**Abstract—Linguists and developmental psychologists have long-recognized distinctive speech patterns in adults speaking to infants and toddlers. Referred to as child-directed speech, these features include hyper-articulation, a distinctive lexicon, reduced structural complexity, and increased fundamental frequency (*f0*).**

**In a previous paper [1] we used increased *f0* as a proxy for motherese. Though to our knowledge the principle findings presented in that paper—that mothers, but not fathers—raise *f0* when speaking to their infants and toddlers had not been presented elsewhere, our intention in extracting from nearly 500 hours of recorded speech was to demonstrate soft computing techniques could be applied to areas that previously been the province of controlled laboratory experiments and qualitative ethnographic research. The current paper is another proof of concept but on a scale of over an order of magnitude larger than that reported in [1]. We argue in this paper that 1) fathers but not mothers raise *f0* when speaking to their infants and toddlers; and 2) that soft computing techniques can be used to investigate linguistic data on a scale inconceivable only a decade ago. Specifically, we use automatic speech recognition techniques embodied in software from the LENA research foundation to classify over 7,000 hours of recorded speech, and our own software to locate nearly four million individual instances of child-directed speech from which we extract *f0* and analyze results**

**Keywords—child-directed speech, motherese, fundamental frequency**

I. Introduction

The decades of the 1950s, 1960s, and 1970s, in tandem with the cognitive revolution, saw serious research into language development in children (cf. [3], [4], [5], [6]). Data collection, whether through field observation or controlled laboratory experiment, was time-consuming and expensive. The largest part of early and subsequent research into language acquisition depended upon controlled experiments or interactions in a laboratory among parents and children. The sample sizes were small and crucially dependent upon trained transcribers, who both err and bring their own biases to the observations. Perhaps the strongest objection to laboratory investigation into child speech concerns ecological validity. Beyond the small sample, how can we be sure that what we find in the laboratory has not been altered by the setting?

Cost places limits on both approaches. In the mid-nineties Hart and Risley [8] argued that the number of hours of conversation parents have with their children is the strongest predictor of future academic success, stronger than any of the usual contenders, race, ethnicity, and socioeconomic status. The constraints under which Hart and Risley labored would be familiar to just about any developmental psychologist or field linguist, namely the expense of collecting, transcribing, and classifying data. Hart and Risley studied only 42 children for an hour each month over three years.

At the same time that the cognitive revolution was encouraging researchers to approach language as a computational process in the mid-fifties, researchers began to investigate the use of computers to transcribe speech, a field that has come to be known as automatic speech recognition (ASR). If soft computing can be construed as addressing that set of problems whose solutions are probabilistic in nature, ASR is one of its genuine successes, with error rates for large vocabulary speech recognition systems dropping dramatically since the introduction of Bayesian inference techniques in the 1990s and, most recently, neural networks [11]. The LENA (Language Environment Analysis) Research Foundation, by applying modern ASR and data analysis techniques to day-long acoustic recordings of children at home has made it possible to take an ethnographic approach to language data collection. We noted in a preliminary paper [1, p. 1349] that our data consisted of “491.2 hours of recorded speech, a volume that would have been difficult to manage even a decade ago.” The data set for the current study comprises over 7,000 hours of recorded speech, a volume that would not have been just difficult to manage a decade ago, but impossible to conceive.

In this paper, we use the phenomenon of child-directed speech (CDS) to illustrate the extraordinary advances in soft-computing. CDS is the well-known and well-attested manner in which mothers’ (the choice of gender is intentional) speak with their infants and toddler. Though CDS can be characterized lexically, syntactically, and pragmatically, we confine ourselves to a single parameter, the parents’ vocal fundamental frequency (f0), a parameter that can be extracted from the speech stream and analyzed by computer. This last is important, since it implies an objective measure rather than fuzzier impressions of, for example, reduced syntactic complexity. We ask a single question (does the CDS of fathers, using the proxy of raised *f0* differ from that of mothers) and in the process show that soft computing techniques can be used to process over 7,000 hours of recorded speech and over 3,000,000 individual instances of CDS speech from sixty-two families.

