Fundamental Frequency of Child-Directed Speech Using Automatic Speech Recognition

Mark VanDam
Department of Speech & Hearing Sciences
College of Medical Sciences
Washington State University
Spokane, WA, USA
mark.vandam@wsu.edu

Paul De Palma
Department of Speech & Hearing Sciences
College of Medical Sciences
Washington State University
Department of Computer Science
Gonzaga University
Spokane, WA, USA

***Abstract*—It has long been recognized that speech directed to infants and toddlers—child-directed speech (CDS) or motherese—is characterized by a collection of features including hyperarticulation, a distinctive lexicon, reduced structural complexity, and increased fundamental frequency (f_0). This paper examines f_0 in nearly 500 hours of recordings from 33 families, showing that mothers, but not fathers, distinguish CDS with consistently higher f_0. We also found that, compared with parents of typically-developing children, parents of children who are hard-of-hearing do not differ in production of f_0 to their children. Results are relevant for improved ASR application and better understanding of mothers' and fathers' speech to their children.**

Keywords—automatic speech recognition, motherese, child directed speech, LENA

I. Introduction

If we think of think of soft computing as addressing that collection of problems whose solutions are probabilistic in nature, then one of its real success stories is the application of Bayesian techniques to automatic speech recognition (ASR) over the past two decades. Though work in ASR dates to the earliest days of computing, significant breakthroughs did not occur until the adaptation of probabilistic machine learning techniques to the problem of transcribing continuous speech to text. In fact, ASR has begun to assume the characteristics of what Thomas Kuhn has called "normal science," that is "research firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges for a time as supplying the foundation for further practice" [11, p. 10]. In Kuhn's sense of past achievements, here we refer to Bayesian inference, in particular, the hidden Markov model (HMM).

The success of soft-computing approaches to ASR has opened speech recognition to applications far beyond what the pioneers in modern ASR might have imagined. Here, for example, is Frederick Jelinek on the very first page of his classic text: a "speech recognizer … can be thought of as a voice-actuated 'typewriter' in which a computer program carries out the transcription and the transcribed text appears on the word display" [8, p. 1]. True enough, but this formulation surely understates what humans do when they converse. By broadening the definition of ASR—in particular, a written transcript fails to capture important aspects of human speech communication, including prosodic features conveyed by f_0—we can not only hope to improve its accuracy [4], but also use it to investigate fundamental questions about human language itself [5].

This is where LENA (Language ENvironment Analysis; LENA Research Foundation, Boulder, CO, USA) enters the picture [12], [13]. One of the more serious constraints faced by researchers into child language acquisition is the expense and difficulty of hand-transcribing and classifying recordings of speech. LENA applies modern ASR and other techniques to day-long acoustic recordings collected in a natural familial setting. This introduction of soft computing techniques to language research dramatically increases the amount of conversational data that may be investigated. This study, for example, consists of 491.2 hours of recorded speech, a volume that would have been difficult to manage even a decade ago.

In this paper we illustrate the extraordinary advances of soft computing techniques with an investigation of the phenomenon known as child-directed speech (CDS), the distinctive way in which parents talk to their infants and toddlers, and certain well-known speech differences when talking to listeners with perceived language or hearing deficiencies, such as when talking with a person with hearing loss. We limit ourselves here to a single parameter—parent production of f_0—and examine if we can predict the use of CDS based on the sex of the parent or the hearing status of the child through two research questions. First, do fathers and mothers use f_0 differently during CDS? Second, do parents of hard-of-hearing (HH) children use f_0 to mark CDS differently from parents of typically developing (TD) children?

