Signal Processing for Young Child Speech Language Development

Dongxin Xu, Umit Yapanel, Sharmi Gray, Jill Gilkerson, Jeff Richards, John Hansen*
Infoture Inc. 5525 Central Avenue #100, Boulder, CO 80301, USA
dongxinxu,umityapanel,sharmigray,jillgilkerson,jeffrichards@infoture.org
*CRSS: Center for Robust Speech Systems, The University of Texas at Dallas,
*800 W. Campbell Rd., Richardson, TX 75080, USA.
*john.hansen@utdallas.edu

ABSTRACT
Speech signal processing and other man-machine interaction technologies have been developed for improved child-computer interaction for education, entertainment, as well as other applications [1, 2]. However, for very young children (in the age range of 0 to 4 years old, and especially 0 to 2), such interaction is not encouraged [3, 4]. Instead, parent-child interaction is highly recommended [3, 4] since it promotes improved language development. In this study, a new system entitled LENA™ (Language Environment Analysis) and its associate processing technologies will be introduced. LENA provides parents/caregivers with quantified statistical information concerning the language environment and development status of children in order to allow for the determination of what needs to improve and how to improve. The adult word count (AWC) estimation algorithm is shown to reduce the relative Root Mean Square Error from an initial 42% to 7-8% after 5 hours of measuring time. If LENA's feedback suggests any potential development problem, parents can take action at a crucial early stage. LENA is a new processing system not only for parents/caregivers but for pediatricians, speech language pathologists, child development psychologists, and other researchers as well. This system represents one of the first breakthroughs in assessing early childhood language development and child environment conditions.

1. INTRODUCTION
This paper provides an introduction to the processing elements which make up the LENA system and its technologies [5]. The fundamental idea behind LENA is to utilize technologies to encourage human-human interaction for very young children. Hart and Risley's ground-breaking study of child development [3] showed the relationship between the amount of adult talk and interaction with children and children's language development and even their IQs and future success at school and in the workplace. The American Academy of Pediatrics also recommends parent-child interaction and is concerned about TV's impact on children under 2 [4].

The LENA system utilizes signal processing technologies to monitor the natural language environment of children, especially the home environment. A small light-weight digital recorder (DLP – digital language processor) is worn in the pocket of specially designed clothing of a child [6]. Currently, the DLP can hold up to 16 recording hours per day. All sounds in the child's environment, including his/her own voice, are generally recorded in an unobtrusive way. The LENA software processing system analyzes the recording and estimates the adult word-count, the adult-child-turn-taking-count, the child vocalization count, the TV time, and other traits to provide feedback regarding the language environment. With this hardware and software combination, parents/caregivers can now obtain information about a child's stage of language development, as well as how they may improve.

Another important feature of the LENA processing system is the automatic child vocalization assessment (AVA). Early detection is a crucial issue for any child language development delay or problem. In contrast to the current clinical approach which relies on parent report and relatively short observation time (often 30 to 60 minutes) in clinical settings [7, 8], AVA is based on recordings of up to an entire day, made in the natural home environment. AVA extracts the composition or complexity of a child's voice from the recording and links it to his/her speech language development status. Previously, there has been only limited research with small data sets in this area [9]. AVA is the first automatic screening tool for parents/caregivers, pediatricians and other professionals.

In the following sections, the complexity of the data, the difficulty of the task, and the task goals will be discussed. An overall system diagram will be presented. Since the overall system involves a range of processing sub-systems which in combination address a number of language development issues, we focus here on the more core fundamental ones in this paper, including segmentation and segment-ID, adult-word-count estimation and AVA.

