A longitudinal study of prosodic exaggeration in child-directed speech

Soroush Vosoughi, Deb Roy

The Media Laboratory
Massachusetts Institute of Technology, Cambridge, MA USA
soroush@mit.edu, dkroy@media.mit.edu

Abstract

We investigate the role of prosody in child-directed speech of three English speaking adults using data collected for the Human Speechome Project, an ecologically valid, longitudinal corpus collected from the home of a family with a young child. We looked at differences in prosody between child-directed and adult-directed speech. We also looked at the change in prosody of child-directed speech as the child gets older. Results showed significant interactions between speech type and vowel duration, mean F0 and F0 range. We also found significant changes in prosody in child-directed speech as the child gets older.

Index Terms: prosody,prosodic exaggeration, duration, fundamental frequency, child-directed speech, adult-direct speech, longitudinal, naturalistic, English

1. Introduction

Previous studies have shown how the prosodic aspects of child-directed speech (CDS) differs significantly from adult-directed speech (ADS) (e.g. [1, 2, 3, 4]). One theory is that prosodic exaggeration in CDS helps attract the attention of the child [2]. Other studies have shown that infants are sensitive to the prosodic aspects of speech [5, 6, 7] and that prosody plays an important role in child language acquisition [8, 9, 10, 11, 12, 13]. However, there have not been many studies that look at the prosody in CDS of caregivers over a period of several months, as the child ages.

In this study we examine prosody in hundreds of child-directed and adult-directed utterances from the Human Speechome Project's corpus [14] to first reinforce the picture of prosodic exaggeration in CDS (compared to ADS) and to more importantly use our unique, longitudinal corpus to study how the prosodic aspects of child-directed speech change as the child ages from 9 to 24 months, which includes the child's first productive use of language at about 10 months, all the way to 24 months when the child has learned more than 517 words and is starting to use multi-word sentences.

2. Method

2.1. Corpus

This study uses the corpus collected for the Human Speechome Project (HSP). The HSP corpus is high-density, longitudinal and naturalistic. The corpus consists of high fidelity recordings collected from boundary-layer microphones embedded throughout the home of a family with a young child [14]. For this study we look at a subset of the data collected continuously from ages 9 to 24 months which contains about 4260 hours of 14-track audio of which about 1150 hours contain speech. The data consists of adult-directed and child-directed speech from the three primary caregivers. All child-directed speech is directed at the same child. Two of the caregivers, the mother and the father, are native English speakers while the third caregiver, the female nanny, speaks English as a second language (however all the utterances used in this study are in English).

Of the more than 2.3 million utterances in the corpus we analyze an evenly-sampled 2500 utterances that have been hand-transcribed using new, semi-automatic methods and for which the speaker has been automatically identified with high confidence using our automatic speaker-identification system [15]. The 2500 utterances were distributed between 4 annotators who then used an annotation tool to identify the utterances as child-directed or adult-directed. This annotation tool allowed the annotators to listen to an utterance while reading the corresponding transcription and then making a decision on whether the speech was directed at the child or at an adult. In order to measure the accuracy of the human annotations, a total of 150 utterances were randomly chosen from the 2500 utterances and were given to all the 4 annotators. Table 1 shows the inter-annotator agreement between pairs of annotators. The average pairwise inter-annotator agreement was 0.97 which shows a high level of consistency and accuracy in the human annotated data. Of the 2500 utterances, the annotators identified 1925 as child-directed and 575 as adult-directed.

Table 1: Pairwise inter-annotator agreement for all 4 annotators.

	A2	A3	A4
A1	.97	.96	.96
A2		.98	.99
A3			.97

2.2. Measuring prosody

In our analysis we looked at three prosodic variables, a standardized measure of mean word duration, fundamental frequency (F0) and intensity. Below we give our operational definition for each of these variables and explain how they were extracted.

