31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Child Vocalization Composition as Discriminant Information for Automatic Autism Detection

Dongxin Xu, Jill Gilkerson, Jeffrey Richards, Umit Yapanel, Sharmi Gray

Abstract—**Early identification is crucial for young children with autism to access early intervention. The existing screens require either a parent-report questionnaire and/or direct observation by a trained practitioner. Although an automatic tool would benefit parents, clinicians and children, there is no automatic screening tool in clinical use. This study reports a fully automatic mechanism for autism detection/screening for young children. This is a direct extension of the LENA™ (Language ENvironment Analysis) system, which utilizes speech signal processing technology to analyze and monitor a child's natural language environment and the vocalizations/speech of the child. It is discovered that child vocalization composition contains rich discriminant information for autism detection. By applying pattern recognition and machine learning approaches to child vocalization composition data, accuracy rates of 85% to 90% in cross-validation tests for autism detection have been achieved at the equal-error-rate (EER) point on a data set with 34 children with autism, 30 language delayed children and 76 typically developing children. Due to its easy and automatic procedure, it is believed that this new tool can serve a significant role in childhood autism screening, especially in regards to population-based or universal screening.**

I. Introduction

AUTISM SPECTRUM DISORDER (ASD) has gained considerable attentions over the last decade [1]. Significant increases in research grants from 1997 to 2006 have been reported, with a clear shift from basic science to clinical and translational research [2]. In clinical practice, diagnosis is the first step. Early diagnosis of autism is important in order for young children with autism to access effective early intervention services [3, 4, 7]. The American Academy of Pediatrics recommends autism screening for all children at the 18 month and 24 month checkups [5]. However, a survey completed in 2004 indicated that only 8% of primary care pediatricians routinely screened for Autism [6]. For parents with concerns, it typically takes at least 6 months to obtain a clinical diagnosis [3] due to the laborious nature of the existing screening/diagnostic procedures and an insufficient number of trained personnel relative to the large number of children in need of evaluation. Efficient and/or automatic tools for autism detection can help facilitate the evaluation process. This study reports a fully automatic mechanism for autism detection for young children using the LENA™ (**L**anguage **EN**vironment **A**nalysis.) system and child vocalization composition as discriminant information.

Autism is characterized by: (i) *qualitative impairments in social interaction* shown by the abnormalities in such behaviors as eye gaze, body posture, sharing interests and emotions; (ii) *qualitative impairments in communication* shown by language development issues such as delayed status, problems initiating and sustaining conversations, repetitive patterns; (iii) a *restricted repertoire of interests, behaviors and activities* shown by an adherence to certain topics, routines, rituals, motor manners, parts of objects and sensory abnormalities [7]. In recent years, increased research efforts have been made towards early identification of autism. For instance, [8] reported on the discovery of early attention differences that may lead to early identification and new therapies; [9] reported unusual use of toys in infancy as an indicator of later autism; [10] reported vocal differences and abnormalities in high risk infants at 12 months; [11] reported less responsiveness to their names in 12-month-old high-risk children; [12, 13] focused on specific abnormality in prosody of children with autism. These findings were based primarily on subjective observations and rarely related to automatic or machine-generated objective measures. [14] showed the potential of an automatic measure for prosodic quality rating in a laboratory setting. In addition to the efforts associated with detection, there are reports on intervention employing a computer or robot. [15] and [16] described a computer-animated tutor for vocabulary and language learning and a robotic prosody therapist, respectively.

The current standard diagnostic tools in clinical practice include the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule-Generic (ADOS-G) [3]. Some of the existing screens for early identification of autism include CHAT (the Checklist for Autism in Toddlers), the quantitative CHAT, the Modified CHAT, STAT (the Screening Test for Autism in Toddlers), PDDST-II (the Pervasive Developmental Disorders Screening Test-II) and ESA (the Early Screening for Autism questionnaire) [3]. Because these instruments either require parent participation and/or direct observation, rating and scoring by a trained practitioner, they are labor-intensive and necessarily include some degree of subjectivity. Evaluation in an unfamiliar clinical setting may also influence child behavior and potentially influence the evaluation.