2. DATA COMPLEXITY & TASK GOAL
The natural language environment of a child is one without any constraints since anything could happen. A recording from the natural home environment usually contains a significant amount of overlapping signals, various kinds of background noise, presence of TV/Radio and other media sounds, other children that may be confused with the targeting child (key child), near or far field signals (clear or faint signals), outdoor versus indoor settings, big rooms with strong echo versus small rooms with less echo, cocktail-party cases versus one-to-one story-telling in a quiet environment, infant-directed adult speaking style known as parentese versus adult-directed style. In short, recording data is tremendously variable and the processing of such data is difficult and challenging. High performance at the micro-level (e.g., for every 5 seconds, every 1 minute, etc.) may be difficult to achieve. Different from normal speech and speaker recognition, the accuracy on every bit of detail may not be the most important

concern and could be a long-term goal. Instead, the overall quality of the language environment and the general development status of a child are the most interesting topics. Thus, the major goal here is the estimation count at the macro-level (e.g., 1 hour, or a 12 hour recording), which makes it possible to achieve a high standard and create a viable tool for speech and language experts focused on child language development. As an example, it will be shown how the performance of the adult-word-count estimation improves with the increase of the length of observation time. In addition to the unique task goal, the very large recording database collected to date is another important factor that makes it possible to develop new speech processing sub-tasks for the overall system. Thus far, more than 65,000 hours of recording data from ordinary American families have been collected. The development and research of the LENA hardware and software system has further benefited from the use of a cluster-computer with more than 170 processors and 30 Terabytes of fast storage disk space.

3. SYSTEM OVERVIEW

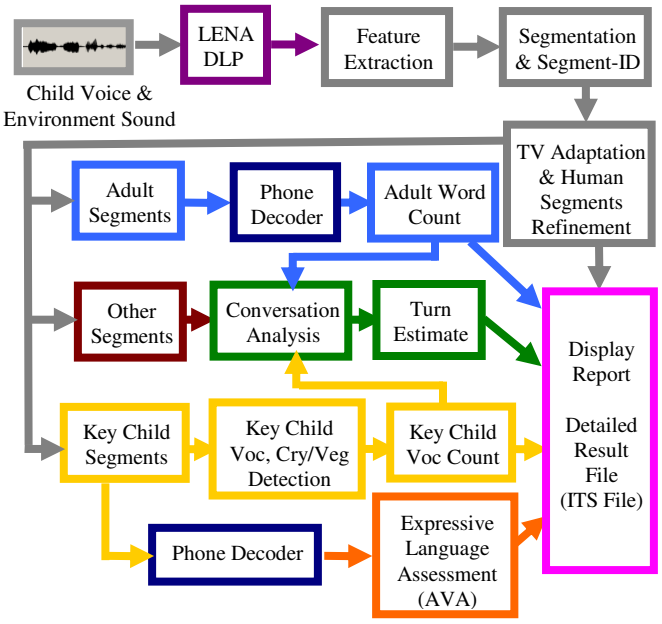

Figure 1: Overall speech processing framework for the LENA system.

A simplified LENA system diagram is shown above in Figure 1. The processing system starts with the acoustic feature extraction from recording data. Various acoustic features are extracted for different purposes, including MFC (mel-frequency cepstrum) [17], PMVDR (perceptual minimum variance distortionless response) [13] and SSC (Spectral Subband Centroids) [18].

One of the first steps for the LENA processing system is to locate and identify each sound class referred to as the segmentation and segment-ID process. All sounds in actual environmental recordings are categorized into one of 8 classes: key child, adult male, adult female, other child, TV (including Radio and other electronic media sound), noise, silence and overlap. All non-silence classes are further categorized into clear/faint sub-classes (related to near/far field). Overall, there are 15 sub-classes.

Once the segmentation and segment-ID process is performed, clear-adult-segments are further processed to produce adult-word-count estimates. Key-child segments are further processed to delineate normal vocalizations from cries, vegetative sounds and fixed signals [10]. Key-child segments are also decoded using a phone-decoder to extract their composition for AVA. Eventually, the conversation analysis is done based on the combination of adult-speech processing results, key-child processing results and other segmentation results. As a practical consideration, it is also required that the full processing time be within 0.5 real-time.