In order to calculate mean word duration, we first extracted duration for all vowel tokens in our corpus using a forced-aligner. We next converted these to normalized units for each vowel separately (via z-score), and then measured the mean standardized vowel duration for all the vowels in our child-directed and adult-directed utterances (this was done separately for each of the three speakers). For example, a high score on

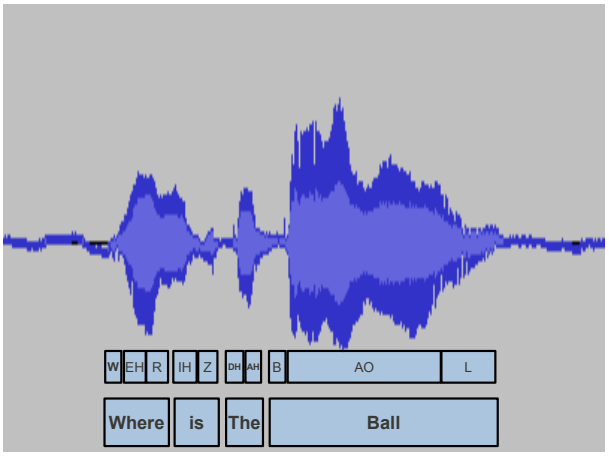

Figure 1: *A sample phoneme level alignment of an utterance generated by the HTK forced-aligner.*

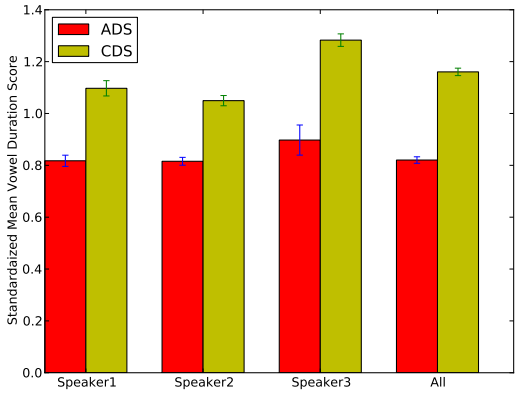

Figure 2: *Standardized measure of duration of vowels in CDS and ADS for all speakers.*

this measure for an utterance would reflect that the vowels that occurred in that utterance were on average often long relative to comparable vowel sounds that appeared in other utterances spoken by the same speaker. We grouped similar vowels by converting transcripts to phonemes via the CMU pronunciation dictionary[16].

The forced-aligner that was used for this task uses the Hidden Markov Model Toolkit (HTK) [17] and works as follows. Given a speech segment and its transcript, the Viterbi algorithm in HTK is applied to align the transcript to the speech segment on the phoneme level, using the CMU pronunciation dictionary[16] and an acoustic model. Since each transcript is associated with a speaker ID label that is generated automatically using our speaker-identification program[15], we can use speaker-specific acoustic models (which we have trained using thousands of samples from our corpus) to get more accurate alignments [18]. Figure 1 shows a sample phone alignment of an utterance done by our forced-aligner.

F0 and intensity contours for each utterance were extracted using the PRAAT system [19]. Using these contours we measured the range, variance and mean F0 and intensity for each of 2500 utterances in our data-set.

3. Results

3.1. Comparison of prosody in CDS and ADS

Differences in prosody between CDS and ADS were assessed using an ANOVA. We found statistically significant interactions between speech type (CDS /ADS) and mean duration ($F(1, 2498) = 144.45, p < 0.001$), F0 mean ($F(1, 2498) = 81.63, p < 0.001$), F0 range ($F(1, 2498) = 35.97, p < 0.001$). We did not find statistically significant interactions between speech type and any of the intensity measurements.

Overall the duration of vowels in CDS was shown to be longer than in ADS across all speakers, as shown in Figure 2. On average, the duration of vowels were about 40 percent longer in CDS than in ADS.

The difference in F0 mean in CDS and ADS was smaller than that of duration but still statistically significant for two of the speakers (Figure 3). On average, mean F0 in CDS was about 20 percent higher than in ADS. F0 range was also significantly higher in CDS compared to ADS in two of the speaker (Figure

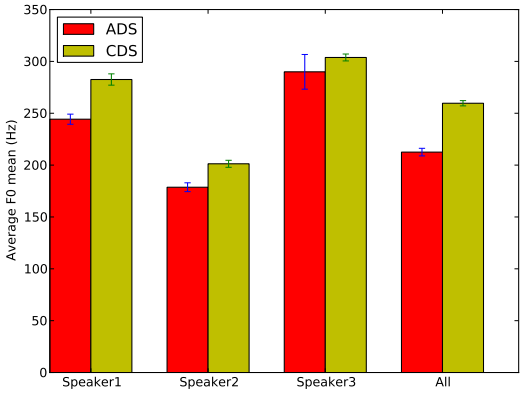

Figure 3: *Average F0 mean in CDS and ADS for all speakers.*

4). Interestingly, speaker 3 who did not show any significant difference in F0 mean between CDS and ADS had the greatest difference in F0 range between CDS and ADS. On the other hand, speaker 1 who showed the most difference in F0 mean did not show any significant change in F0 range. On average, F0 range was about 35 percent greater in CDS compared to ADS.