II. From ASR to LENA

Modern ASR treats speech, the perturbations of air molecules by the human vocal apparatus, as a noisy channel. In this view, an acoustic waveform is a noisy version of a word string produced by a talker. The task of ASR is to develop a probabilistic mapping from the waveform to the string. Suppose O is an acoustic signal and S a word string, each can be decomposed, O into regular acoustic observations, S into some convenient linguistic unit (such as the written text):

$$O = o_1, o_2 \ldots, o_n$$

$$S = s_1, s_2 \ldots, o_n$$

Since S is conditionally dependent upon O and we are looking for the most probable S given O, Equation 1 establishes the relationship:

$$hyp(S) = \frac{max}{S \in L} P(S|O) \quad 1$$

This can be read, "the hypothesized word string, S, is the most probable, among all possible legal word strings in the language L, given the acoustic observation, O." We can use Bayes' Rule to transform Equation 1 to Equation 2:

$$hyp(S) = \max_{S \in L} \frac{P(O|S) * P(S)}{P(O)} \quad 2$$

Since the denominator does not change among candidate word strings in any given language, it can be eliminated, leaving:

$$p(S) = \max_{S \in L} P(O|S) * P(S) \quad 3$$

The well-known Bayesian terms, *likelihood* and *prior probabilities* (P(O|S) and P(S)) are known, in the world of ASR, as the acoustic and language models. P(O|S) is the probability of an acoustic signal given a word string. P(S) is the probability that a word string would be uttered by a speaker of the language under observation.

Modern speech recognizers, using standard digital signal processing (DSP) techniques, typically construct sets of 39 feature vectors, known as mel frequency cepstrum coefficients, from periodic samples of the acoustic waveform. The acoustic model is computed from the feature vectors. It represents the likelihood of observed feature vectors given something we cannot observe in an acoustic signal, namely a word string. The probabilistic relationship between an observed sequence—in this case feature vectors—and something that is not observed, i.e., a word strong, is represented using HMMs, a machine learning technique found throughout modern ASR [3], [9], [7]. Viewed this way, speech recognition is an instance of a generalized classification task: map subcomponents of an acoustic signal into the word (or phone or subphone) classes where they best fit.

One of the hurdles traditionally faced by researchers into human speech has been both the expense of obtaining data and the difficulty of transcribing and classifying data once it has been obtained. For example, in the mid-nineties, Hart & Risley [6] famously argued that the single most important predictor of future academic success—more important than race, more important than ethnicity, more important than socioeconomic status—is the amount of conversation parents have with their young children. This study, like all other experimental work in child language acquisition, was constrained by the effort involved in recording, transcribing, and classifying data. Hart & Risley, for example, studied only 42 children for about an hour a month over a three year period. LENA was developed to reduce the cost of obtaining this kind of data by automating the classification and tagging of speech segments [5].

LENA allows researchers to collect and tag language and development data from children from 2 months to 4 years of age. It consists of an acoustic recording device worn by children in specially designed vests and a software suite that performs various DSP, ASR-related, and statistical tasks. Similar to an ordinary speech recognizer, the system transforms the time-stamped audio stream into a collection of feature vectors that it segments and labels at centisecond resolution. But with LENA the labels are not words (or phones) but a collection of about 60 *a priori* labels indicating the provenance of the sound. These labels include *key child*, *other child*, *adult male*, *adult female*, *overlapping sound*, *electronic sound*, etc. Labeled segments can further be concatenated into a number of "vocalization activity blocks," such as *key-child-conversing-with-adult*, *female-adult-monologue*, etc. Summary reporting includes estimates of total conversational turns and vocalization duration, among many other details [13], [17].

LENA employs segmentation and labeling techniques familiar to researchers in automatic speech recognition and machine learning including HMMs and probabilistic Gaussian mixture models. In addition, some rule-based techniques are used in classification. For example, spectral acoustic energy levels above some pre-set threshold may be used to distinguish a non-speech cry from a speech-like child vocalization [17].

Finally, it is well-known that task-related factors affect the accuracy of ASR. The presence of noise, the use of an unconstrained vocabulary, and conversational as opposed to read speech all make the ASR task more difficult and reduce the accuracy of classification. The LENA approach, however, knowingly trades these difficulties for ecologically valid, natural recordings of families *in situ*—a novel advance. LENA is designed to eliminate signal types that might result in false positives, namely vegetative or non-speech vocalizations of children, overlapping speech, and faint speech. Several reports show a mean agreement of 76.25% over four categories between the LENA software and human transcribers [16], [23], [24], [26], [27]. This is roughly in line with the state of the art standard ASR systems [3], [19]. We are conducting our own reliability study of the LENA system over a subset of the data described below.