4. SEGMENTATION & SEGMENT-ID

As mentioned previously, LENA recording data is extremely complex and the acoustic characteristics of the signal vary dramatically. To be successful for the segmentation and segment-ID for such rich content data, the acoustic features need to encode all necessary information to distinguish the 8 classes mentioned above. We experimented with FFT-based MFC [17] and PMVDR [13]. It turned out that non-parametric MFC is slightly better than model-based PMVDR for this complex data. Experiments have also shown that for this specific type of data, MFC-36 is better than MFC-19 or MFC-20 which is usually used for speaker identification or recognition in the literature.

With the acoustic feature determined, the segmentation and segment-ID algorithms were investigated for this unique task. There is an existing 2-step method with BIC-based homogeneous segmentation as the first step and segment labeling as the second step [12]. The weakness of this method is the disjoint segmentation and labeling which could be otherwise optimized jointly. To this end, Hidden Markov Model (HMM) [11, 16] was considered. However, the uniqueness of the task makes it less appealing. LENA recordings contain segments with huge differences in duration, e.g. 1-hour silence or background noise versus 2-second speech. HMM has implicit exponential type duration modeling which may not be appropriate for this duration variation. A Minimum Duration Gaussian Mixture Model (MDGMM) was proposed and developed under a maximum likelihood framework for this unique task. The state transitions in HMM are now all set to 1 so that there is no implicit exponential type of duration model, which makes it flexible for big duration variation. The removal of implicit duration modeling also introduces the problem of too short segments and noisy results. The minimum-duration constraints can prevent too short segments and smooth out the maximum likelihood result. It is not difficult to show that with a minimum duration constraint, any segment can be decomposed into several segments with the same ID whose durations are all between the minimum duration and twice the minimum duration. Consequently, under a set of minimum duration constraints, any segment sequence could be expressed by segments with no more than twice the minimum durations. This will guarantee that the active nodes in a maximum likelihood search space will be confined to twice the minimum durations. Thus, MDGMM can globally achieve maximum likelihood decoding for a 16-hour recording without any pruning.

Table 1: Comparison between MDGMM and 2-step method

Acoustic Class	T2-BIC	MDGMM
Key child	70.6%	76.7%
Adult	77.8%	85.5%

Experiments have shown that MDGMM performs better than the 2-step T2-BIC-based method. Table 1 shows the result of a specific comparative experiment, where only two major detection rates of concern are compared for simplicity.

As noted above, TV and other media sounds (radio) need to be distinguished from live human speech. For today's high quality media, the available audio makes this task less straightforward. The frequency band of signals may not be an effective distinguishing feature. Some high quality media audio may have a broad frequency band while some relatively faint live speech may reveal certain narrow band characteristics. In addition, the varied types of media sounds and live speech may differ greatly from recording to recording, from one moment to another. Currently, TV detection is further refined by a fast localized adaptation after the initial MDGMM segmentation/segment-ID. During the adaptation, each 10-minute recording section will be recalculated for likelihoods using the adapted model locally derived from the surrounding 30 minutes of data. With updated likelihood for each frame, MDGMM will be applied again to find maximum likelihood results. Currently, there are 3 iterations of adaptation.

Another important issue is the existence of faint sounds which generally may not contribute to the language development of a child, and should not be credited for language environment measures. Currently, a likelihood-ratio-test (LRT) method is used for this purpose. For any non-silence segment, its likelihood is compared with the silence-likelihood of the same segment. If the ratio is low enough, then it is considered as faint sound. The LRT-threshold was tuned based on a relatively small test-set. With the LRT faint/clear sound detection, the 15 subclasses mentioned above are eventually generated.