3.2. Longitudinal study of prosody in CDS

The longitudinal nature of our corpus allowed for the study of how prosodic aspects of CDS changed as the child grew. As mentioned previously, the data-set used for this paper contains 2500 utterances evenly-sampled between 9 to 24 months from the Speechome corpus. We divided our 2500 utterances into 5 age groups, each covering 95 days of the child's life from 9 to 24 months. Since the utterances were evenly sampled, each group had roughly 500 utterances of which around 77% was CDS and 23% ADS.

For each of the three prosodic variables discussed in the last section (Duration, F0 mean and F0 range) we ran ANOVA on the 5 age-groups to see if there are any statistically significant interactions between the prosodic variables and the age of the child. As one would expect, when looking at

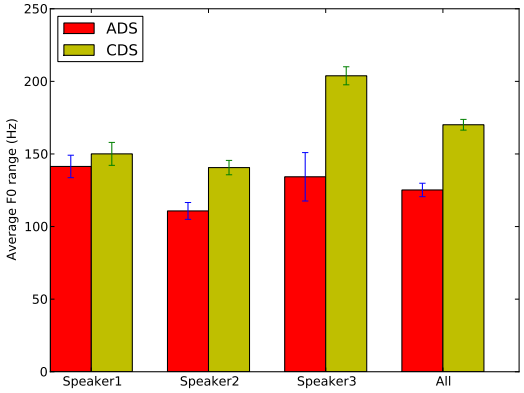

Figure 4: *Average F0 range in CDS and ADS for all speakers.*

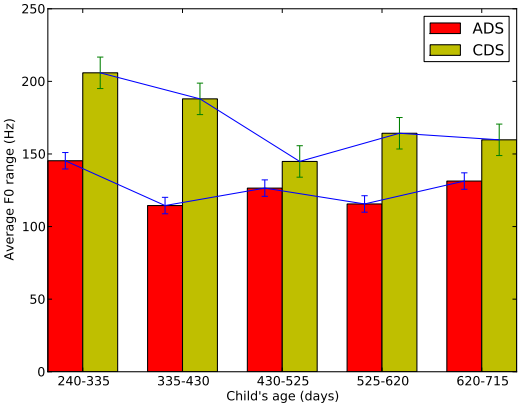

Figure 6: *Average F0 range of all speakers from 9-24 months in CDS and ADS*

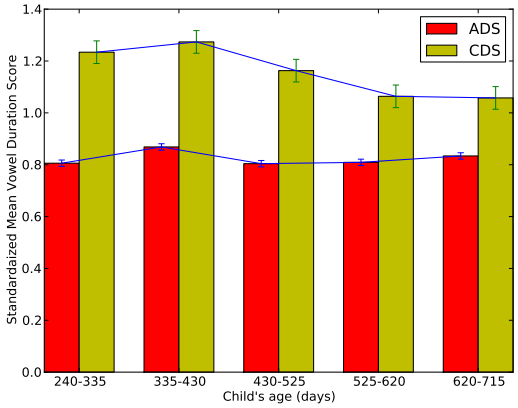

Figure 5: *Standardized measure of duration of vowels from 9-24 months in CDS and ADS*

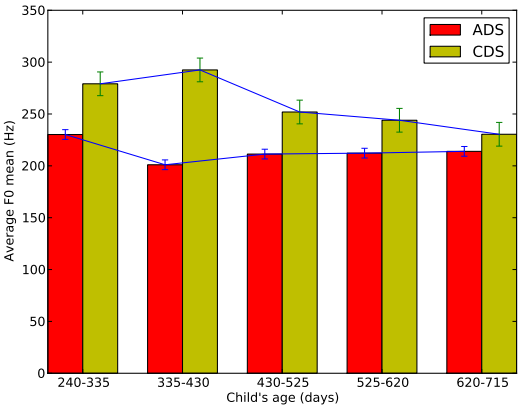

Figure 7: *Average F0 mean of all speakers from 9-24 months in CDS and ADS*

ADS there were no significant interactions between age of the child and duration ($F(4, 570) = 0.75, p = 0.56$), F0 mean ($F(4, 570) = 0.86, p = 0.49$) or F0 range ($F(4, 570) = 0.96, p = 0.43$). For CDS however, we found weak but statistically significant interactions between age of the child and duration ($F(4, 1920) = 9.38, p < 0.001$), F0 mean ($F(4, 1920) = 22.18, p < 0.001$) and F0 range ($F(4, 1920) = 8.22, p < 0.001$).

