of the potential underlying behavior or developmental problems that contributed to the categorization. To rectify this situation, the current study attempts to explore the details of the features mentioned above and at the same time to search for more behavioral information buried in the audio streams of a large number of daylong recordings. It is also necessary to point out that behavioral features can be discovered either by starting from top-down approaches such as theory-based ones or starting from bottom-up approaches such as data-driven ones.

This exploratory research shows that many child vocal behaviors can be automatically and objectively measured by applying signal processing and pattern recognition technologies to a large number of daylong audio recordings, including phone-level sound category features, phonetic

*Research supported by the LENA Research Foundation.

Dongxin Xu is with the LENA Research Foundation and is adjunct faculty of University of Colorado, Boulder, CO 80301 USA. (phone: 303-441-9012; fax: 303-545-2166; e-mail: dongxinxu@lenafoundation.org).

Jill Gilkerson is with the LENA Research Foundation and is adjunct faculty of University of Colorado, Boulder, CO 80301 USA (e-mail: jillgilkerson@lenafoundation.org).

Jeffrey Richards is with the LENA Research Foundation, Boulder, CO 80301 USA (e-mail: jeffrichards@lenafoundation.org).

development features, unsupervised self-organized sound category features, sound-sequence features, features about interactive behavior with environments, prosodic features and spectrum features of child vocalization and so on. This study demonstrates age-related child development trends of these behavioral features. Moreover, group differences of these behavior features among different child diagnostic groups are also demonstrated (i.e., children of typical development-TD, children with language delay but not autism-LD, and children with autism-ASD). The overall autism risk estimation with all these features is also studied. The new approach achieves 94% accuracy at equal-error rate points on a data set with 226 children.

II. STUDY SAMPLE

This study included N=106 TD children, N=49 LD children and N=71 ASD children. The ages of TD children were 8-to-48 months but mainly above 12 months; LD children were 10-to-44 months but mainly 14-to-41 months; ASD children are 16-to-48 months but mainly 25-to-48 months. There were a total of 1363 naturalistic daylong audio recordings with 802 for TD children, 333 for LD children and 228 for ASD children. No recording was under 9 hours with about 99% being 16-hour recordings. The details about how participants were recruited and demographic and other characteristics such as ethnicity, human assessment scores of PLS-4, REEL, CDI and CBCL can be found in [4]. One difference is that in the original study the ASD group included recordings with therapy time which were removed from this study to avoid any potential confounding effects.

III. CHILD BEHAVIORAL FEATURES & MEASUREMENT

As mentioned above, daylong audio recordings were processed to generate the sequence of sound segments, and key-child segments were further processed with a phone recognizer or other approaches. The child behavior-related features can be extracted either from sound segment sequences (interaction effects) or directly from child vocalizations (phone-level sound categories). We discuss a sampling of these features in the following to demonstrate child development trends and group differences. Correlations with age and Welch two sample t-tests (2-sided) were used to show respectively developmental trends and group differences. Each age-month in the following graphs represents a range of ±5 months within which recordings from one child were averaged; mean and standard errors were estimated based on child-level averages. Separately, correlations were estimated based on recordings with weights such that each child has a total weight of 1 from all his/her recordings. To compare overall group differences without the interference of feature variation across ages, the mean value of TD recordings within each age-month was removed from all recordings of the same month.

A. *The percentage frequency of phone-level categories*

Key-child segments were processed with the open source Sphinx phone recognizer [1,2] to generate 46 phone-level sound categories, including both speech and non-speech sounds based on acoustic similarity. Figure 1 shows the frequency of consonant-like sounds in child vocalizations. Children in all three groups produced more and more consonant-like sounds as they aged. Relatively, TD children developed most rapidly and ASD children least rapidly. Broadly, consonant production involves blocking and releasing air flows, typically requires finer motor control and thus can be considered a more advanced articulatory skill than vowel production. ASD children have been shown to develop language more slowly, and this may be reflected in the slower consonant production. Both correlations with chronological age and overall group differences are statistically significant.