The LENA system reported here introduces an objective, unobtrusive and easy-to-use automatic system for autism detection based on audio recordings from the natural home environment. An overview of the system, methods, and the experiment results are provided in the following sections.

Authors are with LENA Foundation, Boulder, CO 80301 USA. (e-mail: dongxinxu, jillgilkerson, jeffrichards, yapanel@lenafoundation.org).

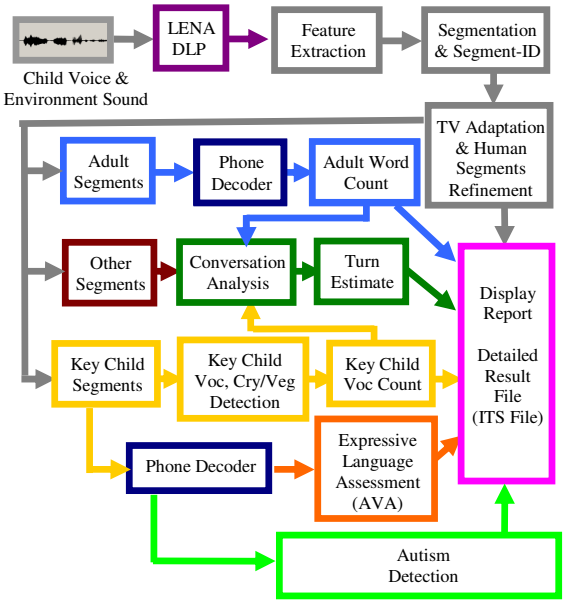

Fig. 1. Diagram of the LENA System.

II. System Overview

As shown in Figure 1, the LENA system starts with a small light-weight digital recorder (DLP – digital language processor) worn by the child in the pocket of specially designed clothing [17]. The DLP can hold up to 16 hours of audio. All sounds in a child's environment, including his/her own voice, are recorded in an unobtrusive way. The LENA system first transfers the audio data from the DLP into a computer, then analyzes the data, producing estimates for the adult word-counts, the adult-child interactions (turns) and the child vocalizations. The system also produces the estimates for the amount of audible TV/electronic media, an automatic expressive language assessment (AVA) score, the automatic autism detection result and other traits to provide feedback regarding the language environment and the development of the child. This hardware and software combination allows parents/caregivers to obtain information about a child's development as well as improvement over time, providing parents and clinicians with the opportunity to intervene when indicated at an early stage [18].

As described in [18], all sounds in the 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 based on likelihood ratio test. Overall, there are 15 sub-classes. After this segmentation and segment-ID process is performed, clear-adult-segments are further processed to produce adult-word-count estimates. Clear key-child segments are further processed to delineate normal vocalizations from cries, other fixed signals and vegetative sounds. Clear key-child segments are also decoded using a phone-decoder to extract the child's phone-level composition for AVA and automatic autism detection. The system processing time is required less than 0.5 real-time. The segmentation/segment-ID accuracy varies from 70.5% to 82.0%; the adult word-count performance in terms of the Relative Root Mean Square Error varies from 42% for 1 minute measuring length to below 7-8% after 5 hours of measuring time; the AVA scores achieves 0.75 correlations with the scores assessed by human speech language pathologists using standard language assessments. More detailed information can be found in [18]. The rest of the paper focuses on the automatic autism detection using child vocalization composition.

III. Child Vocalization Decomposition

As mentioned above, childhood autism is characterized by the abnormalities in social interaction, communication, language development and repetitive stereotyped behavior. It is reasonable to assume that certain characteristics of these abnormalities could be exhibited and detected within a day-long audio recording. Specific abnormalities of vocalization and prosody in children with autism have been reported [10, 12, 13]. Although modeling social interaction in an audio recording is not straightforward, modeling child vocalization is relatively easier. The question may naturally be raised regarding whether children with autism tend to produce certain types of sounds more often and certain types of sounds less frequently than other children. Are there any discernible differences in the vocalization composition between children with and without autism? Using composition analyses to distinguish different materials is common in Chemistry and other scientific areas. To test child vocalization composition, it is necessary to obtain sufficient samples and to create an efficient, consistent and objective method to "decompose" the vocalization samples into different "components" (or categories). It would be virtually impossible to obtain similar large-scale quantitative comparisons relying solely on human observation. The LENA system now provides the means to perform such tests.