III. CHILD-DIRECTED SPEECH

Anyone who has been around infants and toddlers recognizes the distinctive features of CDS. Its components include shorter utterances, hyper-articulation, decreased linguistic complexity, along with the prosodic feature we are most interested in, increased f_0. CDS has been extensively attested in several European languages as well as Japanese, suggesting to some scholars that not only might it be found universally, but it might also serve an important function in language acquisition [15].

CDS can be viewed as a special case of the Lombard effect, which is the tendency for talkers to produce systematically altered speech in the presence of perceived noise or listener deficiency [14]. Anyone who has eaten in a crowded, noisy restaurant or spoken with someone wearing headphones has experienced the Lombard effect. In one formulation, "the tendency to produce motherese [i.e., CDS] may thus involve the same kind of unconscious adjustment to speech that is necessary to communicate with the listener" [10, p. 156]. To illustrate the promise of ASR in arenas outside those typically associated with speech recognition, we have used the LENA system to investigate whether fathers as well as mothers use similar speech patterns with their toddlers.

Another well-known special case of the Lombard effect is speech to HH listeners, in which talkers make predictable speech adjustments to listeners known (or assumed) to be hearing-impaired. In the present work, we compare the CDS from parents who have children who are HH with parents of TD children who have normal hearing. It is possible that parents of children who are HH, compared with parents of TD children, produce increased f_0 while speaking to their children.

IV. MATERIALS AND METHODS

CDS can be described acoustically (change in f_0), lexically (specialized vocabulary), phonologically (hyperarticulation), and syntactically (reduced structural complexity), among other ways. We have confined our investigation to the first parameter, variation in f_0, because it can be defined and observed objectively. Since the pioneering work of Gunnar Fant in the 1960s, speech scientists have viewed the vocal tract as an acoustic filter that alters the waveforms generated through vocal fold vibration. Vocal fold vibration produces complex but periodic waveforms which can be analyzed through Fourier analysis. The lowest frequency component of the complex waveform is the f_0. We can now restate our hypotheses with precision:

1) Mothers and fathers will produce higher mean f_0 during CDS than non-CDS. 2) Mothers and fathers of HH children will produce higher f_0 during CDS than mothers and fathers of TD children.

To investigate these hypotheses, nearly five hundred hours of familial speech were recorded using the LENA recording device and labeled with the associated ASR classification software [17], [25], [27]. The mean f_0 of all speech segments labeled *adult-male* and *adult-female* was then computed using software described in [18] and [19]. Adult speech was distinguished from CDS by context. A speech segment labeled, for example, *adult-male*, and found adjacent to another adult segment was defined as adult-directed, or non-CDS. An adult segment adjacent to a speech segment labeled *key-child* was defined as CDS. Table I shows the study details for HH and TD families.

Table I. Participants & Materials

Participants	22 families with TD children 11 families with HH children
Sex	51.5% male
Child Age	M_{all} = 30.1 months (SD=2.3 months) M_{TD} = 31.1 months (SD=2.3 months) M_{HH} = 29.4 months (SD=1.9 months)
Secondary Disability	None, by parent report
Data	Unprocessed whole-day recordings (single channel, 16kHz, 16-bit, PCM) using LENA wearable recorders
Time	491.2 total hours, 14.8 mean hours/fam
Coding	Target child, adult-female, adult-male, other-child, other (electronic, overlappi vocals, silence, noise)
Software	1) LENA Software for coding 2) software to compute f_0 [18], [19] 3) Custom-built software to compute statistics

V. RESULTS

The results of the study are shown in Tables II through VII and graphically in Fig. 1, where f_0 of CDS is on the ordinate and f_0 of non-CDS is on the abscissa for both mothers and fathers, with an equal-f_0 bisector also shown. An observation on the bisector indicates equal f_0 between CDS and non-CDS speech, above the bisector indicates increased f_0 in the CDS condition, and below the bisector indicates increased f_0 in the non-CDS condition. Parents of TD and HH children are individually indicated in the figure.

