Objective Child Vocal Development Measurement with Naturalistic Daylong Audio Recording

Dongxin Xu, Jill Gilkerson, Jeffery A. Richards

LEAN Research Foundation, Boulder, Colorado 80303, USA
Dongxin Xu, jillgilkerson, jeffrichard@lenafoundation.org

Abstract

Child vocal development is a subject that touches many areas. Its measurement is based mainly on subjective approaches. This study demonstrates an objective and unobtrusive measurement and monitoring approach using daylong audio recordings of the natural home environment. One purpose of this study is to explore the underlying features about child vocal development buried within the audio streams of daylong recordings using automated approaches with signal processing and pattern recognition technologies. Various child vocal development features are studied, including phonetic development features, features based on unsupervised and self-organized sound categories, features about phone sequence information, prosodic features, spectrum features and so on. We are interested in the developmental trends and the group differences of the features. Based on the developmental trends of the features, the vocalization age can be estimated. A correlation of 0.84 between the estimated vocalization age and the chronological age for children of typical development is achieved. Based on the group differences, the features can be used for childhood autism identification. 94% accuracy at equal-error rate point is achieved for this purpose. 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.

Index Terms: child vocal development, objective measurement, child phonetic development, childhood autism identification.

1. Introduction

Objective data about child vocal behavior and development 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 and development related to phonetic perception and production, language and verbal communication, social-emotional interaction and so on. There are advantages of audio signals over other signals in terms of convenience, required physical conditions (such as light for visual signals), etc. This study 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 vocal behavior and development, 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. Speech signal processing and pattern recognition technologies are used to automatically detect different sound segments, including key-child (who wears the recorder), adult, noise, silence and so on, 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 of daylong recordings [1,2,3]. Unsupervised approaches can also be explored to categorize key child vocalizations.

Our previous research has focused on data-driven modeling and the overall performance for child vocalization analysis and childhood autism identification using the frequency features of phone-level sound categories, and achieved good performance overall [1,2,3]. 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 sometime work as a black box, whereby it is sometimes 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 delayed vocal development or 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 information buried in the audio streams of a large number of daylong recordings.

This exploratory research shows that many child vocal behaviors can be automatically and objectively measured using a large number of daylong audio recordings, including phone-level sound category features, phonetic development features, unsupervised self-organized sound category features, sound-sequence features, prosodic features and spectrum features of child vocalizations and so on. This study demonstrates the age-related child developmental 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 predicted vocalization age for TD children correlates well with the actual chronological age (0.84). The overall autism identification achieves 94% accuracy at an equal-error rate point.
2. Samples of the Study
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 [3]. One difference is that in the original study the ASD group contained recordings with therapy time which were removed from this study to avoid any potential confounding effects. In addition, it is necessary to point out that the features studied in [3] were based on top-down approaches while the features explored in this study are mainly derived from bottom-up approaches.

3. Child Vocalization Features and 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, and the child vocalization features can be extracted subsequently. 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 the development trends and the group differences. Each age-month in most of the following graphs represents a range of ±5 months within which recordings from one child were averaged; mean, standard deviation or 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 the overall group differences by t-tests 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.

3.1. 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-like and non-speech-like 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 develop 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.

We found that the frequencies of different phone-like sounds in child vocalizations contain rich information about child vocal and phonetic development. Figure 2 demonstrates the landscape of the frequency features (means and standard deviations) for consonant-like, vowel-like and non-speech-like sounds in child vocalizations, providing a panoramic view of how the landscape change over child age and how different the percentages of the 3 categories could be for TD, LD and ASD children. As can be seen, the relative percentages change rapidly with age. As they grow, more and more consonant-like and vowel-like sounds and less and less non-speech-like sounds are produced by children. Relatively, TD children produce less non-speech sounds and more consonant-like and vowel-like sounds than both LD and ASD children.

We could look into more details of the phone-like sounds in child vocalizations. Figure 3 further demonstrates the relative
percentages of nasal-like sounds ([M], [N], [NG]) in child vocalizations. Again, this percentage landscape informatively demonstrates the developmental trends and the differences among TD, LD and ASD children. Relatively, [M] of 3 nasal phones is developed first as in child babbling of “Ma Ma” and takes most of the percentage, and then [N] is developed later. [NG] is developed latest. When children develop to produce more and more [NG] and [N] like sounds, the relative percentage of [M] is reduced and the relative percentage of [NG] is increased. Our data show that the relative percentage of [N]-like sounds remains relatively flat but the absolute numbers is increased with age. LD and ASD children develop slower with less [NG]-like sounds and more [M]-like sounds in their vocalizations.