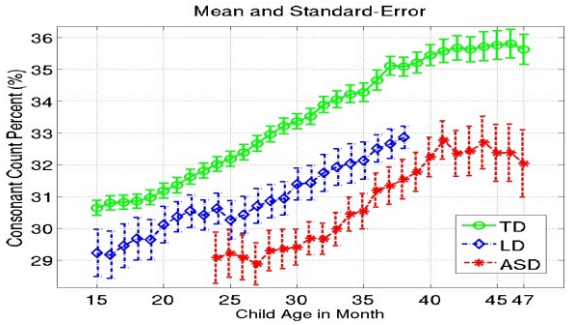

Figure 1. The frequency (percentage) of consonant-like sound in child vocalization. Horizontal axis is child age in month. Mean and standard error were estimated for each age with a range of ±5 months. Separately, correlations with age were estimated and Welch two sample t-tests (two-sided) were conducted. This same approach was applied for all figures. Abbreviation: TD – Typical Development, LD – Language Delay but not Autism, ASD – Autism. Correlation with age: TD=0.67***; LD=0.42**; ASD=0.32**. Group differences t-tests: TD-LD: t(68)=5.52***; TD-ASD: t(90)=7.95***; LD-ASD: t(118)=2.62**. All correlations and group differences for this feature are statistically significant ⊗

Figure 2 depicts the frequency of non-speech sounds in child vocalizations. As shown, ASD children tended to produce more non-speech sounds than children in the other groups. The three groups also differed significantly on this feature. Contrary to consonant-like sounds, all three groups tended to produce fewer non-speech sounds with age, though at different levels (and possibly slopes).

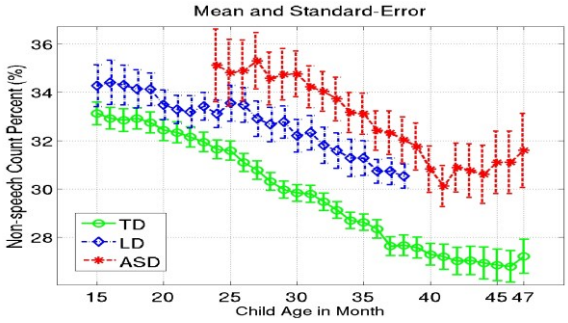

Figure 2. The frequency (percentage) of non-speech sound in child vocalization. Correlation with age: TD = -0.57***; LD = -0.33*; ASD = -0.27*. Group differences t-tests: TD-LD: t(79) = -4.59***; TD-ASD: t(93) = -5.87***; LD-ASD: t(116) = -2.07*.

B. *Interaction with the environment*

The temporal sequence of sound segments for key-child and environment sound categories should provide unique

⊗ p-value notation for this paper: *: p<0.05; **: p<0.01; ***: p<0.001

information about how a child interacts with the environment. Here is one example using such features. Under naturalistic conditions, it is inevitable that a speaker's voice will overlap (or collide) with those of other speakers and with environmental sounds. We observed that the vocalizations of young children frequently collide with other sounds in the environment, which may be related to their coordination capabilities or general sensitivity to other sounds in their environment. The probability of a child's vocalization colliding with external environment sounds may be used to quantify such coordination capability. More specifically, we refer here to the conditional probability that, given a current key-child segment, the preceding or subsequent sound segment is an overlapped sound. In Figure 3 we see that the vocal activity of ASD children demonstrates a significantly higher rate of these environmental collisions, suggesting less well-developed coordination capability with the environment. Furthermore, no significant differences between TD and LD children are evident, indicating the specificity of this feature to ASD children. As well, TD and LD children show no changes with age, whereas ASD children demonstrate a downward trend. Although our sample is relatively sparse at the younger ages, this pattern does suggest that this feature could potentially enhance early identification or risk estimation of autism. However, much more data is needed to draw any solid conclusions.

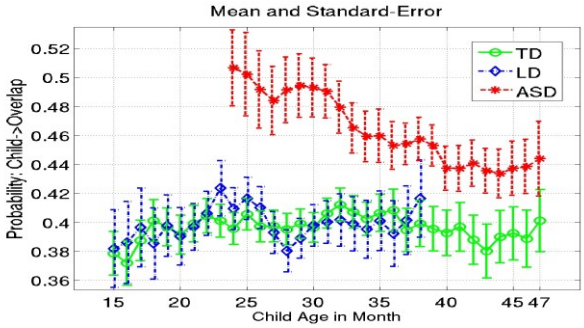

Figure 3. The probability of child vocalization "colliding" with other sounds. Correlation with age: TD = 0.002; LD = 0.126; ASD = -254*. Group differences t-tests: TD-LD: t(90) = -0.13; TD-ASD: t(132) = -3.66***; LD-ASD: t(111) = -2.94**.

C. Categorization of Child Overlapped Sounds

Given that child vocal production frequently collides with other types of sounds, it may be asked whether there exist meaningful differences across types of overlapped sound. Similar to child vocalizations, child overlapped sounds can be processed and categorized using a phone recognizer or other approaches and the frequency features can be similarly obtained. Figure 4 shows one such result, the frequency of [p]-like sounds in child overlapped segments. Note that significant differences between ASD children and TD/LD children are observed in the graph, though we make no particular interpretation of these specific patterns here.