As expected, mothers' mean f_0 was significantly higher than fathers' ($t_{(32)}$=34.3, $p<10^{-26}$; see Table II). Fig. 1 and Table III show that mothers have significantly increased f_0 in the CDS condition ($t_{(32)}$=18.6, $p<10^{-18}$), but fathers do not show difference in f_0 production by condition ($t_{(32)}$=.55, p>.5).

Production of f_0 between the TD and HH groups was not found to be different for mothers ($t_{(64)}$=1.33, p>.1), but HH fathers showed significantly greater f_0 than the TD fathers ($t_{(64)}$=4.62, $p<10^{-4}$; see Tables IV and V). That is, in their production of f_0, mothers were not sensitive to the hearing status of their children, but HH fathers had a higher f_0 than TD fathers.

Finally, mothers of TD and HH children showed a significant increase in f_0 for CDS speech ($t_{(10)}$=10.0, $p<10^{-6}$ and $t_{(21)}$=16.1, $p<10^{-12}$, respectively; see Table VI). Neither TD nor HH fathers showed a difference in f_0 for CDS speech ($t_{(10)}$=.39, p>.5 and $t_{(21)}$=.41, p>.5, respectively; see Table VII).

Table II. f_0 of Mothers and Fathers

	M (Hz)	*SD* (Hz)
Mothers	230.5	22.3
Fathers	173.9	12.5

Table III. f_0 of all Mothers and Fathers

	M (Hz)	*SD* (Hz)
$\text{Mothers}_{\text{CDS}}$	250.2	13.3
$\text{Mothers}_{\text{non-CDS}}$	210.8	5.6
$\text{Fathers}_{\text{CDS}}$	174.4	13.4
$\text{Fathers}_{\text{non-CDS}}$	173.5	11.9

Table IV. f_0 of TD & HH Mothers

	M (Hz)	*SD* (Hz)
TD	225.4	21.1
HH	233.1	22.6

Table V. f_0 of TD & HH Fathers

	M (Hz)	*SD* (Hz)
TD	165.1	15.0
HH	178.3	8.3

Table VI. f_0 of TD & HH Mothers by Condition

	M (Hz)	*SD* (Hz)
TD_{CDS}	242.9	14.7
$\text{TD}_{\text{non-CDS}}$	207.8	6.0
HH_{CDS}	253.8	11.2
$\text{HH}_{\text{non-CDS}}$	212.1	4.2

Table VII. f_0 of TD & HH Fathers by Condition

	M (Hz)	*SD* (Hz)
TD_{CDS}	164.7	15.5
$\text{TD}_{\text{non-CDS}}$	165.6	15.3
HH_{CDS}	177.9	6.3
$\text{HH}_{\text{non-CDS}}$	178.8	10.0

VI. CONCLUSION AND CURRENT RESEARCH

With respect to the hypotheses developed in section IV: 1) mothers, but not fathers, increased f_0 when talking to their toddlers; and 2) fathers of HH children had a higher f_0 than fathers of TD children, but mothers of neither HH children nor TD differed significantly in f_0.

This is the first study of its kind to use soft computing methods on a very large, ecologically-valid database of hundreds of hours of recorded family speech. By looking at well-known phenomena of CDS and a special case of the Lombard effect using modern methods, we are able to explore aspects of human speech as never before.

This study accomplished two main things. First, it is proof-of-concept evidence that soft computing methods can be profitably applied to the study of human language and language development on a massive scale. Second, the

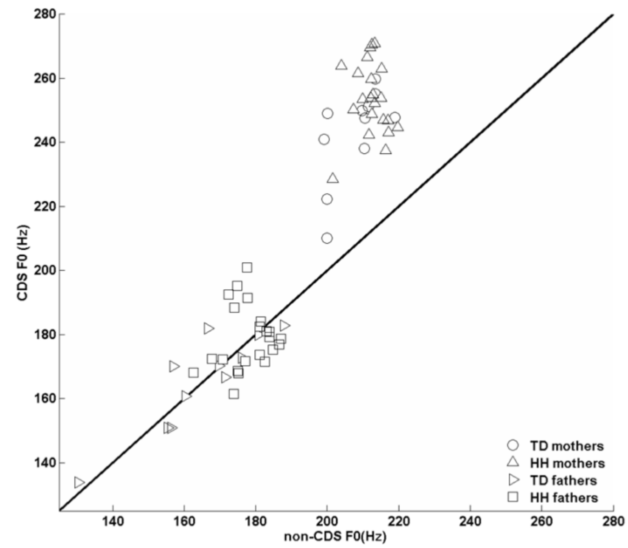

Fig 1. f_0 of CDS and non-CDS for mothers and fathers of TD and HH children.

