Using Automatic Speech Processing to Analyze Fundamental Frequency of Child-Directed Speech Stored in a Very Large Audio Corpus

Paul De Palma
Department of Computer Science
Gonzaga University
Spokane, WA USA
depalma@gonzaga.edu

Mark VanDam
Department of Speech & Hearing Sciences
Elson S. Floyd College of Medicine
Washington State University
Hearing Oral Program of Excellence (HOPE)
Spokane, WA, USA
mark.vandam@wsu.edu

***Abstract*—Child-Directed-Speech (CDS) is associated with raised fundamental frequency (f_0). In a previous paper we claimed that f_0 could be extracted from 500 hours of audio recordings using soft computing techniques and that mothers, but not fathers, increase f_0 in CDS. Using an audio corpus more than ten times larger, this paper reports that fathers do raise f_0 but not as much as mothers. The principle finding is a proof of concept: 1) very large speech corpora, unavailable until recently, can be processed using soft computing techniques; 2) the use of very large corpora may force revisions of conclusions based on smaller datasets.**

Keywords—child-directed speech; automatic speech processing; LENA; big data; fundamental frequency

I. INTRODUCTION

The 1950s, 1960s, and 1970s, saw the start of serious research into child language development (cf. [1]-[4]). Data collection, whether through field observation or controlled laboratory experiments, was time-consuming and expensive. In both cases, the sample sizes were small and dependent upon trained transcribers, who erred and brought their own biases to the observations. In addition, laboratory investigation raised questions of ecological validity: 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-1990s, for example, Hart and Risley [5] argued that the number of hours of conversation parents have with their children is the strongest predictor of future academic success. The constraints under which Hart and Risley worked 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.

In the mid-1950s, at the same time that the cognitive revolution was encouraging researchers to view language computationally, work began on 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 [6].

The LENA Research Foundation (Boulder, Colorado, USA), by applying modern speech processing 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 [7, 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 have been not just been 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 manner in which mothers (the choice of gender is intentional) speak with their infants and toddlers. Though CDS can be characterized in many ways, we confine ourselves to a single parameter, the parents' vocal fundamental frequency (f_0), a parameter that can be extracted from the speech stream and analyzed by

978-1-5090-4917-2/17/$31.00 ©2017 IEEE

computer [8]. This last is important, since it implies an objective measure rather than fuzzier impressions of, for example, reduced syntactic complexity when speaking with children. We ask a single question: does the CDS of fathers, using the proxy of raised f_0, differ from that of mothers? In answering this question, we show that soft computing techniques can be used to process over 7,000 hours of recorded speech and nearly 4,000,000 individual conversational instances 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 [7].

II. CHILD LANGUAGE RESEARCH, AND LENA

Samples of children's speech are usually collected in the laboratory or during home visits. Researchers are 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 are costly to obtain. LENA was developed to solve the problems of cost, ecological validity, and bias by removing, through automatic speech processing, the human component from the data collection and coding process.

Since all readers 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 introduction, see [9]-[12]. Speech is the perturbation of air by the human vocal apparatus. Modern ASR treats speech as a noisy version of an idealized speech string intended by the speaker. ASR systems produce a probabilistic mapping from the acoustic signal to the speech string. They do this through familiar Bayesian inference techniques. If we let O represent a sequence of acoustic observations, and S a sequence of words from a language L, we can state the speech recognition problem as a conditional probability:

$$G(S) = \max(P(S|O)) \text{ s.t. } S \in L \quad (1)$$

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:

$$G(S) = \frac{\max(P(O|S) \times P(S))}{P(O)} \text{ s.t. } S \in L \quad (2)$$

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

$$G(S) = \max(P(O|S) * P(S) \text{ s.t. } S \in L \quad (3)$$

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, use standard digital signal processing techniques to extract feature vectors from periodic samples of an acoustic waveform. These are 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 distinct sound. A typical English phone is the initial *p* in 'pan,' known as a *bilabial stop* and produced by closing the lips, adding pressure to the closed oral cavity, then releasing the increased air pressure by opening the lips. A triphone is a phone with its left and right subphonetic contexts. Its use is an attempt to model coarticulation, the property exhibited in the English vowel *eh*, for example, which could produce a somewhat different set of acoustic features, depending on whether it appears in *wed*, *yell*, or *Ben* [10]. Taken together, the feature extraction and subsequent statistical mapping, 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 described 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 [10]-[11]. 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.

Using LENA and software we developed, we have collected, labeled, and analyzed over 7,000 hours of speech data from infants, toddlers, and their parents. There are two components to LENA: an acoustic recording device and software that performs digital signal processing and classification (i.e., labeling) 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. LENA uses techniques common to most ASR systems until the very recent introduction of neural networks [6], but

978-1-5090-4917-2/17/$31.00 ©2017 IEEE

with a crucial difference. Audio streams are mapped not to words, as in standard ASR, but to over sixty labels that indicate the source of the sound, labels such as *key child*, and *adult male near* [13]-[14].