II. CHILD LANGUAGE RESEARCH, AUTOMATIC SPEECH RECOGNITION, AND LENA

Samples of children’s speech are usually collected in the laboratory or during visits to their homes. Researchers try to elicit speech through games, questions, and simple tasks. Researchers making recordings in this way find themselves in a double bind. If the recordings are made in the laboratory under formal scientific protocols, the samples are necessarily small and decontextualized, by definition. On other hand, recordings made in the home, though somewhat more ecologically valid, are costly to obtain. Once again, the sample size is necessarily small. LENA (Language Environment Analysis), developed at the LENA Research Foundation in Boulder, CO, USA was developed to solve the problems of cost, ecological validity, and bias by removing the human component from the data collections and coding process through the use of automatic speech recognition.

Since all conference attendees may not be familiar with ASR and since we argue that the adoption of ASR has changed the landscape of child language research, we offer a short introduction to ASR. For a more technical introduction, see [9], [16], or [17]. Speech is the perturbation of air molecules by the human vocal apparatus. Modern ASR treats speech as a noisy version of an idealized speech string intended by the speaker. ASR, in essence, produces a probabilistic mapping from the acoustic signal to the speech string. It does this through familiar Bayesian inference techniques. Suppose an acoustic signal, O, and a word string S, represent a sequence of acoustic observations, o1, o2, … on and a sequence of words, s1, s2, … sn, and L a language, we can state the speech recognition problem as the conditional probability found in Equation 1:

G(S) = max(P(S|O)) s.t. S ∈L 1

Equation 1 is read, “G(S) is the most probable word string, among all candidate word strings, S, given acoustic observation O and such that S is a legal string in the language. Invoking Bayes’ rule this becomes:

G(S) = s.t. S ∈L 2

Since the acoustic observation does not change for candidate word strings, equation 2 becomes:

G(S) = max(P(O|S) \* P(S) s.t. S ∈L 3

In the language of ASR, the first term on the right-hand side—the likelihood—is known as the *acoustic model*. The second term—the prior—is known as the *language model*. Modern speech recognizers, using standard digital signal processing techniques, extract feature vectors from periodic samples of an acoustic waveform. These are then probabilistically mapped to speech units, usually triphones, a term that deserves some explanation. Each language has its own inventory of phones, where a phone is a speech unit. Typical English phones include word initial *p* in pan, known as a *bilabial stop* and produced by closing the lips and then releasing the air that has built up behind them. A triphone is a phone with its left and right sub-phonetic contexts. Its use is an attempt to model co-articulation, the property exhibited in the English vowel [eh], which may produce a somewhat different set of acoustic features, depending on whether it appears in *wed*, *yell*, or *Ben* [16]. Taken together, the feature extraction and subsequent statistical mapping, known as Gaussian Mixture Models, allow us to express the likelihood of an acoustic observation given a word string.

At a slightly higher level, the probabilistic relationship between something that is observed—here an acoustic signal—and something not observed, here a word string, can be modeled using an HMM, a machine learning technique that, as much as anything else, has been responsible for the success of ASR in the past two to three decades [16], [17]. Viewed this way, automatic speech recognition is an instance of generalized classification: place subcomponents of the acoustic signal into the word (or phone or subphone) bucket where they best fit. As we will shortly see, the LENA system classifies speech signals but, instead of classifying them into words, it groups them by conversational role in the language of infants, toddlers, and their parents.

The LENA system has allowed us to collect and label over 7,000 hours of speech data from infants and toddlers—two months to four years—and their parents. There are two components. An acoustic recording device and a suite of software that performs digital signal processing and classification tasks. The LENA recorder weighs less than 60 grams, holds up to 16 hours of audio recordings, and is designed to be worn in a specialized vest, a toddler’s bib overalls, or an infant’s onesie. LENA software, like most ASR systems until very recently, uses Gaussian Mixture Models (GMM), but with a crucial difference. Time-stamped audio streams are transformed into feature vectors that are segmented and labeled at centisecond resolution. With LENA, however, the labels are not words but over sixty categories that indicate the source of the sound. These include *key child*, *other child*, *adult male near*, *adult female near*, *overlapping sound*, and *electronic sound*. Labeled speech segments are grouped into “vocalization activity blocks,” like *key-child-conversing-with-adult*, *female-adult-monologue*, plus many others [13], [18]. In addition, to hidden Markov Models and Gaussian mixture models, LENA employs rule-based classification techniques. For instance, spectral acoustic energy that exceeds a pre-determined threshold is used to distinguish non-speech-like crying from speech-like child vocalization [18].