Table 2: Confusion matrix without segment-ID refinement

	Key Child	Clear Adult	TV	Others
Key Child	***75.9***	*12.1*	0.1	11.9
Clear Adult	*3.1*	***81.0***	3.9	12.1
TV	0.4	8.1	***70.5***	21.0
Others	4.7	14.1	6.2	***75.0***

The diarization performance of the above segmentation/segment-ID method is shown in the confusion matrix of Table-2, where each row adds up to 100% and is based on human transcription, and each column is the machine result. The testing data were selected from 70 recording files from natural home environments, with child age ranging from 2-month to 36-month and each month-age containing 2 recordings from 2 different children. 1-hour high speech activity regions were further selected randomly from each recording file for human transcription. So, there are 70 total hours of test data from 70 recordings of 70 children.

For LENA's language environment monitoring and automatic child vocalization assessment purposes, the accurate detection of key-child and adult is important. The detailed analysis of the above confusion matrix showed that there is confusion among key-child, other-child and adult which may have a negative impact on the goal of the LENA system. To further reduce such confusion, a model-based feature, PMVDR [13], is utilized, which potentially works better for live human segments because the model assumption fits better to human speech. The final segmentation/segment-ID confusion matrix with PMVDR-based refinement is shown in Table 3. As can be seen the confusion among human segments is reduced.

Table 3: Confusion matrix with PMVDR-based refinement

	Key Child	Clear Adult	TV	Others
Key Child	***76.0***	*7.3*	0.1	16.6
Clear Adult	*1.9*	***82.0***	3.9	12.3
TV	0.5	7.8	***70.5***	21.1
Others	4.5	13.6	6.2	***75.7***

Overall, the segmentation/segment-ID performance varies from 70.5 – 82.0%, where chance is 25%. The computation load of segmentation/segment-ID mainly comes from the Gaussian calculation for likelihoods. A fast Gaussian calculation method was implemented which improved the speed by more than 5 times without losing segmentation/segment-ID accuracy. This is the major factor which assures 0.5 real-time overall processing speed.

5. ADULT WORD COUNT ESTIMATION

As pointed out by Hart and Risley [3], adult word count (AWC) is one of the most important factors contributing to child language environment. As one tries to speak more to a child, the topic and content tend to be diversified, and so is the vocabulary. Consequently, AWC is one of the major LENA measurements for child language environment.

Although word count is the by-product of word speech recognition, it is quite a unique task. The focus is to obtain count not content. How to utilize the uniqueness to optimize the problem for high performance, especially at the macro-level, is our major consideration. Because of the different focus, word recognizer based AWC estimation is naturally considered "heavy-weight". The potential problems of this approach are the choice of vocabulary and grammar. A prior lexicon that contains all available words is needed as well, and depending on each family environment, as well as specific proper names, places, expressions used for the child, this is a major research task itself. For highly spontaneous speech in a natural environment, it is difficult to find appropriate choices. In addition, it is highly language dependent and susceptible to dialects, accents, and family specifics.

The method used in the current proposed solution is based on phone-decoding and Least-Squares linear regression with vowel, consonant counts and their nonlinear variants, along with the durations in order to target human transcription word counts. This proposed method can overcome many problems associated with the word-based method, providing flexibility over different environment/family cases. More importantly, one more layer of modeling actually provides a chance for adjustment which could even "remedy" the error caused by incorrect segmentation. Theoretically, Least-Squares linear regression is unbiased under the Gauss-Markov condition [14], which represents another benefit for consideration. It has been shown that the asymptotic performance of the count estimation will be lower bounded by the bias of the estimator. Achieving an unbiased estimation of the adult word count then becomes critical.

Specifically, based on experiments, the formula for AWC estimation for an adult segment (or utterance) is determined as:

$$AWC = b_1 c + b_2 v + b_3 \sqrt{c} + b_4 \sqrt{v} + b_5 d + b_6 d_s$$

where b_i are coefficients trained with Least-Squares, c is consonant count, v is vowel count, d is duration of the segment, d_s is the duration without silence.

To measure the performance, two quantities are used. They are the relative error mean (or relative average error) and the relative Root Mean Square Error, defined as $e_m = E[e]/E[t] \times 100\%$ and $rmse = \sqrt{E[e^2]}/E[t] \times 100\%$ respectively, where E is the expectation, e and t are the error and true-value of AWC for a specific measuring region. So, $rmse$ is related to error standard deviation or variance, while e_m is about error mean.

