Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Age-dependent intonational changes in child-directed speech

Daniil Kocharov and Okko Räsänen

Speech and Cognition Research Group, Signal Processing Research Centre,
Tampere University, Finland

daniil.kocharov@tuni.fi, okko.rasanen@tuni.fi

Abstract

The linguistic properties of child-directed speech (CDS) change over time as children get older and their language skills develop. The focus of this research is on prosodic changes of CDS within the earliest years of children's life, especially on the changes in melody. We analyzed mothers' speech from Providence corpus, a collection of longitudinal (bi-monthly) recordings of mother-child spontaneous speech interactions from six English-speaking children between 1.0–3.5 years of age (363 h of audio). Raw prosodic features were extracted from speech using OpenSMILE toolkit. Timing of prosodic events with respect to segmental content was estimated with automatic alignment of orthographic transcripts and the speech signals. Analyses of prosodic features in the data show that mothers' voice in CDS changes during the second and the third years of their children life, as the mean fundamental frequency lowers significantly, while the within-utterance fundamental frequency variability doesn't change.

Index Terms: child-directed speech, utterance length, intonation, pitch, age-dependency

1. Introduction

Child-directed speech (CDS) is a particular speaking style that adults use when addressing infants or young children. CDS differs from adult-directed speech (ADS) in several ways. Syntactically and lexically CDS tends to be much simpler than ADS, resulting in shorter utterances and a smaller vocabulary [1]. From prosodic point of view, CDS is characterized by higher volume [2], higher pitch and its greater variability [3], [4], longer pauses [5], lengthening of vowels [2], and a generally slower speech rate. CDS of mothers also tends to be stronger in emotional expression than when they address adult listeners [6].

Earlier research on CDS shows that all its linguistic properties, including syntactic, lexical, and prosodic characteristics, change as children become older. Soderstrom and colleagues published a case study for two American-English mothers, showing a monotonic increase of the syntactic complexity of the speech addressed to their children, as expressed in terms of number of clauses and full noun phrases per utterance [7]. As a consequence, the mean length of utterances also increases [7, 8]. Simultaneously there is an increase of speech rate in CDS. Raneri and colleagues reported a monotonic increase of speech rate from child age of 0;7 until 2;0 [9]. Ko studied speech rate of six American-English-speaking mothers in Providence Corpus, finding that the speech rate increase is not linear, but that there is a rapid increase of the rate until 24 months of child age, followed by a much slower increase or even slight decrease after that [10].

While there are no contradictory results on syntactic, lexical, and duration properties of CDS obtained in longitudinal research, the changes in the melody within the first years of children's life are not so obvious. McRoberts and Best published a case study of one child, showing that there was a significant monotonic decrease of American-English-speaking mother's F0 from 347 Hz to 228 Hz between ages of 0;2 and 1;5 [4]. Stern with colleagues published similar results, showing that there is a decrease of F0 from 0;4 to 2;0 [8]. There are some published results that show an increase of mother's F0 during first years after child birth. Kitamura and Burnham published results on twelve Australian-English mothers, showing an increase of both mean and range of F0 between child ages of 0 to 12 months [11]. A longitudinal study of eighteen Dutch-speaking mothers showed that their F0 mean and range are higher when their children are 15 months compared to 11 months [12].

To complement the existing literature, the current study investigates age-dependent changes in both general F0 and local within-utterance F0 properties of mothers' when they are speaking to their children during the first years of life. For this purpose, we use the longitudinal Providence Corpus of American-English CDS [13]. Providence Corpus is a valuable speech dataset of longitudinal recordings. As opposed to many other longitudinal CDS datasets it contains tens of recordings over several years for each speaker, more over the recordings were done at home, making them as natural as possible.

2. Method

2.1. Material

The research is based on publicly available Providence Corpus [13]. The corpus consists of twice-monthly recordings of hour-long mother-child spontaneous speech interactions from six American-English-speaking children between approximately 1–4 years. There are from 44 to 88 recording sessions per speaker. The corpus gives possibility to track personal continuous changes in mother's speech while her child gets older. The whole corpus is manually transcribed.

A stationary video-camera with a microphone was used to record mother-child free interactions from a distance of several meters. The recordings were made both inside and outside the house. Participants were not restricted in their movements, thus they were moving around recording area. The conversations were natural and vivid. The background noise was not high enough to disturb the quality of pitch detection.

