Extracting and Analyzing Fundamental Frequencies From a Very Large Corpus of Audio Recordings

Paul De Palma[[1]](#footnote-1) Mark VanDam  
Department of Computer Science Department of Speech and Hearing Sciences  
Gonzaga University College of Medical Sciences

Spokane, WA USA Washington State University

[depalma@gonzaga.edu](mailto:depalma@gonzaga.edu) Spokane, WA, USA

mark.vandam@wsu.edu

**Abstract—Linguists and developmental psychologists have long-recognized distinctive speech patterns in adults speaking to infants and toddlers. The collection of patters is referred to as *child-directed speech* or *motherese.* The 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 that 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) although mothers have a stronger tendency to raise f*0*than fathers, fathers do raise f0. This is an amendment of the findings from [1] and underscores the need for the use of very large corpora in computational linguistic research; and 2) 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, big data**

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). In the process we 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. Further we show the value of large corpora linguistic research, since the results reported here using just over 7,000 hours of speech vary from the same experiment on just under 500 hours of speech drawn from the same corpora and reported in [1].

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, through automatic speech recognition, the human component from the data collection and coding process.

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 complete introductions, see [9], [16], [17], [24]. 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 systems, in essence, produce 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 human language has its own inventory of phones, where a phone is a speech unit. A typical English phone is the 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 hidden Markov models (HMM). The HMM, 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—reduce 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 of 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 motherese. 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 are 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 to, for example, 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 is 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 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.

We have made 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.

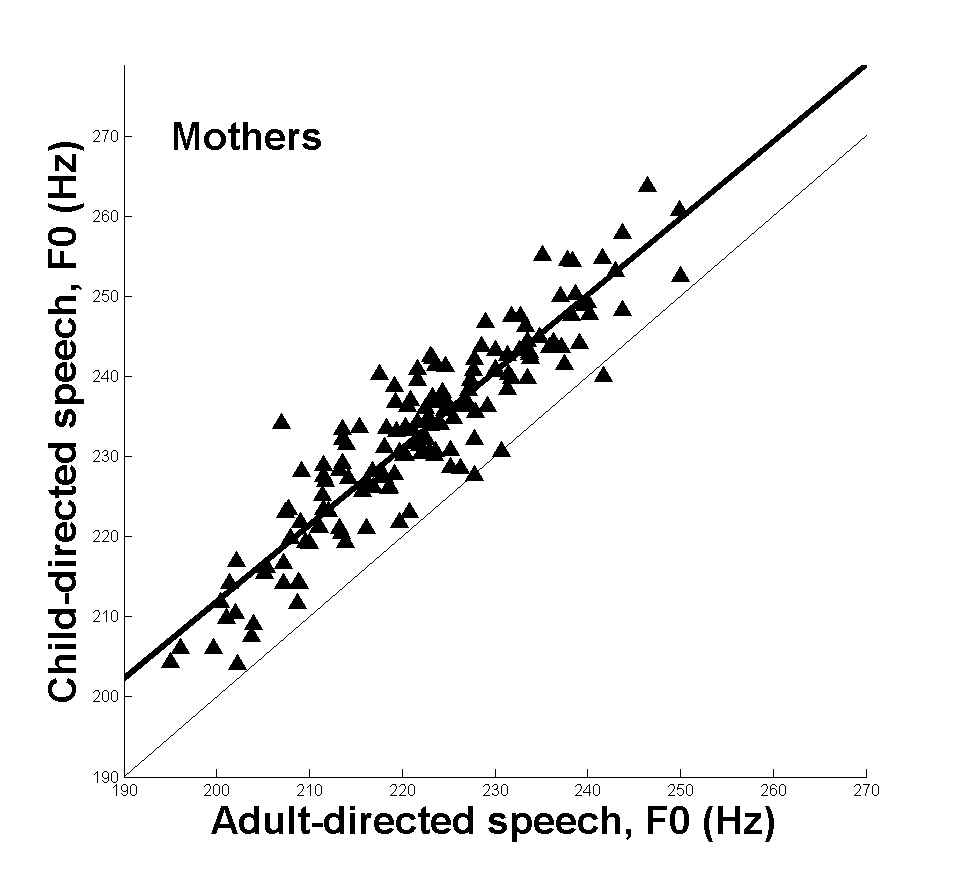
**Table 1. Participants & Materials**

|  |  |
| --- | --- |
| 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 Recordings | 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 Fig 1. In accord with our expectation, mean *f*0 values for mothers and fathers were consistent with known values 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). In Figs 1 and 2, adult-directed speech (ADS) and CDS is on the Y-axis. An observation on the bisector, the lighter line, indicates equal *f0* in the ADS and CDS situations. During periods of 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 Figs. 1 and 2, the heavier line is the least squares fitted linear regression for the distribution shown in each figure. The fit of the line was significant for both mothers (*R*2=.844, *p*<10-61) and fathers (*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, although the relationship was stronger for mothers than for fathers as is shown in Tables V and VI.