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. Several studies show a mean agreement of 76.25% between LENA and human transcribers [15]-[17]. This is consistent with standard ASR systems [9], [12]. More recently, a four-alternative, forced-choice task with 24 judges and 2,340 segments of LENA-labeled speech data found an agreement between judges and LENA of 79%, again, consistent with standard ASR systems [18]-[19].

III. CHILD-DIRECTED SPEECH

Child-Directed Speech (CDS), or *motherese,* can be described syntactically (reduced complexity), phonologically (hyperarticulation), lexically (specialized vocabulary), and acoustically (raised fundamental frequency). 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 over standard speech. 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 [20]-[22]. Indeed, at least one study implicates CDS in the evolution of language itself [23].

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 f_0. Said another way, f_0 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 the two are correlated [8], we use the terms f_0 and pitch interchangeably. Here, we use fundamental frequency as the dependent measure to describe 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, f_2, or any among many acoustic correlates of behaviors. 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 speech scientists for forty years, namely ecological validity and the cost of data collection and coding. Because our speech data were collected using inexpensive digital recording and storage devices and analyzed using automatic speech processing techniques, we have examined 7000+ hours of speech as opposed to, for example, the nine hours reported in [24].

IV. MATERIALS AND METHODS

Though CDS can be described in multiple 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 can be extracted from audio files using a pitch extraction algorithm and analyzed computationally [8], [25]. In a word, the determination of raised pitch is *objective*, in a way that other language features such as syntax can never be. We can now state the null hypotheses with precision: *Mothers and fathers will produce higher mean f_0 during CDS than during non-CDS.*

To investigate this hypothesis, over 7,000 hours of intra-family speech were recorded and labeled using LENA, and stored in a conventional Linux file system. Specially constructed software traversed the file system, building nearly four million 1-2 second instances of conversation as WAV files. Adult speech was distinguished from child speech by context. A speech segment, whose source LENA determined to be an adult, was considered adult speech if it was found adjacent to another adult segment; it was considered to be CDS if it was found adjacent to a segment whose source LENA determined to be a child. The f_0 of each adult participating in a conversation was extracted using RAPT [25]-[26] and analyzed with the specially constructed software mentioned above. Table I shows the study details.

V. RESULTS

The results of the study are shown in Tables II, III and IV, and graphically in Figs. 1 and 2. As expected, 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) is on the abscissa and CDS is on the ordinate. An observation on the bisector, the lighter line, indicates equal f_0 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

978-1-5090-4917-2/17/$31.00 ©2017 IEEE

(*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 ordinary least-squares fitted linear regression for the distribution shown in each figure. The fit of the line was significant for both mothers ($R^2=0.844$, $p<10^{-61}$) and fathers ($R^2=0.373$, $p<10^{-16}$). 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.

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, it shows that a very large database of wild-type auditory data can be successfully captured and processed. Researchers in data base construction, algorithms, speech science, automatic speech recognition, speech and language disorders, digital signal processing, and bioacoustics may benefit from and contribute to the techniques described here. Theoretical implications include applying this technology and approach to better understand fields from data 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. The fundamental frequency shift described here for mothers has been similarly described by other researchers, but with many fewer observations and outside of the naturalistic environment described here. Another question of interest, addressed but not examined thoroughly in the literature, is the differential speech behavior that fathers show in the presence of adults and children. 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 patterns are not identical. We point out here that the sample included both typically developing children and those with some hearing loss.

Further, the results presented here differ from results the authors reported 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 f_0 during CDS [7]. Not only is it possible to use very large corpora of auditory data in research, their use can correct problems that appear in work based on much smaller corpora. We take this to be a significant finding, since 491.2 hours of recorded data is itself a large corpus. 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 collection of audio recordings through a fully explicated example, we expect this research program to have several fruitful avenues in the future.

TABLE I. Participants & Materials

Participants	62 Families
Child Sex	52% female 48% male
Child Age	M = 2.53 yrs (SD = .69 yrs)
Data	Unprocessed whole-day recordings (single channel, 16KHz, 16 bit, PCM)
Total Recordings	7,541.23 hours in 641 sessions 10.34 mean sessions/family 117.83 mean hours per family
Conversations	Total: 1,414.51 hours Total Conversations: 3,829,565 Mean Conversations 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 [13] 2) Software for f_0 extraction [24]-[25] 3) Custom-built software to find and extract CDS. Available from HomeBank [26] 4) Custom-built data analytic software

978-1-5090-4917-2/17/$31.00 ©2017 IEEE

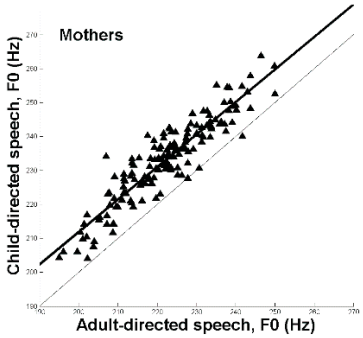

Fig. 1. Fundamental frequency of mothers' speech.