As we noted in our previous paper [1] and is well-understood in the ASR community (c.f., [16]), the environment in which speech is collected can dramatically affect accuracy. Ambient noise, an unconstrained vocabulary, conversational as opposed to read speech—all characteristics of the environments in which LENA is intended to be used—affect classification accuracy. Unlike generalized ASR, LENA is specifically designed to eliminate false positives, namely non-speech vocalizations and indistinct speech. Several studies show a mean agreement of 76.25% between LENA and human transcribers [20], [21], [22], [23]. This is consistent with standard ASR systems [9], [24]. More recently, Van Dam used a four-alternative forced-choice task with 24 judges and 2,340 hours of LENA-labeled speech data. The agreement—more precisely the Fleiss kapp reliability coefficient—between the coders and LENA was found to be 79%, again, consistent with standard ASR systems.

III. CHILD-DIRECTED SPEECH

Child-Directed Speech (CDS), often called *motherese,* is the collection features frequently found in adult speech to infants and toddler. These include higher pitch, exaggerated articulation, a distinct lexicon, and decreased linguistic complexity. CDS has been attested in Japanese and several European languages. One study showed that the forty-eight infant subjects preferred the speech register commonly associated with CDS. Another demonstrated that infants prefer the distinctive prosody of motherese and that this distinctive prosody corresponds to clausal boundaries. These and other results have led some researchers to argue that CDS plays a role in language acquisition [25], [26], [27]. Indeed, at least one study implicates CDS in the evolution of language itself [28].

Since the break-through research of Gunnar Fant in the 1960s, linguists have modeled the vocal tract as an idealized acoustic filter that modulates the waveforms generated by vocal fold vibrations. These vibrations produce complex and periodic waveforms that can be decomposed through Fourier analysis. The lowest frequency component of the vocal waveform is called fundamental frequency or *f0*. Said another way, *f0*is the first harmonic of the speech signal. The term *pitch* usually refers to what the listener perceives as opposed to fundamental frequency which is what the talker produces. Since our data consists of digitized speech, and since the two correlated [34] we use the terms *f0* and pitch interchangeably and use fundamental frequency as a proxy for motherese. It is important to point out that motherese, is not our primary interest. Nor is fundamental frequency. We might just have easily extracted phrase duration, amplitude, *f2* or any among many acoustic correlates. This paper is a proof of concept. It shows that soft computing techniques along with very large data collections can be used to solve problems that have bedeviled linguists and developmental psychologists for forty years, namely ecological validity and the cost of data collection and coding. Because our speech data was recorded using the now available inexpensive digital recording and storage devices and analyzed using machine learning techniques, we have been able to examine 7000 hours of speech as opposed the nine hours reported in [31].

V. MATERIALS AND METHODS

CDS can be described syntactically (reduced complexity), phonologically (hyperarticulation), lexically (specialized vocabulary), acoustically (raised pitch), among other ways. We have confined our investigation to raised pitch, because it one of the most easily recognized features of CDS, but most importantly, because pitch—actually variation in pitch—can be extracted from wav files using a pitch extraction algorithm and analyzed computationally. In a word, raised pitch is objective, in a way that, that simplified syntax (what do we mean by “simplified”) can never be. We are aware, of course, that raised pitch occurs in many situations beyond CDS, in anger, for example or when surprised. LENA, however, allows us to extract just those speech segments where a parent is speaking directly to a child. Some of these may reflect anger, of course, given that the recordings were done in the home. Nevertheless, one of our operating assumptions in that in over 7,000 of in-home speech, most will not be spoken in anger (or with other non-typical mental states associated with raised pitch. We can now state the null hypotheses with precision: *Mothers and fathers will produce higher mean f0 during CDS than during non-CDS.*