As shown in Figure 5, overall the duration of vowels in CDS became shorter and moved closer to the duration of vowels in ADS as the child matured while the duration of vowels in ADS did not change significantly.

Though there is a lot more variability in F0 range over time compared to duration, F0 range in CDS got smaller and closer to F0 range in ADS as the child got older (Figure 6), while the F0 range in ADS did not change significantly. We see the strongest change in F0 mean. As shown in Figure 7, there is significant reduction in F0 mean in CDS as the child gets older, while F0 mean in ADS does not show any significant change.

It should be noted that even though all three prosodic variables in CDS changed as the child grew, the are still significantly different (more exaggerated) than ADS at all ages.

The figures discussed above show how each of the three prosodic variables in CDS changed with age. It would also be useful to see the combined change in these variables. In order to visualize that, we created Figure 8. The x-axis of the graph is the age of the child, the y-axis is the average F0 mean, the width of the rectangles represent the average vowel duration score and the height of rectangles represent average F0 range (all normalized for this graph). In other words, the area and the vertical position of the rectangles represent the average prosodic emphasis for a given age. The more exaggerated and salient the caregivers' prosody, the bigger and the higher the rectangles are. As you can see, while the rectangles representing CDS are higher and bigger than the ones representing ADS at all ages, they got lower and smaller in area as the child got older while the rectangles representing ADS barely moved or changed size.

4. Discussion and Conclusions

Our study quantified several prosodic variables in child-directed and adult directed caregiver speech: duration, F0 (mean, range and variance) and intensity (mean, range and variance). We found that there are significant differences in duration, F0 mean and F0 range between child-directed and adult-directed speech.

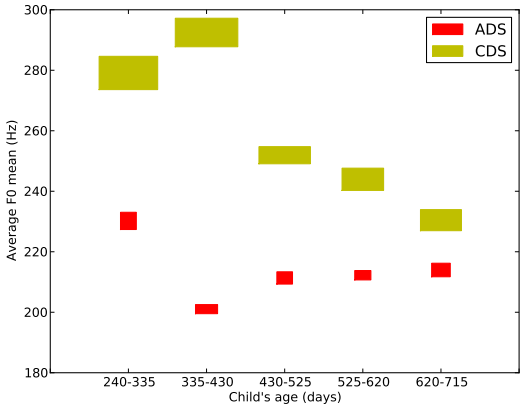

Figure 8: *Change in vowel duration, F0 mean and F0 range of all speakers from 9-24 months in CDS and ADS. The x-axis of the graph is the age of the child, the y-axis is the average F0 mean, the width of the rectangles represent vowel duration and the height of rectangles represent F0 range.*

We showed CDS to be characterized by elongated vowels, raised pitch and a wider pitch range. These results reinforce previous findings on prosodic exaggeration in child-directed speech and the differences between prosody in child-directed speech and adult-directed speech([20] [1], [2], [3], [4], [21]).

The longitudinal nature of our data-set allowed us to examine how prosodic aspects of child-directed speech change as the child in our study gets older. We found that as the child gets older there is less prosodic emphasis in child-directed speech (though there is significant difference between child-directed and adult-directed speech at all times from 9-24 months).

It has been suggested that the prosodic aspects of CDS allow caregivers to elicit the child's attention in order to better facilitate language acquisition when the child is of language acquisition age [22]. Our results validate these claims in that the prosodic features of CDS became less exaggerated as the child matured. Given the short time frame studied here, the changes observed are minimal yet notable.

This paper only examined prosody in caregiver speech, in the future we plan to examine the child's own prosodic development. Interactions between CDS and prosodic development warrant further inquiry.