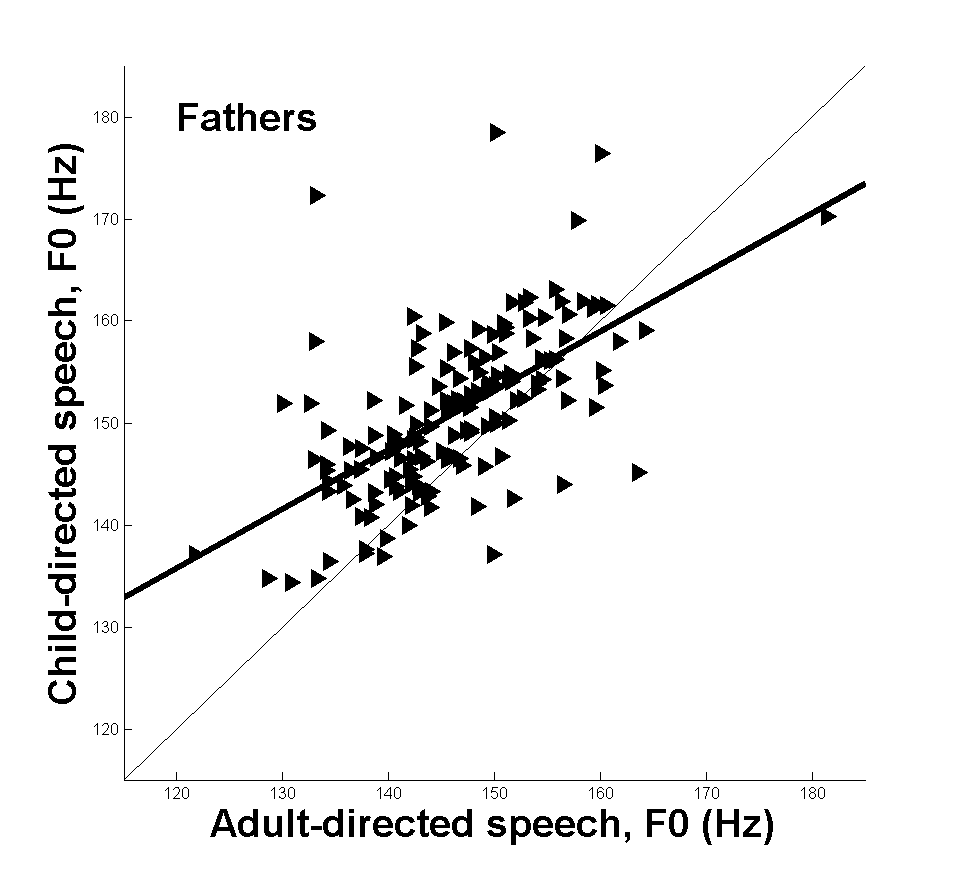


*Fig. 1*. The fundamental frequency of mothers’

adult-directed speech by child-directed speech is shown.

The bisector is shown by the light line, and the least-squares

linear regression is shown by the heavy line.



*Fig. 2*. The fundamental frequency of fathers’

adult-directed speech by child-directed speech

is shown. The bisector is shown by the light line,

and the least-squares linear regression is shown by

the heavy line.

**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 |
|  |  |  |

VI. Conclusion and Future Directions

In this paper, we showed that a very large database of naturally-collected audio can be processed and analyzed for features known to be important for human communication. Here we analyzed hundreds of daylong recordings collected from the auditory perspective of a preschool child in his or her normal family routine. Thousands of hours of audio were collected *in situ*, diarized with automatic speech processing techniques from the LENA Foundation, then further processed by our algorithms to extract a speech feature, fundamental frequency, of specific talkers in the context of the diarization coding.