A large quantity of child vocalization samples are automatically collected from day-long recordings by the LENA segmentation/segment-ID subsystem. Because the key child continuously wears a DLP in a pocket near his/her chest, the task of identifying key child audio segments becomes easier. After the key child segments are detected, the phone-decoding subsystem is used to recognize the phone-like sounds within the segments. A phone-decoder based on the open-source Sphinx system [19] is used, which contains 39 regular English phone models such as [t], [a] and 7 filler models to absorb pause, breath, hesitation, possibly crying and other categories in key child segments. There are in total 46 categories collectively referred to as uni-phones in the study. The frequency of a uni-phone is defined by the count of that uni-phone normalized by the total count of all uni-phones in a recording. All such frequencies constitute the probability density function (pdf) for uin-phone distribution.

The composition of a child's vocalization can be quantified by this pdf function. In addition to the uni-phones, to make use of the dynamic information contained in phone-sequences, uni-phone-pairs (called bi-phones) are also tested. Since the bi-phone pdf function has high dimensionality (roughly 46x46 = 2116), Principal Component Analysis (PCA) is used to reduce the dimension to 50 (called bi-phone-50 in the study) [18]. Similarly, tri-phone and longer phone-sequences could potentially be utilized.

Young child vocalization/speech recognition is a difficult task and somewhat ill-defined due to the immature nature and large variation of child pronunciation. One advantage of using composition information rather than recognizing specific vocalization abnormalities for autism detection is that the fine detail accuracy of phone-decoding may be less important for the ultimate goal of autism detection. As long as the system works consistently and objectively to produce the decomposition with high enough "resolution" for autism detection, it is less important whether a particular vocalization is recognized as [a] or [i] or other categories.

IV. Detection and Data Analysis Methods

Unlike most autism detection research in which only a few variables are examined (e.g. attention [8], pitch range [13]), the uni-phone pdf and bi-phone-50 approaches utilize high dimensional features. Although each individual component in the uni-phone or bi-phone-50 may not contain significant discriminant information, the combination of them can be powerful enough to achieve much better performance. A data-driven approach is used to find the optimal transform to convert high-dimensional data into low- or 1-dimensional space. Specifically, Linear Discriminant Analysis (LDA) [20] is utilized to obtain the linear projection with optimal Fisher-Ratio. Under certain assumptions, the posterior probability of a child's recording belonging to the autism-class can be estimated. The formal description of the method is as follows.

For a day-long recording of a child, the uni-phone pdf or bi-phone-50 parameters are calculated, annotated as $X_i = (x_{i1}, x_{i2}, ..., x_{id})^T$ where i is the index of the recording and d is the dimension of the parameters. The child class-ID is coded 1 for autism and 0 otherwise, annotated as c_i. A linear projection $W = (w_1, ..., w_d)^T$ with optimal Fisher-Ratio $f = W^T S_B W / W^T S_W W$ is searched, where S_B and S_W are between-class and within-class scatter matrices respectively. With the optimal W, the multi-dimensional input data X can be converted into one-dimensional value: $y = W^T X$. Under the assumption that y is Gaussian distributed for autism-class and non-autism-class with an equal variance (this is actually the underlying assumption of LDA), the means of m_1 for autism and m_0 for others, and the variance σ^2 can be estimated. With the a priori probabilities p_1 for autism and p_0 for others, the posterior probability of a recording belonging to the autism-class given the input X could be calculated as:

$$P(c=1 \mid X) = \frac{p_1 G(y - m_1, \sigma^2)}{p_1 G(y - m_1, \sigma^2) + p_0 G(y - m_0, \sigma^2)}$$

where $G(y - m, \sigma^2)$ is the Gaussian function with mean m and variance σ^2.

With a decision threshold t, any recording with the posterior probability above t could be considered belonging to the autism-class. By varying t from 0 to 1, the ROC curve can be obtained and the equal-error-rate (EER) point on ROC could be determined, i.e. the point with the miss-detection-rate equal to the false-alarm-rate. EER is used as the performance measure for comparison of different cases. It should be noted that the choice of the prior p_1 or p_0 does not affect the ROC and the corresponding EER.