principal findings reported here have not, to our knowledge, been reported elsewhere. Namely, the broad finding that mothers and fathers differ in their speech strategies when talking to children has received very little attention in the literature. The details of this finding could be profitably exploited by certain ASR (or DSP) approaches used on real speech to either improve machine performance or quantify certain error types. Another broad finding here is that parents do not systematically alter f_0 when directing speech to HH children as compared with TD children. This result is somewhat surprising given well known talker effects based on assumed or perceived listener qualities, but may reflect the limitations of the Lombard effect in certain communicative contexts or may reflect characteristics of a population of HH children that are not well understood. For example, HH children in the present sample are the beneficiaries of modern intervention strategies [28], [29] that may render them a different sample from what has been previously described in the literature.

To date, we have collected more than 10,000 hours of in-family speech using the LENA recording device, a subset of which is reported here. We are currently examining other family dynamics including child speech production data and the role of conversational interactions in both TD and HH preschoolers.

The essential fact is that work like ours would not have been possible even a decade ago. The availability of inexpensive computing and data storage devices along with the development of modern probabilistic and machine-learning techniques—all falling under the umbrella of soft computing—has shifted what is possible for investigators into child speech among both TD and disordered populations.

For instance, LENA ASR technology has been used to objectively characterize the vocalizations of TD and at-risk populations across large data sets [2], [5], [16], [22], [24]. Preliminary results have demonstrated the utility of this particular ASR application, and suggest the possibility that

ASR used in conjunction with other parameters (e.g., disability status, sex, age, spectral and temporal vocal characteristics, engagement in conversation, etc.) might improve our understanding of language acquisition, child speech, and family vocal behavior [16], [21]. The resulting classification models hold promise for early identification and intervention. Some preliminary results have been able to successfully classify children with autism spectrum disorder and language delay from TD children based on ASR analyses of audio recordings alone [16].

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health Grants NIH/NIDCD 5 R01 DC009560 (co-principal investigators, J. Bruce Tomblin, University of Iowa and Mary Pat Moeller, Boys Town National Research Hospital). The content of this project is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NIDCD. This work was supported by the WSU Spokane Seed Grant Program. We thank the families who contributed, and the children and families of the HOPE School of Spokane.

REFERENCES

[1] Ambrose, S. E., VanDam, M., & Moeller, M. P. 2014. "Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss." Ear and Hearing, vol. 35(2), pp. 139-147.

[2] Caskey, M., Stephens, B., Tucker, R., & Vohr, B. 2011. "Importance of parent talk on the development of preterm infant vocalizations." Pediatrics, vol. 128(5), pp. 910-916.

[3] De Palma, P. 2014. "Probabilistic methods in automatic speech recognition." In M. Khosrow-Pour (ed.), Encyclopedia of Information Science and Technology. Hershey, PA: IGI Global.

[4] De Palma, P., & Wooters, C. 2012. "Automatic speech recognition with syllables and concepts." Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems and the 13th International Symposium on Advanced Intelligent Systems, Kobe, Japan, November, 2012.

[5] Gilkerson, J., & Richards, M. 2008. "The power of talk: impact of adult talk, conversational turns, and TV during the critical 0-4 years of child development." Retrieved 7/17/2014 from: www.lenafoundation.org/TechReport.aspx/PowerOfTalk/LTR-01-2.

[6] Hart, B., & Risley, T.R. 1995. Meaningful Differences in the Everyday Experience of Young American Children. Baltimore: Paul H. Brookes Publishing Co., Inc.

[7] Huang, X., Acero, A., & Hsiao-Wuen, H. 2001. Spoken language processing: a guide to theory, algorithm, and system development. Upper Saddle River, NJ: Prentice Hall.