This work has two main goals. First, this is a proof of concept that a very large database of wild-type auditory data can be successfully captured and processed in a way that is useful to a wide variety of researchers. Researchers in computer science, algorithms, speech science, automatic speech recognition, speech and language disorders, engineering, bioacoustics and other fields may benefit from techniques such as those described here. Theoretical implications include applying this technology and approach to better understand fields from database management to the implementation of language in human communication systems. Practical implications of this work include better understanding of early human communication, improving algorithms and processing techniques for automatic speech processing and automatic speech recognition, and identifying communication characteristics of children who may be at risk of developmental delay or disorder.

Second, this work addresses the question of how fathers and mothers control their speech in different communicative contexts, namely when talking with adults or when talking with children. The fundamental frequency shift described here for mothers has been similarly described by others, but it has not been demonstrated with a very high number of observations or in highly naturalistic environments as shown here. Another question of interest here that has not been examined thoroughly in the literature is the speech behavior that fathers show in the presence of adults and children. Here again we show that fathers’ speech patterns are similar to mothers’ in gross respect—that is, on average fathers use higher *f*0 in CDS as compared to ADS—but the pattern is not identical to mothers. Further, the results presented here differ from results done in an identical experiment on a 491.2 hour subset of the corpora. That experiment indicated that mothers (*t(32)*=18.6, *p*<10-18) but not fathers (t(32)=.55, *p*>.5) raise *f0* during CDS. Not only is it possible to use very large corpora of auditory data in research, their use can correct problems that appear in work with smaller corpora. A detailed description of the difference is beyond the scope of this report but may reveal important differences between mothers and fathers.

Having demonstrated the fundamental utility of a very large database speech corpus through a fully explicated example, we expect this research program to have several fruitful avenues in the future. First, there is a need to refine and extend existing approaches to data collection, analysis, and processing. Researchers have reported LENA system performance, and we expect this work to continue. Nevertheless, there remain certain proprietary aspects of the system that are inaccessible to researchers. Further, the LENA system may not be possible or appropriate to use in some applications, such as for children with sensory or other disorders. To date, the LENA system has no fully functional alternatives. To address this, we are working toward developing an alternative system without proprietary restrictions. This work also includes improving algorithms in the pre- and post-processing stages of raw data analysis. Due to the large volume of data to be processed, improved efficiency and reliability are needed.

Second, this technology and approach have great potential to impact at-risk populations including children with developmental disorders and children and families from low socio-economic or other disadvantaged backgrounds. In one project, we are looking at the effect of mild-to-moderate hearing loss on the speech development of preschoolers. We are using the automatic methods here to assess speech production characteristics and compare between preschoolers with and without hearing loss.

Third, this technology can lead to better understanding of typical development. As wearable biotechnology rapidly grows and changes, researchers have a dramatically different ability to observe and document typical development, not only in the domain of communication and language but also in domains such as motor control or sociobiological characteristics, to name a just two. It is currently not well understood how observable patterns in various domains interact or may be related. For example, the work reported here suggests that fathers may use different speech characteristics than mothers in the speech they engage in with their children. Exploring these differences in a variety of contexts will help researchers better understand the role of fathers.

Fourth, despite the advantages, data collection and analysis remains a challenging task. To reduce the burden

and positively leverage the results of many researchers working in this field, there are efforts to archive and document audio data, associated metadata, and processing tools (such as the pitch determination algorithm extractor used in this work along with the code used to traverse the file system and analyze the extracted data). The accessible online repository HomeBank (<http://homebank.talkbank.org/>) [14] makes a wide variety of data available to researchers to explore new possibly applications, improve the technology, and contribute additional (raw) data.

The work reported here is an early demonstration of new, exciting technology and its application to a practical question of interest to researchers in speech and allied fields. The methods and procedures hold great promise to advance both the theoretical underpinnings and the practical application of this emerging science.

ACKNOWLEDGMENT

P.D. wishes to thank the School of Engineering and Applied Science for sabbatical leave that enabled him to pursue this work. He also thanks is very able research assistant Rianne Lyons who assisted in the preparation of the manuscript. The recordings were supported by National Institutes of Health Grants NIH/NIDCD 5 ROI 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. The work was supported ty the WSU Spokane Seed Grant Program. We thank the families who contributed, and the children and families of the HOPE School of Spokane, WA.

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.

[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.

[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.

[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..

[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).

[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.

[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.

1. This paper represents the equal contribution of both authors. [↑](#footnote-ref-1)