One important issue of data-driven approaches is the generalization or the potential of models over-fitting with training data. To obtain realistic performance estimation, cross-validation is needed. To make full use of the data available, the leave-one-out-cross-validation scheme [21] is utilized. In the actual performance analysis, various levels of targets are left out for cross-validation, including recording, child and recorder. In the recording-left-out test, the posterior probability (pp) of a recording is calculated with the LDA and Gaussian models described above trained using all recordings but the targeted one itself. This is circulated through all recordings to obtain the pp for all of them. Similarly, in the child-left-out test, in addition to the target recording, all other recordings from the same child are left-out for the model training. In the child-and-recorder-left-out test, all recordings from the same child or the same recorder of the target recording are left out for its model training. By performing various levels of left-out-cross-validations, we are attempting to ensure that it is the signature of autism captured by the models and reflected in the performance report, not the confounding signature of a child or a recorder.

Because young children develop rapidly, the characteristics of different month-ages could be significantly different. To further improve the performance and test month-age effects, age-normalization is tested. For each month-age a, the mean and variance are estimated for each input parameter x_j using the recordings from typically developing children of ages $a - band \le a \le a + band$. Normalization results in the transformed parameters with 0-mean and unit-variance for each month-age: $\bar{x}_j = (x_j - age\_mean_j) / age\_std_j$. This can be regarded as part of the modeling process and is also tested with leave-out-cross-validation.

As indicated above, one child may contribute multiple recordings. The posterior probability of a child being the autism-class can be estimated by assuming the independence of different recordings: $pp = \sqrt[n]{\prod_{i=1}^{n} pp_i}$ where pp_i is the posterior probability of i-th recording and n is the number of recordings for the child.

V. Experiment Data and Test Results

The current study includes 34 children with autism (225 recordings), 30 children with language delay (290 recordings) and 76 typically developing children (712 recordings). All recordings are at least 12-hour long. There are in total 140 children (1227 recordings). Figure 2 shows the recording distribution over age (note: a child may have multiple recordings at different month-ages).

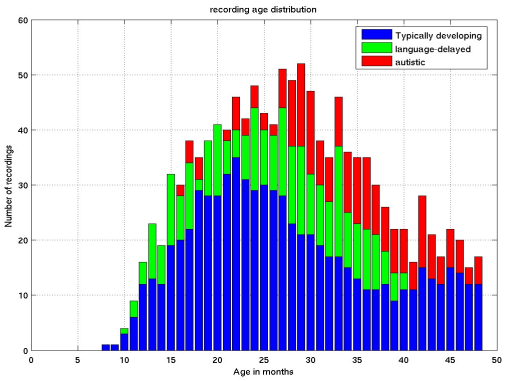

Fig. 2. Recording distribution over age.

3 different detection tasks were tried: (1) autism versus language-delay (a vs d); (2) autism vs. typical-development (a vs t); (3) autism vs. language-delay and typical-development (a vs d + t). The goal of the tasks is to detect autism from other cases. This was tested at the recording-level or child-level (shown in the columns of Table-1,2,3). Various levels of left-out-cross-validations were tested. First, without age-normalization, an experiment was tried to compare the cross-validations of leave-recording-out and leave-child-out. The EERs are shown in Table-1 with significant differences between these 2 cross-validation schemes, suggesting that leave-recording-out may have child signature interfered with the result and is not realistic; instead, the child-left-out-cross-validation should be used. The second experiment tested age-normalization (done before cross-validation) and the effect of recorders (DLPs). By comparing Table-1 and Table-2, it is verified that age-normalization can significantly improve the EERs, which, from the opposite perspective, demonstrates the significant age differences in young children. The EERs of leave-both-child-and-recorder-out are very close to those of leave-only-child-out, indicating that the recorders used have no effect on vocalization composition features and may not necessarily be included in cross-validation. The third experiment tested age-normalization inside cross-validation. As shown in Table-3, there is basically no difference to include age-normalization inside cross-validation compared with the EERs in Table-2 where age-normalization is done before cross-validation. Figure 3 gives an example of ROC curves and EER points. The above results show that bi-phone-50 performs slightly better than uni-phone with age-normalization; child-level performance is better than that of recording-level, suggesting that collecting more recordings (and/or longer recordings) could enhance the performance.