To investigate this hypothesis, over 7,000 hours of inter-family speech were recorded using the LENA recording, labeled with associated LENA software [12], [13], and stored using a conventional Linux file system. Specially constructed software written in Python 2.7 traversed the file system, constructing nearly four million of 1 – 2 second instances of CDS as wav files. Adult speech was distinguished from child speech by context. For example, a speech segment which LENA determined to be that of adult male was considered adult speech if it was found adjacent to another adult segment and CDS if it was found adjacent to a segment LENA determined to be child speech. The f0 of each CDS segment was extracting using RAPT ([32], [33]) and analyzed with the specially constructed software mentioned above. Table 1 shows the study details.

The recorded data and the software used to traverse it and extract and analyze *f0* is freely available to researchers through HomeBank. HomeBank is an online database of daylong audio recordings of child speech in a naturalistic environment. HomeBank was designed to give researchers’ access to “large-scale data and tools, linking the acoustic, auditory, and linguistic characteristics of children’s environments with a variety of variables including socioeconomic status, family characteristics, language trajectories, and disorders” [14, p. 128]. The software developed to study CDS is modular in design. It can be used with the data repository to study many other aspects of child speech.

**Table 1. Participants & Materials**

|  |  |
| --- | --- |
| Participants | 62 Families  20 Families with TD Children (12 boys, 8 girls)  42 Families with HH Children (18 boys, 24 girls) |
| Sex | 52% female (57% in TD sample, 40% in HH sample)  48% male (43% in TD sample, 60% in HH sample) |
| Child Age | Mall = 2.53 yrs (SD = .69 yrs)  MTD = 2.39 yrs (SD= .79 yrs)  MHH = 2.60 yrs (SD= .64 yrs) |
| Secondary Disabilities | None (by parent report) |
| Data | Unprocessed whole-day recordings  (single channel, 16KHz, 16 bit, PCM) |
| Total Recording Time | 7,541.23 hours in 641sessions  (Available from HomeBank [14])  10.34 mean sessions/family  117.83 mean hours per family |
| Child Directed Speech | Total: 1,414.51 hours  Total CDS Instances: 3,829,565  Mean CDS Instances per family: 61,767.2 |
| LENA coding used to  determine adjacency | CHN: child near MAN: male adult near  FNN: female adult near |
| Software | 1. LENA software for coding 2. Software for f0 extraction   (coded in C) [32], [33]   1. Custom-built Python2.7   Software to find and extract CDS  (Available from HomeBank [14])   1. Custom-built Python2.7 and   MatLab software to compute statistics |

**V. Results**

The results of the study are shown in Tables II, III and IV, and graphically in Figure 1. In accord with our expectation, mean *f*0 values for mothers and fathers were consistent with well- known ranges for adult women and men (*M*mothers=227.5 Hz, *SD*mothers=54.2 Hz; *M*fathers=148.5 Hz, *SD*fathers=40.4 Hz).   During periods of adult-directed-speech (ADS), mothers’ mean *f*0 was 222.1 Hz (*SD*=53.6 Hz) and during CDS it was 233.0 Hz (*SD*=54.7 Hz).  The difference between mothers’ ADS and CDS was significant (*t*(151)=27.89, *p*<10-60).

During periods of ADS, fathers’ mean *f*0 was 146.1 Hz (*SD*=39.4 Hz) and during CDS it was 150.9 Hz (*SD*=41.3 Hz).  The difference between fathers’ ADS and CDS was significant (*t*(151)=8.07, *p*<10-12).  In both figures, the lighter line is a linear bisector of the equally scaled square figure.  The heavier line is the least squares linear regression for the distribution shown in each figure.  For the mothers, the regression was significant (*R*2=.844, *p*<10-61).  For fathers, the regression was significant (*R*2=.373, *p*<10-16).  Both mothers and fathers used higher mean *f*0 values in the CDS condition than in the ADS condition, though the relationship, as shows in Tables V and VI was much stronger for mothers than for fathers.