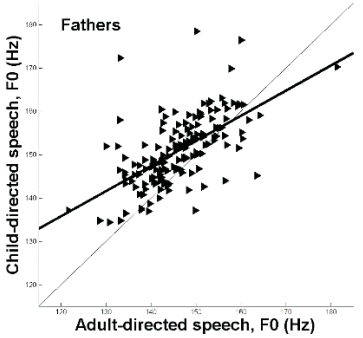

Fig. 2. The fundamental frequency of fathers'

TABLE II. f_0 Mothers and Fathers

	M(HZ)	SD(HZ)
MOTHERS	227.5	57.2
FATHERS	148.5	40.4

TABLE III. f_0 Mothers' ADS & CDS

	M(HZ)	SD(HZ)
MOTHERS ADS	221.1	53.6
MOTHERS CDS	233.0	54.7

TABLE IV. f_0 Fathers CDS

	M(HZ)	SD(HZ)
FATHERS ADS	146.1	39.4
FATHERS CDS	150.0	41.3

TABLE V. ADS CDS Compared

	t	p
MOTHERS	t(151)=27.89	$<10^{-60}$
FATHERS	T(151)=8.07	$<10^{-12}$

TABLE VI. Regression

	R^2	p
MOTHERS	0.844	10^{-61}
FATHERS	0.373	10^{-16}

First, there is a need to refine and extend existing approaches to data collection, analysis, and processing. Researchers have reported on LENA system per-formance, and we expect this work to continue. Nevertheless, proprietary aspects of the system are inaccessible to researchers. Further, the LENA system may not be appropriate for 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 has 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 another 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 to assess speech production characteristics and compare them 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 just two. It is currently not well-understood how observable patterns in various domains interact. 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 demonstrated here, 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. The accessible online repository HomeBank makes a wide variety of data available to researchers to explore new possible applications, improve the technology, and contribute additional data. The data used in this experiment along with the custom-built software are available through HomeBank [27].

978-1-5090-4917-2/17/$31.00 ©2017 IEEE

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

This paper represents equal contributions of the authors. P.D. thanks Gonzaga University for sabbatical leave and thanks his research assistant, Rianne Lyons, for her help in preparing this manuscript. This research was supported by NIH/NIDCD 5R01-DC009560, NIH/NIDCD 5R01-DC009560-01S1, NSF-SBE RIDIR-1539133, and the Washington Research Foundation. The content of this project is solely the responsibility of the authors and does not necessarily represent the official views of the supporting entities. We thank the families who contributed and the children and families of the Hearing Oral Program of Excellence (HOPE) School of Spokane, WA.

REFERENCES

[1] B.F. Skinner, *Verbal Behavior*. New York: Appleton-Century-Crofts, 1957.

[2] N. Chomsky, "A review of B. F. Skinner's *Verbal Behavior*." Language, 35(1), 26-58, 1959.

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

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

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

[6] A. Maas, Z. Xie, D. Jurafsky, and A. Ng., "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, 2015, pp. 345-354.

[7] M. VanDam and P. De Palma, "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.

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

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

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

[11] X. Huang, A. Acero, and H. Hsiao-Wuen. 2, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper Saddle River, NJ: Prentice Hall, 2001.

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

[13] LENA, "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, 2011.

[14] T. Paul, D. Xu, and A. Richards, "System and Method for Expressive Language Assessment." Patent Number US 8844847 B2, retrieved 7/21/2014 from: http: www.google.com.ar/patents/US874484, 2014.

[15] S. F. Warren, J. Gilkerson, J.A. Richards, D.K. Oller, D. Xu, U. Yapanel, and S. Gray, "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, 2010.

[16] D. Xu, U. Yapanel, S. Gray, and C. Baer, "Reliability of the LENA Language Environment Analysis System in Young Children's Natural Home Environment." Retrieved 7/21/2014 from:http://www.lenafoundation.org/TechReports.aspx/Reliability/LTR-05-2, 2009.

[17] D. Xu, J.A. Richards, and J. Gilkerson, "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, 2014.

[18] M. VanDam and N. Silbert, "Precision and Error of Automatic Speech Recognition," *Proceedings of Meetings on Acoustics*, 19: 060006, 2013, doi:10.1121/1.4798466

[19] M. VanDam and N. Silbert, "Fidelity of automatic speech processing for adult and child talker classifications." *PLOS ONE*, 11(8): e0160588, 2016 doi:10.1371/journal.pone.0160588

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

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

[22] N. Masataka, "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, 1999.

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

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

[25] D. Talkin, "A robust algorithm for pitch tracking." In W. Kleijn, K. Paliwal (eds.), *Speech Coding and Synthesis*, Atlanta, GA: Elsevier Science B.V., 1995.

[26] D. Talkin. "David Talkin's pitch tracker, get_f0, modified to be a stand-alone binary using dpwelib for sound I/O," https://github.com/dpwe/get_f0_snd, 2017.

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