Overall, the cross-validation tests achieved about 77% to 83% recording-level accuracy and 85% to 90% child-level accuracy at EER points for the task of discriminating autism from language-delay and the task of discriminating autism from language-delay and typical-development.

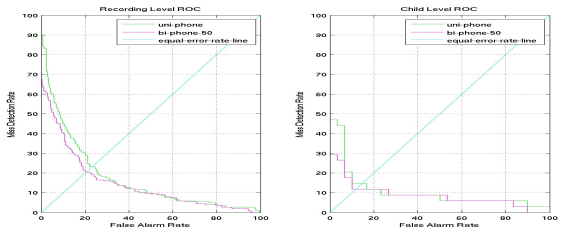

Fig. 3. Example ROC curve of detecting autism from delayed-ones, using leave-child-out and age-normalization inside cross-validation. Left: recording-level ROC; Right: child-level ROC. The straight lines in both plots are EER lines. Green lines are uni-phone; Red Lines are bi-phone-50.

TABLE -1
EQUAL-ERROR-RATE (%) WITHOUT AGE-NORMALIZATION
(THE COLUMNS OF "RECORDING" AND "CHILD" MEAN RECORDING-LEVEL AND CHILD-LEVEL EER. SAME MEANING IN TABLE-2, 3)

Leave-out	Detection case	Uni-phone		Bi-phone-50	
		recording	child	recording	child
Recording	a vs d	14.83	3.33	13.33	5.88
	a vs t	14.22	8.82	12.92	8.82
	a vs d + t	14.87	8.82	14.17	7.55
child	a vs d	24.44	8.82	24.00	17.65
	a vs t	20.00	14.47	21.33	14.71
	a vs d + t	19.56	12.26	20.16	17.65

TABLE -2
EQUAL-EEOR-RATE (%) WITH AGE-NORMALIZATION
(AGE-NORM IS NOT INSIDE LEAVE-OUT-CROSS-VALIDATION)

Leave-out	Detection case	Uni-phone		Bi-phone-50	
		recording	child	recording	child
child	a vs d	23.11	14.71	20.44	11.76
	a vs t	14.67	11.76	12.89	9.21
	a vs d + t	17.33	12.26	16.97	9.43
child + DLP	a vs d	23.45	14.71	20.44	11.76
	a vs t	15.11	11.76	12.92	9.21
	a vs d + t	17.78	12.26	17.33	9.43

TABLE -3.
EQUAL-ERROR-RATE (%) WITH AGE-NORMALIZATION INSIDE CROSS-VALIDATION

Leave-out	Detection case	Uni-phone		Bi-phone-50	
		recording	child	recording	child
child	a vs d	23.11	14.71	20.44	11.76
	a vs t	15.45	11.76	13.34	10.53
	a vs d + t	17.78	13.21	17.33	9.43

VI. Conclusion and Discussion

This study reports a fully automatic autism detection method using the LENA system. By combining the DLP hardware and the speech signal processing software, child vocalization samples can be collected in an automatic and unobtrusive way in the natural home environment with large quantity. This capability, by itself, has great potential for many research areas and general applications. Specifically in this study, such large quantity of naturalistic young child vocalization data makes it possible to test child vocalization composition. For the first time, it is shown that children with autism are significantly different from other children in terms of vocalization composition at the phone-level. Quantitative models for autism detection have been built up using pattern recognition and machine learning approaches. The fully automatic method described here performs well. The following are some points for conclusion and discussion:

- Compared with clinical observation, this method is unobtrusive and objective.
- It is repeatable whereas human observation may not.
- It is capable of incorporating large quantity of data while human observation is usually limited in samples.
- It is capable of dealing with large number of variables and their joint effects while human observation may be limited to only a few variables and joint effects may be difficult to observe.
- It is a data-driven method while observational approach may start from theory or intuition or human intelligence.
- The automatic method is associated with machine error (e.g. segmentation error, phone recognition error) while human observation may be affected by subjectivity.
- Machine error may be compensated for in part by increased sampling while human intelligence can make full use of a more limited number of samples in observation. Our experience and preliminary analyses showed the trend toward machine error compensation via increased sampling. More rigorous analyses and experiments will be done and reported in the future.
- Future directions may include modeling of other types of information in recording, such as social interaction, emotion, prosody, etc. Combining with other existing screening instruments may be important. More data is necessary to generate more rigorous tests and more robust and detailed modeling for better performance.