5. Acknowledgments

We would like to thank Brandon Roy and Matt Miller for their work on the speaker identification system. Thanks to Rupal Patel for her helpful comments about the paper. Special thanks to Jennifer L Bustamante, Carolyn Hsu, Halle Ritter and Katie Sheridan for their help with annotating the data. Finally, thanks to Karina Lundahl for her administrative support.

6. References

[1] D. L. Grieser and P. K. Kuhl, "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese." *Developmental Psychology*, vol. 24, no. 1, pp. 14–20, 1984.

[2] A. Fernald and T. Simon, "Expanded intonation contours in mothers' speech to newborns." *Developmental Psychology*, vol. 20, no. 1, pp. 104–113, 1984.

[3] D. D. Albin and C. H. Echols, "Characteristics of stressed and word-final syllables in infant-directed speech: Implications for word-level segmentation." *Infant Behavior and Development*, vol. 19, pp. 401–418, 1996.

[4] O. Garnica, "Some prosodic and paralinguistic features of speech to young children," *Talking to children: Language input and acquisition*, pp. 63–88, 1977.

[5] R. Cooper and R. Aslin, "Developmental differences in infant attention to the spectral properties of infant-directed speech," *Child Development*, vol. 65, no. 6, pp. 1663–1677, 1994.

[6] A. DeCasper and W. Fifer, "Of Human Bonding: Newborns Prefer Their Mothers' Voices," in *Readings on the Development of Children*, M. Gauvain and M. Cole, Eds. Worth Publishers, 2004, ch. 8, p. 56.

[7] J. Mehler, P. Jusczyk, G. Lambertz, H. Nilofar, J. Bertoncini, and C. Amiel-Tison, "A precursor of language acquisition in young infants," *Cognition*, vol. 29, pp. 143–178, 1988.

[8] S. Vosoughi, B. C. Roy, M. C. Frank, and D. Roy, "Contributions of prosodic and distributional features of caregivers' speech in early word learning," in *Proceedings of the 32nd Annual Cognitive Science Conference*, 2010.

[9] L. Gleitman and E. Wanner, "The state of the state of the art," in *Language acquisition: The state of the art*. Cambridge University Press, 1982, pp. 3–48.

[10] K. Hirsh-Pasek, K. Nelson, G. Deborah, P. Jusczyk, K. Cassidy *et al.*, "Clauses are perceptual units for young infants." *Cognition*, vol. 26, no. 3, pp. 269–286, 1987.

[11] P. Jusczyk, K. Hirsch-Pasek, D. Kemler Nelson, L. Kennedy *et al.*, "Perception of acoustic correlates of major phrasal units by young infants." *Cognitive Psychology*, vol. 24, no. 2, pp. 252–293, 1992.

[12] N. Kemler, K. Hirsh-Pasek, P. Jusczyk, and K. Cassidy, "How the prosodic cues in motherese might assist language learning." *Journal of Child Language*, vol. 16, no. 1, pp. 55–68, 1989.

[13] S. Vosoughi, B. C. Roy, M. C. Frank, and D. Roy, "Effects of caregiver prosody on child language acquisition," in *Proceedings of the 5th International Conference on Speech Prosody*, 2010.

[14] D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, M. Levit, and P. Gorniak, "The Human Speechome Project," in *Proceedings of the 28th Annual Cognitive Science Conference*. Mahwah, NJ: Lawrence Earlbaum, 2006, pp. 2059–2064.

[15] B. C. Roy and D. Roy, "Fast transcription of unstructured audio recordings," in *Proceedings of Interspeech*, Brighton, England, 2009.

[16] H. Weide., "The CMU Pronunciation Dictionary, release 0.6," Carnegie Mellon University, 1998.

[17] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Cambridge University Engineering Dept, 2001.

[18] S. Vosoughi, "Interactions of caregiver speech and early word learning in the speechome corpus: Computational explorations," MIT M.Sc. thesis, 2010.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.01)," http://www.praat.org/, 2009.

[20] A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl, "Mothers, adults, children, pets - towards the acoustics of intimacy," in *Proceedings of the 33rd International Conference on Acoustics, Speech and Signal Processing* , Las Vegas, USA, 2008.

[21] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. de Boysson-Bardies, and I. Fukui, "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," *Journal of Child Language*, vol. 16, no. 3, pp. 477–501, 1989.

[22] A. Fernald, "human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective," in *Language Acquisition: Core Readings*, P. Bloom, Ed. Cambridge, MA: MIT Press, 1994, pp. 51–94.