D. Features Based on Mathematical Approaches

From a data-driven perspective, child vocalization categorization can be based on unsupervised self-organization techniques such as K-means algorithms. Figure 5 depicts the frequency of one such cluster derived directly from child vocalization data. Figure 6 shows a feature of principal component analysis for bi-phone sequences. It is more difficult to interpret the physical meaning of such features. However, these features, based on bottom-up data-driven approaches, do provide different perspectives when compared with top-down theory-based ones, even though there is some overlap in the discriminant information provided. It is possible to trace down the related physical meanings of such features generated via mathematical transformations.

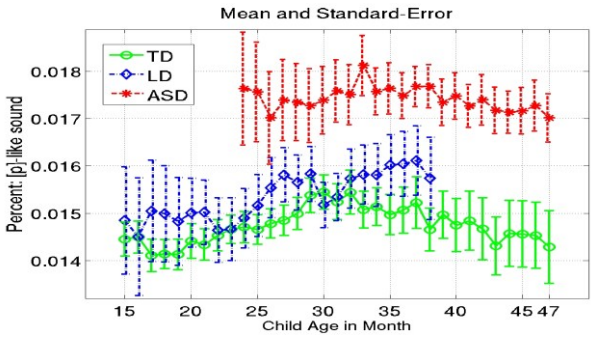

Figure 4. One example of environment sounds a child may be more likely to collide with: frequency (percentage) of "[p]-like" sound in child overlapped sound. Correlation with age: TD = 0.065; LD = 0.076; ASD = -0.033. Group differences t-tests: TD-LD: t(75) = -0.93; TD-ASD: t(129) = -5.20***; LD-ASD: t(100) = -3.04**.

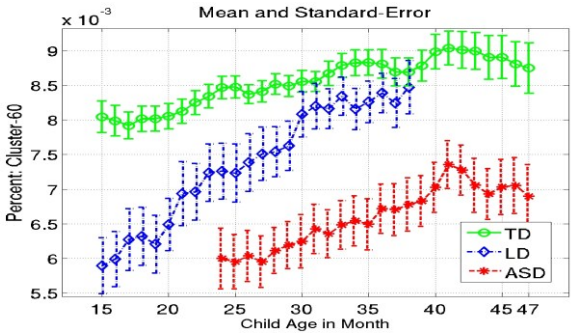

Figure 5. An unsupervised self-organized child vocalization cluster obtained with K-means algorithm. Correlation with age: TD = 0.287**; LD = 0.311*; ASD = 0.197. Group differences t-tests: TD-LD: t(70) = 3.29**; TD-ASD: t(110) = 8.09***; LD-ASD: t(107) = 3.30**.

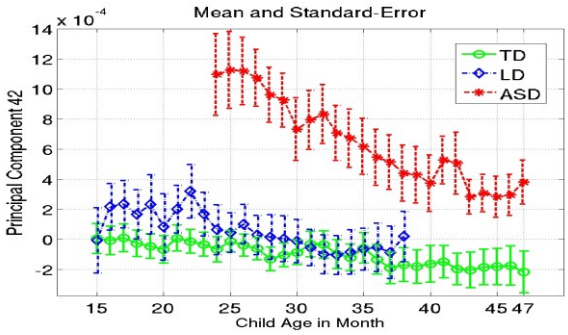

Figure 6. A principal component of bi-phone sequence (phone-level sounds include non-speech-like and consonants-like and vowel-like sounds). Correlation with age: TD = -0.039; LD = -0.067; ASD = -0.279*. Group differences t-tests: TD-LD: t(93) = -1.54; TD-ASD: t(109) = -6.47***; LD-ASD: t(118) = -4.62***.

E. Prosodic Features of Child Vocalization

Prosodic features may include the statistics of duration, volume, pitch (f0), pauses, etc. in recordings. Here we include examples for duration and volume. Figure 7 shows the variation of the duration of child consonant-like sounds within each recording. As we know, consonant duration is very stable due to the nature of consonant production (e.g., plosive consonants). Thus, a degree of variation may reflect physical control of the production of consonant-like sounds. As shown, this variation reduces with age for TD/LD children, but ASD children do not show as much variation reduction (no significant negative correlation with age for ASD while both TD and LD have significant negative correlation with age), possibly indicating different control capability.