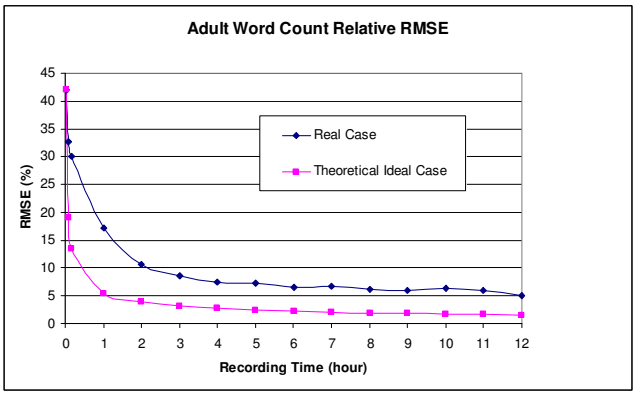

Figure 2: Adult Word Count (AWC) Performance Measured in terms of Relative Root MSE (in %) versus Measuring Time.

Currently, the relative overall word-count error for the 70-hour test set is 2%. A Leave-One-Out-Cross-Validation [19] experiment yielded similar results. As noted above, adult word count estimation at the micro-level for natural environments might be difficult. However, the under-estimations and over-estimations over time are expected to cancel out at a larger scale. Experiments have shown that the relative RMSE will decrease with the increase of the measurement time. The graph in Figure 2 shows the relative RMSE result for our current AWC processing algorithm on the 70-hour test set. The blue-diamond line is the actual real case. The result shows that the relative RMSE for AWC for 1-minute regions is 42%. It reduces to 33% for 5-minute regions, 30% for 10-minute regions, 17% for 1-hour regions, and so on. When the measurement region is above 5 hours, a level of relative RMSE near 7-8% is achieved and the overall trend is decreasing all the time. The red-square line in Figure 2 is the theoretical ideal case where estimation is unbiased and the estimation errors in different recording regions are uncorrelated and identically distributed. It can be shown that under this ideal case, the relative RMSE is going to decrease at the rate of the square root of r when the measurement time length increases at the rate of r. In other words, the decreasing factor is $1/\sqrt{r}$. Starting with the same 42% relative RMSE for 1-minute and applying the factor of $1/\sqrt{r}$, the theoretical performance curve under the ideal case can be derived. We can see that the actual relative RMSE has the same trend as the theoretical one. The difference between them is bigger when measurement time length is relatively short, and the difference becomes smaller when measurement time length increases. This is probably because in a real case, the estimation errors of adjacent recording regions tend to have some correlation while the estimation errors of regions some time apart tend to be uncorrelated. These curves show how AWC macro-level performance is related to and improved from micro-level performance.

6. AUTOMATIC VOC ASSESSMENT

Child language development is a process involving both cognitive and articulation algorithm development. The related status of a child will somehow be "encoded" in his/her voice. At the very early stage, crying is the way to communicate; next quasi-vowels, squeals, growls and other protophones will be developed, followed by babbles, syllables and eventually words, phrases and sentences [10]. Naturally, the composition of a child's speech (vocalization) could provide some indication of his/her language development status. How to automatically extract such composition information from recording data is the actual problem. The child vocalization composition could be at different levels, starting from acoustic feature space to phone, syllable, word, etc. It is known that child speech or vocalization recognition is a very difficult task because of the high variation of vocalization and the not-well-formed pronunciation due to the nature of the child speech development process. Child speech recognition at a higher level such as word or phrase may have more uncertainty than that of a lower level such as phone or acoustic feature space. As an initial exploration, we chose to focus on the lower level of phone and acoustic feature space. For early child speech development from 0 to 36 months or so, the lower level composition may be more relevant. This is another reason for our current focus. The phone composition used so far is the phone count distribution or the frequency for each possible phone or similar unit.