**Table II. *f0* Mothers and Fathers**

|  |  |  |
| --- | --- | --- |
|  | m(hz) | Sd(HZ) |
| mothers | 227.5 | 57.2 |
| fathers | 148.5 | 40.4 |

**Table III. *f0* Mothers**

|  |  |  |
| --- | --- | --- |
|  | m(hz) | Sd(HZ) |
| mothers ADS | 221.1 | 53.6 |
| MOthers CDS | 233.0 | 54.7 |

**Table IV. *f0* Fathers CDS**

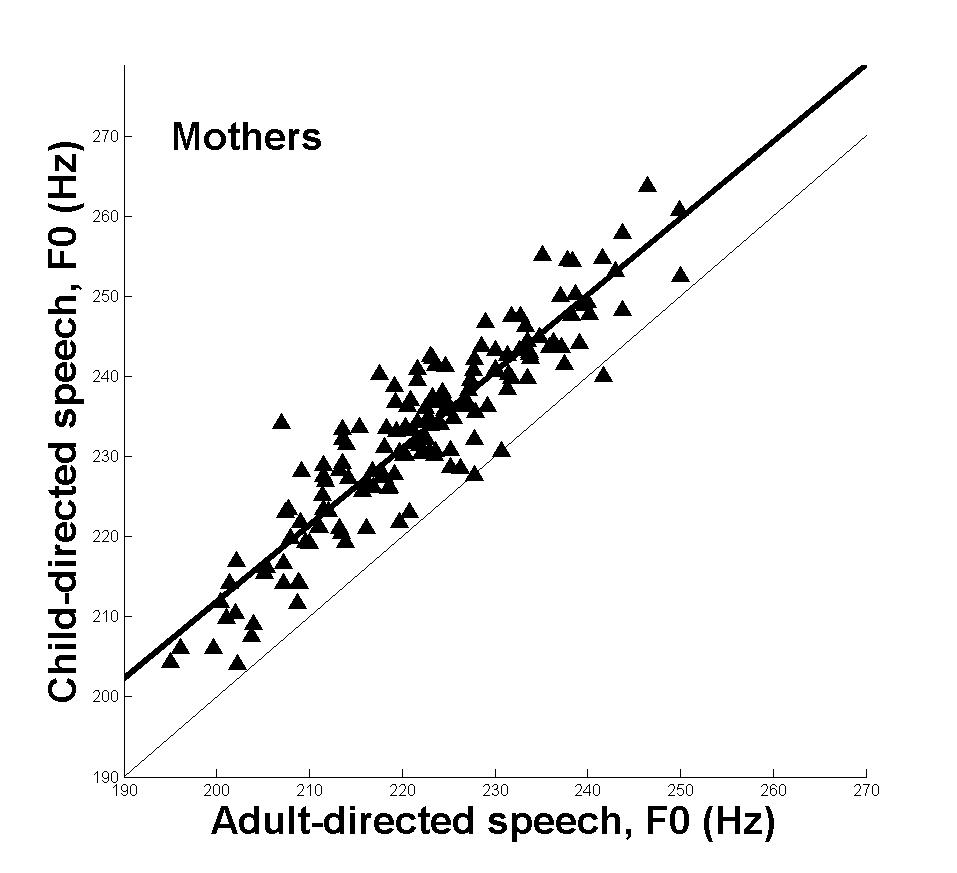
|  |  |  |
| --- | --- | --- |
|  | m(hz) | Sd(HZ) |
| Fathers ADS | 146.1 | 39.4 |
| fathers CDS | 150.0 | 41.3 |