Acknowledgment

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

References

[1] Archive of Abstracts of International Meeting for Autism Research (IMFAR) http://www.autism-insar.org

[2] J. Singh, J. Illes, L. Lazzeroni, J. Hallmayer, "Trends in U.S. Autism Research", 77, 146.2, 7th Annual International Meeting for Autism Research (IMFAR), May 2008, London

[3] Susan E. Bryson, Sally J Rogers, Eric Fombonne "Autism Spectrum Disorders: Early Detection, Intervention, Education, and Psychopharmacological Management", Canadian Journal of Psychiatry, Vol 48, No 8, Sept. 2003

[4] G. Dawson, J. Osterling, "Early Intervention in Autism", in "*The Effectiveness of Early Intervention*", M. Guralnick (Ed.), Baltimore: Brookes, 1997

[5] American Academy of Pediatrics, Council on Children With Disabilities; Section on Developmental Behavioral Pediatrics; Bright Futures Steering Committee; Medical Home Initiatives for Children With Special Needs Project Advisory Committee. "Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening." Pediatrics. 2006; 118: 405-20

[6] C. Johnson, S. Myers and the Council on Children with Disabilities, "Identification and Evaluation of Children with Autism Spectrum Disorders", Pediatrics, Vol 120, No 5, Nov. 2007

[7] T. Charman, G. Baird "Practitioner Review: Diagnosis of Autism Spectrum Disorder in 2- and 3-year-old Children", Journal of Child Psychology and Psychiatry 43:3 (2002), pp 289-305

[8] A. Klin, D. Lin, P. Gorrindo, G. Ramsay, W. Jones "Tow-year-olds with autism orient to non-social contingencies rather than biological motion" Nature advance online publication 29 March 2009

[9] S. Ozonoff, S. Macari, G. Young, S. Goldring, M. Thompson, S. Rogers, "Atypical object exploration at 12 months of age is associated with autism in a prospective sample" Autism, Vol 12, No 5 pp 457-472, 2008.

[10] L. Zwaigenbaum, S. Bryson, J. Brian, W. Roberts, P. Szatmari, B. Mackinnon, S. Mitchell "Early Language Impairments in High-Risk Infants Subsequently Diagnosed with Autism" S4.9, 4th IMFAR, May 2005, Boston

[11] A. Nadig, S. Ozonoff, G. Young, S. Macari, S. Rogers, M. Sigman, A. Rozga "Response to name in 12-month-old siblings of children with autism or typical development" P3B.1.8 4th IMFAR, May 2005, Boston

[12] Joanne McCann, Sue Peppe, "Prosody in Autism Spectrum Disorders: a Critical Review", International Journal of Language & Communication Disorder, Vol 38, No 4, Oct-Dec, 2003

[13] J. Diehl, D. Watson, J. McDonough, C. Gunlogson, E. Young, L. Bennetto "Acoustic and Perceptual Analysis of Prosody in High-Functioning Autism" P1B.2.8, 4th IMFAR, Boston, May 2005

[14] E. Prud'hommeaux, J. Van Santen, R. Paul, L. Black "Automated measurement of expressive prosody in neurodevelopmental disorders" 31 154.31, 7th IMFAR, May 2008, London

[15] A. Bosseler, D. Massaro, "Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning in Children with Autism" Journal of Autism and Developmental Disorders, Vol 33, No 6, December, 2003

[16] E. Kim, E. Newland, R. Paul, B. Scassellati "Robotic Therapist for Positive, Affective Prosody in High-Functioning Autistic Children" 6.114.6, 7th IMFAR, May 2008, London

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

http://www.lenababy.com/

[18] 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. Also available: http://www.lenafoundation.org/DownloadFile.aspx/pdf/SignalProcessing_ChildSpeech

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

[20] R. Duda, P. Hart, "Pattern Recognition and Scene Analysis", A Wiley-Interscience Publication, New York Wiley, 1973

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