The major difficulty in child speech (vocalization) recognition is the modeling of child speech. Even at the phone level, there are many variants of phones compared to adult speech. How to find different variants from data and model them is a challenging problem. For convenience and as an initial exploration, we chose to use the Sphinx adult acoustic model for phone decoding [21]. Obviously, from general speech recognition point of view, using an adult acoustic model for child speech recognition may result in a mismatch. Here again, we argue that different from normal speech recognition, the detail accuracy is not the final goal, what matters is the information about child speech development and whether the phone composition or distribution defined or partitioned by an acoustic model can reserve enough relevant information. Actually, for a specific vocalization, whether it is called (or recognized) as [a] or [i] may not be that important as long as the phone decoder behaves consistently and the resulting phone distribution contains sufficient child speech development information.

The process of child speech development actually gives us another angle to look at the problem. During language development, children learn from the adults around them. Child speech approaches adult speech and eventually converges to adult speech. The difference between child speech and adult speech should become smaller and smaller as child speech development proceeds. Thus, adult speech can actually serve as a reference for

child speech development. From this point of view, using an adult speech model for child speech development assessment makes some sense. Actually, we tested this idea by decoding child and adult speech in our corpus using Sphinx phone decoder with adult acoustic model and measuring the Kullback-Leibler divergence (roughly distance) between the child phone distribution of each recording and the average adult phone distribution of the corpus. Figure 3 shows the result where each point corresponds to each recording and the line is a polynomial fit of all points. It is verified that the difference between child speech and adult speech becomes smaller as child age increases, and the child phone distribution obtained by using an adult speech model does contain information about child speech development.

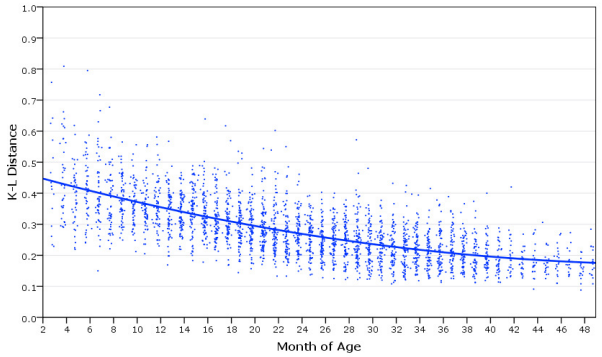

Figure 3: K-L divergence between child and adult phone distributions as a function of child chronological age.

Based on the child phone distribution described above, we tried two tasks. One is to estimate the chronological ages of typically developing children. The other is to predict the child speech development scores assessed by human speech language pathologist (SLP).

For the first task, the experiment data consisted of 243 children with the age ranging from 2 to 48 months. There are 2124 natural home environment recordings from these children. All recordings are above 12 hours. These children were selected based on SLP assessment with the scores within 1-standard-deviation of the mean. Because of the typical status they have, their chronological ages could be considered as their developmental ages. Delayed or advanced children were not selected for the experiment because their developmental ages may deviate from chronological ages to a relatively large extent. This task uses a 2-stage linear regression model to estimate the chronological ages of the children. Since each child may have more than 1 recording at different times, his/her age is changing and should be associated with the recording. Thus, there is a total of 2124 estimated ages. The 2-stage linear regression model could be regarded as a piecewise linear model. At the first-stage, a global linear model is used to estimate a preliminary age:

$$a_p = \sum b_i p_i$$

where b_i is a linear coefficient, p_i is a phone frequency. At the second-stage, a local finer linear model is selected based on the preliminary estimate a_p to give refined final age estimation:

$$a_f = \sum b_i(a_p) p_i$$

where $b_i(a_p)$ is the local linear model parameter associated with a_p. All linear models are trained with Least-Squares. The leave-one-child-out-cross-validations [19] are done for both stage-1 and stage-2 and the overall cross-validation experiment gives 0.90 correlation between the chronological ages and the estimated ones.