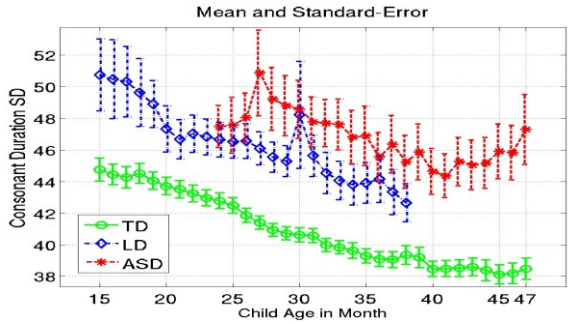

Figure 7. The duration variation (standard deviation) of consonant-like sound in child vocalization. Correlation with age: TD = -0.44***; LD = -0.344**; ASD = -0.082. Group differences t-tests: TD-LD: t(67) = -5.46***; TD-ASD: t(86) = --6.59***; LD-ASD: t(118) = -1.83.

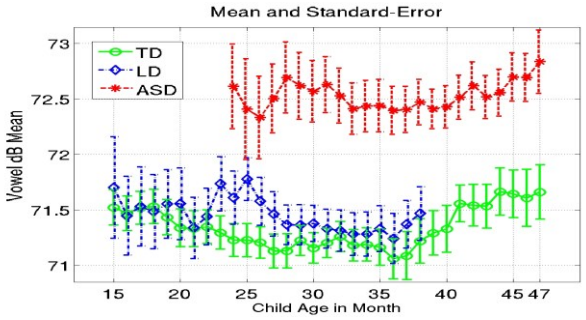

Figure 8. Volume of vowel-like sound in child vocalization. This is the feature of mean value of dB-level in each recording. Correlation with age: TD = -0.046; LD = -0.021; ASD = 0.045. Group differences t-tests: TD-LD: t(97) = -0.45; TD-ASD: t(125) = -5.84***; LD-ASD: t(117) = -4.78***.

Figure 8 tracks the volume of vowel-like sounds with age. As seen, ASD children have significantly higher dB levels for vowel-like sounds than TD/LD children, though none of the groups show clear age effects. We do not know the reason for this pattern, but it may be due to differential attention effects and relationship to the environment. Other spectrum features such as spectrum tilt and spectrum entropy of child vocalizations have also been studied and will be reported in the future.

IV. Autism Identification with All Features

As shown above, vocalization-based child behavioral features exhibit development trends and/or group differences to a greater or lesser degree. Here, we utilize these features to generate a probability for childhood autism identification or risk estimation. We combined 254 features of the above types in this experiment to predict the likelihood of autism using the Adaboost method with simple Gaussian classifier as a weak learner, which can be regarded as an approximation to logistic regression. Table 1 shows the accuracy at the equal-error-rate (EER) point all based on leave-one-child-out cross-validation. The recording level accuracies are based on the estimated autism risk for each recording. The autism risk for each child is the average of the risks of the recordings for the child. The child level accuracies are based on child risks.

TABLE I. Accuracy of Autism Identification at EER-Point

Identification Case	Recording-Level	Child-Level
ASD versus TD	94%	94%
ASD versus LD	86%	89%
ASD versus TD + LD	92%	94%

V. Conclusion and Discussion

This study demonstrates the value and potential of objective child vocal behavior measurement using naturalistic daylong audio recordings. We explored representative features and attempted a preliminary interpretation of each feature to clarify what types of behavior or potential child development issues might be related to these features. Child development trends and group differences were exhibited across various features, indicating the effectiveness of the objective measures. The application of these objective features to childhood autism identification achieved a practical 94% accuracy. The details of certain topics such as the accuracy as a function of child age, the identification of LD versus TD and more cross-validations will be reported in the future. Similar to many emerging non-invasive and telemonitoring technologies in health care, the approach demonstrated in this study is believed to have great potential in child development research, clinical practice and parenting.

Acknowledgment

We greatly acknowledge Terrance Paul for his conception of the LENA System and for personally funding and directing its development as well as the development of the LENA Research Foundation Natural Language Corpus.

References

[1] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, J. Hansen "Signal Processing for Young Child Speech Language Development" 1st Workshop on Child, Computer and Interaction, Oct. 2008, Chania, Crete, Greece.

[2] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, S. Gray "Child vocalization composition as discriminant information for automatic autism detection" International Conference of the IEEE Engineering in Medicine and Biology Society, Sept. 2-6, 2009

[3] S. Warren, J. Gilkerson, J. Richards, K.D. Oller, D. Xu, U. Yapanel, S. Gray "What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism", Journal of Autism Developmental Disorders, Nov.21, 2009

[4] K.D. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, S. Warren "Automated Vocal Analysis of naturalistic recordings from children with autism, language delay and typical development", Proceeding of the Natural Academy of Sciences of the United States of America, July 2010, 107(30), 13354-13359