Another example is the landscape of the relative percentages of voiced-stop-like sounds ([B], [D], [G]) in child vocalizations. Figure 4 shows this landscape. Among the 3 voiced-stops [B], [D] and [G], [B] is relatively easy to produce and is developed first as in child babbling of “Ba Ba”, and takes most of the percentage. Relatively, [D] and [G] are developed later. As children develop to produce more and more [D] and [G] like sounds, the relative percentage of [B] sounds is reduced and the relative percentages of [D] and [G] are increased. The slower developments of LD and ASD children are reflected by the relatively less [D] sounds, less [G] sounds and more [B] sounds in their vocalizations.

3.2. 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 sequence. 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 information provided. If necessary, it is possible to trace down the actual related physical meanings of such features generated with mathematical approaches.

3.3. Prosodic Features of Child Vocalization

Prosodic features may include the statistics of duration, loudness (dB), pitch (f0), pauses, etc. in recordings. Here we include examples for duration and volume. Figure 7 shows the average duration of child non-speech-like sounds within each recording. As shown, the average duration reduces with age significantly for all 3 groups of children. However, on average ASD children produce the longest non-speech-like sounds and the group differences are all significant. The patterns of vowel durations

Figure 3: Mean and standard deviation of the relative percentages for nasal-like sounds ([M]-like sound (green), [N]-like sound (blue) and [NG]-like sound (red)) in the vocalizations of TD, LD and ASD children.

Figure 4: Mean and standard deviation of the relative percentages for voiced-stop-like sounds ([B]-like sound (green), [D]-like sound (blue) and [G]-like sound (red)) in the vocalizations of TD, LD and ASD children.

Figure 5: Unsupervised self-organized child vocalization clusters obtained with K-means algorithm. This is the feature of frequency of sound-cluster-number-36. Correlation with age: TD = 0.516***; LD = 0.339*; ASD = 0.370**. Group differences t-tests: TD-LD: t(99) = 5.62***; TD-ASD: t(138) = 6.88***; LD-ASD: t(114) = 1.07.

Figure 6: Principal Components of bi-phone sequence (phone-level sounds include non-speech-like, consonant-like and vowel-like sounds). This is the feature of principal component of 44. Correlation with age: TD = -0.041; LD = -0.038; ASD = 0.354**. Group differences t-tests: TD-LD: t(70) = 2.60*; TD-ASD: t(92) = 2.60*; LD-ASD: t(118) = 1.41.

3.2. 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 sequence. 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 information provided. If necessary, it is possible to trace down the actual related physical meanings of such features generated with mathematical approaches.

3.3. Prosodic Features of Child Vocalization

Prosodic features may include the statistics of duration, loudness (dB), pitch (f0), pauses, etc. in recordings. Here we include examples for duration and volume. Figure 7 shows the average duration of child non-speech-like sounds within each recording. As shown, the average duration reduces with age significantly for all 3 groups of children. However, on average ASD children produce the longest non-speech-like sounds and the group differences are all significant. The patterns of vowel durations
are actually similar. This may be related to different articulatory motor control of different groups of children.

Figure 7: The average duration of non-speech-like sounds in child vocalization. Correlation with age: TD = -0.584***; LD = -0.499***; ASD = -0.238*. Group differences t-tests: TD-LD: t(71) = -5.67***; TD-ASD: t(88) = -6.68***; LD-ASD: t(116) = -2.08*.

Figure 8: Loudness of vowel-like sounds in child vocalizations. 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 loudness of vowel-like sounds with age. As can be seen, ASD children have significantly higher dB levels of 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 the 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.

Figure 9. Vocalization age estimation for TD children using leave-one-child-out cross-validation

4. Estimation of Vocalization Age

By combining the features discussed, it is possible to estimate child vocalization age as an automated and objective measure for child vocal development. One possible way is to build a linear regression model to predict the chronological age of TD children using the features. The experiment with this method shows 0.84 correlation between the estimated vocalization age and the chronological age for TD children with leave-one-child-out cross-validation. Figure 9 shows the scattering plot of the vocalization age and the chronological age for this experiment. The vocalization age model based on TD children can produces 0.55 and 0.25 correlations between the estimated vocalization age and the chronological age for LD and ASD children respectively, with under-developed vocalization ages.

5. Autism Identification

We can also utilize the features for childhood autism identification. We combined 254 features in our experiment to estimate probability for autism using the Adaboost method (an approximation to logistic regression). The 94% accuracy at the equal-error-rate point using leave-one-child-out cross-validation has been achieved.

6. Conclusions

This study demonstrates the potential of objective child vocal development measurement using naturalistic daylong audio recordings. We explored representative features and attempted a preliminary interpretation of each feature. 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 vocalization age estimation and childhood autism identification achieved good performance. 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.

7. Acknowledgements

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.

8. References