[8] Jelinek, F. 1997. Statistical Methods for Speech Processing. Cambridge: MIT Press.

[9] Jurafsky, D., & Martin, J. 2009. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice Hall.

[10] Kuhl, P., Feng-Ming, T, Liu, H., Zhang, Y, & De Boer, B. 2001. "Language/culture/mind/brain: progress at the margins between disciplines. Annals of the New York Academy of Sciences, vol. 935, pp. 136-174.

[11] Kuhn, T. 1970. The structure of scientific revolutions. Chicago: The University of Chicago Press.

[12] LENA 2014. LENA Research Foundation. Retrieved 7/17/2014 from: www.lenafoundation.org.

[13] LENA 2011. "The LENA advanced data extractor (ADEX) user guide version 1.1.2." Retrieved 7/21/2014 from: www.lenafoundation.org/DownloadPath.aspx/pdf/public/ADEX/v1.1.2/The_LENA_ADEX_User_Guide.

[14] Lombard, E. 1911. "Le signe de de l'elevation de la voix." Annales Des Malades de l'oreille, Larynx, Nez, Pharynx, vol. 37(2), pp. 101-119.

[15] Masataka, N. 1999. "The role of modality and input in the earliest stage of language acquisition: studies in Japanese sign language." In J. Morford, R. Mayberry (eds.), Language Acquisition by Eye, 24. New York, New York: Psychology Press, Taylor & Francis Group.

[16] Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D. & Warren, S. F. 2010. "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development." Proceedings of the National Academy of Sciences, vol. 107(30), pp. 13354-13359.

[17] Paul, T., Xu, D., & Richards, A. 2014. "System and method for expressive language assessment." Patent Number US 8844847 B2, retrieved 7/21/2014 from: http://www.google.com.ar/patents/US874484

[18] Sun, X. 2002. The Determination, Analysis, and Synthesis of Fundamental Frequency. Unpublished doctoral dissertation, Northwestern University.

[19] Sun, X. May, 2002. "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio." Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, USA.

[21] VanDam, M., Moeller, M. P., & Tomblin, B. 2010. "Analyses of fundamental frequency in infants and preschoolers with hearing loss." Poster presented at the 160th Meeting of the Acoustical Society of America, Cancun, Quintana Roo, Mexico, November, 2010.

[22] VanDam, M., Ambrose, S. E., & Moeller, M. P. 2012. "Quantity of parental language in the home environments of hard-of-hearing 2-year-olds." Journal of Deaf Studies and Deaf Education, vol. 17(4), pp. 402-420.

[23] VanDam, M., & Silbert, N. H. 2013. Precision and error of automatic speech recognition. Proceedings of Meetings on Acoustics, 19: 060006. doi:10.1121/1.4798466.

[24] Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. 2010. "What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism." Journal of Autism and Developmental Disorders, vol. 40(5), 555-569.

[25] Xu, D., Yapanel, U., Gray, S., & Baer, C. 2008. "The LENA language environment analysis system: the interpreted time segments (ITS) file." Retrieved 7/21/2014 from: http://www.lenafoundation.org/TechReport.aspx/ITS_File/LTR-04-2.

[26] Xu, D., Yapanel, U., Gray, S., & Baer, C. 2009. "Reliability of the LENA language environment analysis system in young children's natural environment." Retrieved 7/21/2014 from: http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2.

[27]. Xu, D., Richards, J. A., & Gilkerson, J. 2014. "Automated Analysis of Child Phonetic Production Using Naturalistic Recordings." Journal of Speech, Language, and Hearing Research, ahead of print, retrieved May 14, 2014 from: http://jslhr.pubs.asha.org/article.aspx?articleid=1872911.

[28] Yoshinaga-Itano, C., Baca, R. L., & Sedey, A. L. 2010. "Describing the trajectory of language development in the presence of severe to profound hearing loss: A closer look at children with cochlear implants versus hearing aids." Otol Neurotol, vol. 31(8), pp. 1268.

[29] Yoshinaga-Itano, C., Sedey, A.L., Coulter, D.K., et al. (1998). The language of early- and later-identified children with hearing loss. Pediatrics, vol. 102(5): pp. 1161–1171.