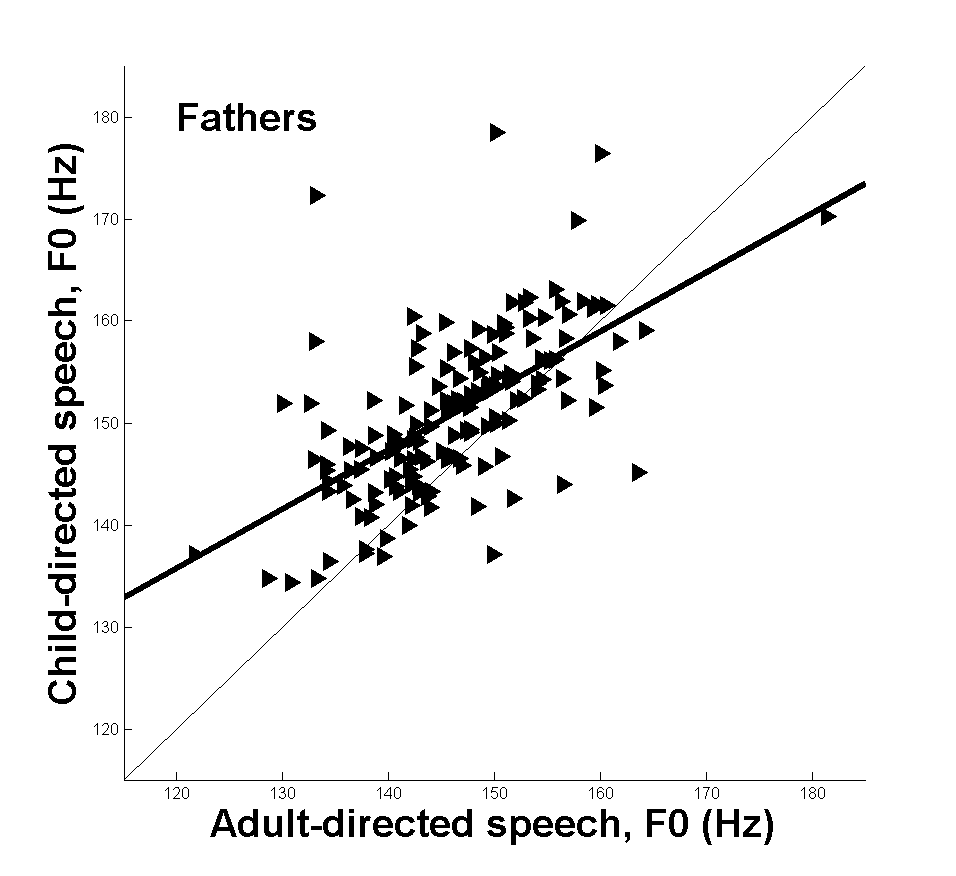
**Table V. ADS CDS Difference**

|  |  |  |
| --- | --- | --- |
|  | t | p |
| Mothers | t(151)=27.89 | <10-60 |
| Fathers | T(151)=8.07 | <10-12 |

**Table VI. Regression**

|  |  |  |
| --- | --- | --- |
|  | R2 | p |
| Mothers  Fathers | .844  .373 | 10-61  10-16 |
|  |  |  |





**V. Conclusions and Current Research**

REFERENCES

[1] VanDam, M., De Palma, P. 2014. “Fundamental Frequency of Child-Directed Speech Using Automatic Speech Recognition.” *Proceedings of the 7th International Conference on Soft Computing and Intelligent Systems and the 15th International Symposium on Advanced Intelligent Systems*, Kitakyushu, Japan, Dec., 2014.

[2] Piaget, J. 1959 (first published, 1923). The Language and Thought of the Child. London: Routledge.

[3] Skinner, B.F. 1957. Verbal Behavior. New York: Appleton-Centurty-Crofts.

[4] Chomsky, N. 1959. “A Review of B. F. Skinner’s Verbal Behavior.” Language, 35(1), 26-58.

[5] Brown, R. 1973. A First Language: The Early Stages. Cambridge, MA: Harvard University Press.

[6] Lenneberg, E, 1967. Biological Foundations of Language. New York: Wiley.

[7] VanDam, M., Warlaumont, A., MacWhinney, B. 2015. “Collaborative Research: Enabling Access to and Analysis of Shared Daylong Child and Family Audio Data.” The National Science Foundation, Award Numbers, 1539133, XXX, XXX.

[8] Hart, B., Risley, T. 1995. Meaningful Differences in the Everyday Experiences of Young American Children. Baltimore: Paul H. Brookes Publishing Co.