For the second task, the experiment data consists of 336 children with the age ranging from 2 to 48 months, including both delayed and advanced children. There are 2910 natural home environment recordings from these children. All recordings are above 12 hours. Each child was assessed by a human SLP for either his/her PLS4 score [7] or REEL score [8] on a day which was within a week from a recording day of that child. Both PLS4 z-score and REEL z-score are age-normalized scores so that the scores of a large set of children from a month-age have zero-mean and unit variance. The PLS4 score or REEL score is the shifted and rescaled version of its z-score so that the mean is 100 and the standard deviation is 15. Thus, SLP assessed scores are all about the relative comparison of the development status of the children with a same month-age. Since the scores for different ages have the same distribution (same mean and variance), they can be compared across different month-ages. For robustness, in the experiments for this task, all the SLP assessed scores for a child were averaged to obtain one single score for the child. Thus, in this task, each child has one averaged SLP score as the "truth" for the development status of the child. The task is to predict the development "truth" for each child based on his/her phone distribution derived from each recording. The predicted score from each recording of a child was averaged together to eventually give the final predicted score for that child. Intuitively, in order to be consistent to the age-normalized nature of SLP scores, the prediction model need to be age-dependent, i.e. for each month-age, there should be a prediction model. Linear models are used for each month-age as:

$$s = \sum w(a,i) p_i$$

where a is a month-age, $w(a,i)$ is the linear coefficient for age a and phone i, s is the predicted score. Least-Squares was used to train all linear models. For the model training of age a, the data from age $a-b(a)$ to age $a+b(a)$ will be used. $b(a)$ is called age-band and could be different for different ages. Optimal age-bands were obtained by Dynamic Programming under age-smooth-constraints. The leave-one-child-out-cross-validation experiment gives 0.72 correlations between the SLP "truth" and the predicted scores.

In order to incorporate the sequence information contained in decoded phone-sequences, bi-phone instead of uni-phone is considered as the inputs of linear models. Since the number of bi-phone is the square of that of uni-phone which is much larger and may cause over-fitting. To resolve this issue, Principal Component Analysis (PCA) is used to reduce the dimension of bi-phone inputs. Experiments showed that the dimension reduction to 50 gives the best cross-validation result. Under this scheme, the leave-one-child-out-cross-validation gives 0.75 correlations

between the SLP "truth" and predicted scores. The final scatter plot of the target scores and the predicted scores of Leave-one-child-out-cross-validation experiment is shown in Figure 4. Since each child used a different prediction model in this cross-validation experiment, the effect of different models could be regarded as some extra "noise". In spite of this "noise" and the potential "noise" of human SLP scores such as the ones due to the subjectivity, limited observation time, etc., AVA scores still highly agree with SLP scores, signifying the validity of the method.

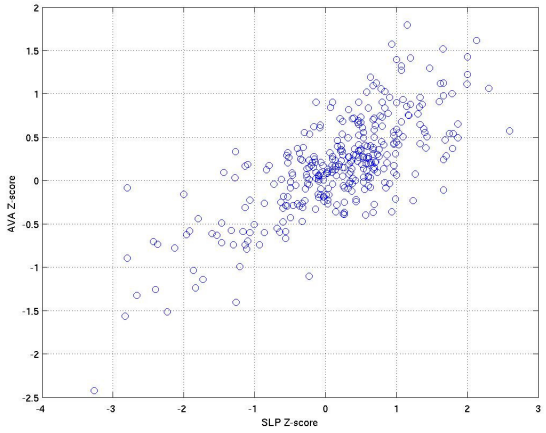

Figure 4: AVA scatter plot. AVA score versus SLP score.

As a summary, the proposed solution here therefore allows for an ongoing, long-term assessment of child vocalizations leading to language development, as well as an assessment of the adult language environment the child is exposed to during their formable years (0-48 months).

7. DISCUSSION

The proposed LENA hardware and speech processing algorithms open a new and exciting field for engineers and scientists interested in child speech/language assessment using signal processing, machine learning, and other technologies. Further detailed technical papers are emerging in the fields of early childhood language development and are expected to be published in the near future. The integrated LENA hardware/software system has addressed a number of new technical issues, and further technical challenges will be addressed as user feedback accumulates. The current system and technologies are by no means final. The proposed processing sub-tasks presented in this study have shown reliable acoustic event detection for key child, clear adult, TV, and other classes, as well as effective adult word count (AWC) estimation over time. These factors will contribute to advances in the field of child language development assessment. The proposed LENA system will continue to evolve with further improvements and new processing features added in the future.

8. ACKNOWLEDGEMENTS

We greatly acknowledge Terry Paul for conceiving of the LENA System and for personally funding and directing its development as well as the development of the Infoture Natural Language Corpus.

9. REFERENCES

[1] http://www.becta.org.uk/etseminars/presentations/2004-10-21/8/slides/slides.pdf. Child-Computer Interaction Overview.

[2] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, P. Barker, "Applications of automatic speech recognition to speech and language development in young children", ICSLP-1996.

[3] B. Hart, T. Risley, "Meaningful Differences in the Everyday Experience of Young American Children", Paul H. Brookes Publishing Co., Inc. 1995

[4] http://www.aap.org/sections/media/ToddlersTV.htm

[5] http://www.lenababy.com/

[6] http://www.lenababy.com/LenaSystem/AboutLena.aspx

[7] I. Zimmerman, V. Steiner, and R. Pond. *Preschool Language Scale, Fourth Edition.* San Antonio: The Psychological Corporation, 2002

[8] K. Bzoch, R. League, and V. Brown. Receptive-Expressive Emergent Language Test, 3rd Ed. Austin: PRO-ED, 2003

[9] H. Fell, J. MacAuslan, L. Ferrier, S. Worst, K. Chenausky, "Vocalization Age As a Clinical Tool", ICSLP-2000

[10] K. Oller, "The Emergence of the Speech Capacity", Lawrence Erlbaum, January, 2000

[11] L. Rabiner, B.-H. Jung, "Fundamentals of Speech Recognition", Prentice-Hall, 1993.

[12] B. Zhou, J.H.L. Hansen, "Efficient Audio Stream Segmentation via the Combined T^2 Statistics and Bayesian Information Criterion", IEEE Trans. on Speech & Audio Processing, vol. 13, issue 4, July, 2005

[13] U. Yapanel, J.H.L. Hansen, "A New Perceptually Motivated MVDR-based Acoustic Front End (PMVDR) for Robust Automatic Speech Recognition", Speech Communication, Vol. 50, February, 2008.

[14] G. Robinson, "That BLUP is a Good Thing: The Estimation of Random Effects", Statistical Science, Vol. 6(1) Feb, 1991.

[15] E.J. Wallen, J.H.L. Hansen, "A Screening Test for Speech Pathology Assessment Using Objective Quality Measure," ICSLP-1996, pp. 776-779, Philadelphia, PA, Oct. 1996.

[16] J. Deller, J. Hansen, J. Proakis, *Discrete-Time Processing of Speech Signals,* IEEE Press, New York, NY, 2000.

[17] S. Davis, P. Memelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on ASSP, vol. 28, pp 357-366, 1980

[18] K. Paliwal, "Spectral subband centroids as features for speech recognition", Proceedings of 1997 IEEE Workshop on Digital Object Identifier.

[19] S. Haykin, "Neural Networks, a Comprehensive Foundation", 2nd-Edition, Prentice-Hall Inc. 1999

[20] M. Murthi, B. Rao, "All-pole Modeling of speech based on the minimum variance distortionless response spectrum", IEEE Trans. on SAAP, vol 8(3), May 2000

[21] http://cmusphinx.sourceforge.net/html/cmusphinx.php

[22] NIST Scoring Toolkit, http://www.nist.gov/speech/tools/