[9] Saon, G., Chien, J. 2012. “Large-vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances.” IEEE Signal Processing Magazine, 29(6):18-33.

[10] Novet, J. 2015. “Apple Claims Siri’s Speech Recognition Tech is More Accurate than Google’s.” VB, 6/8/2015. Retrieved 12/20/2016 from: <http://venturebeat.com/2015/06/08/apple-claims-siris-speech-recognition-tech-is-more-accurate-than-googles/>

[11] Maas, A., Xie, Z., Jurafsky, D., Ng, A. 2015. “Lexicon-Free Conversational Speech Recognition with Neural Networks.” Proceedings Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, Denver, CO, May 31 – June 5, pp. 345-354.

[12] LENA 2014. LENA Research Foundation. Retrieved 12/19/2016 from: [www.lenafoundation.org](http://www.lenafoundation.org)

[13] LENA 2011. “The LENA advanced data extractor (ADEX) user guide version 1.1.2.” Retrieved 12/19/2016 from: <https://cdn.shopify.com/s/files/1/0596/9601/files/The_LENA_ADEX_User_Guide.pdf?416155826784605683>.

[14] VanDam, M., Warlaumont, A., Bergelson, E., Cristia, A., Sodestrom, M, De Palma, P., MacWhinney, B. (2016). “HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. Seminars in Speech and Language.” 37(02):128-142.

[15] Feinberg, R. 1988. “Margaret Mead and Samoa: Coming of Age in Fact and Fiction.” American Anthropologist. 90:656-663.

[16] Jurafsky, D., Martin, J. 2009. Speech and Language Processing. Upper Saddle River, NJ: Pearson/Prentice Hall.

[17] 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.

[18] 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](http://www.google.com.ar/patents/US874484).

[19] 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.

[20] VanDam, M. Silber, N. 2013. “Precision and Error of Automatic Speech Recognition.” *Proceedings of Meetings on Acoustics*, 19: 060006. Soi:10.1121/1.4798466.

[21] 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.

[22] 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/TechReports.aspx/Reliability/LTR-05-2>.

[23] 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 5/14/2014 from: <http://jslhr.pubs.asha.org/article.aspx?articleid=1972911>.

[24] De Palma P. 2014. “Probabilistic Methods in Automatic Speech Recognition.” In M. Khosrow-Pour (ed.), *Encyclopedia of Information Science and Technology*. Heshey, PA: IGI Global.

[25] Fernald, A. 1985. “Four-month-old Infants Prefer to Listen to Motherese.” *Infant Behavior and Development*, vol. 8(2): 181-195.

[26] Kemler Nelson, D., Hirsh-Pasek, K., Jusczyk, P., Wright Cassidy, K. 1989. “How the Prosodic Cues in Motherese Might Assist Language Learning.” *Journal of Child Language*, vol 16(1):55-68.

[27] Masataka, N. 1999. “The Role of Modality and Input in the Earliest State 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.

[28] Falk, D. 2004. “Prelinguistic Evolution in Early Hominins: Whence Motherese?” *Behavioral and Brain Sciences*, vol. 27(4):491-503.

[29] Lombard, E. 1911. “Le signe de l’elevation de la voix.” *Annalses Des Malades de L’Ooreille, Larynx, Nez, Pharynx*, vol. 37(2):101-119.

[30] 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. 935:136:174.

[31] Warren-Leubecker, Bohannon, J. 1984. “Intonation Patterns in Child-Directed Speech: Mother-Father Differences.” *Child Development*, vol. 55(4):1379-1385.

[32] Talkin, D. 1995. “A Robust Algorithm for Pitch Tracking.” In W. Kleijn, K. Paliwal (eds.), *Speech Coding and Synthesis*. Atlanta, GA: Elsevier Science B.V.

[33] Talkin, D. 2017. “David Talkin's pitch tracker, get\_f0, modified to be a stand-alone binary using dpwelib for sound I/O.” Available at: <https://github.com/dpwe/get_f0_snd>.

[34] Hess, W. 1983. *Pitch Determination in of Speech Signals*. Berlin:Springer-Verlag.