For the current research, we used a dataset of all the utterances from the Providence corpus, provided by the authors of BabySLM benchmark (ca. 187 000 utterances) (see [14] for details of the utterance segmentation procedure). Utterances containing phonological fragments (e.g. &, #) or unintelligible speech (e.g., annotated as &, xx, or yy) were excluded from further analysis (see CHAT transcription convention in [15]). We used only fully transcribed utterances, ca. 123 000 utterances. The analyses were limited to age range of 1;0 to 3;6, as there were data for only three children outside this age range.

10.21437/SpeechProsody.2024-46

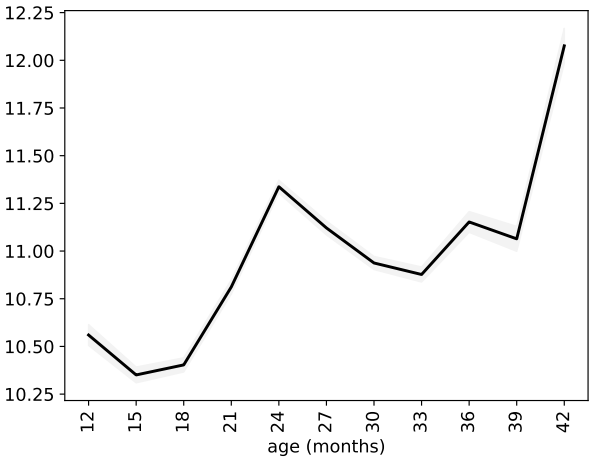

Figure 1: *The mean changes in speech rate (in sounds per second) depending on the age of a child listener in the Providence Corpus.*

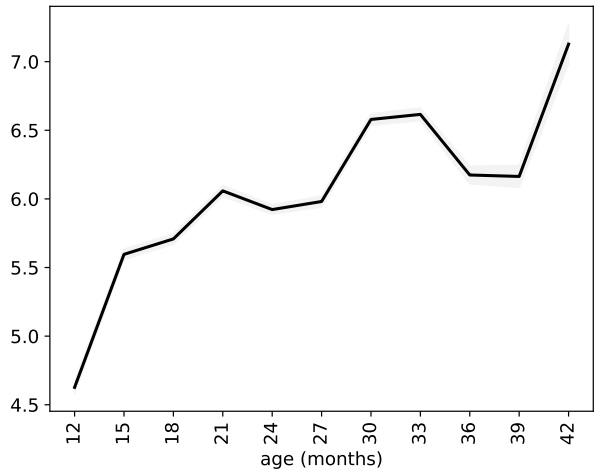

Figure 2: *The mean changes in utterance length (in words) depending on the age of a child listener in the Providence Corpus.*

Table 1: *The list of age-dependent CDS features.*

Feature	Spearman rho	p-value
$F0_{utterance}$ mean	-0.96	< 0.001
$F0_{word}$ first	-0.98	< 0.001
$F0_{word}$ max	0.65	0.03
$F0_{word}$ sd	0.67	0.02
N. of words	0.89	< 0.001
speech rate	0.75	0.01

2.2. Prosodic features

The prosodic analysis of speech data was based on the analysis of F0 values within vowels only. The F0 values were calculated with autocorrelation technique by means of OpenSMILE toolkit [16]. The F0 contour of an utterance was smoothed by Savitzky-Golay filtering using 3rd order polynomials in 5-sample windows. The g2p transcription of orthographic scripts as well as alignment of sound and word segments with speech were performed using WebMAUS online toolkit [17].

We investigated two sets of melodic features: utterance-level and word-level features. The utterance-level features $F0_{utterance}$ included measures of F0 within the whole utterance (vowels only):

- $F0_{utterance}$ max, min, mean, standard deviation within an utterance (in semitones),
- $F0_{utterance}$ range – the difference between $F0_{utterance}$ max and $F0_{utterance}$ min (in semitones),
- $F0_{utterance}$ rise – the max F0 rise within an utterance (in semitones),
- $F0_{utterance}$ fall – the max F0 fall within an utterance (in semitones).

To normalize the data, the $F0_{utterance}$ mean was calculated in semitones relative to an overall speaker F0 mean, while other measures were calculated in semitones relative to a $F0_{utterance}$ mean.

The word-level features were aimed to represent F0 peaks related to accented words within an utterance, which is a technique usually applied for modelling upper declination [18]. $F0_{word}$ of each word was calculated as the maximum F0 within the word. All $F0_{word}$ values in an utterance were calculated in semitones relative to the $F0_{word}$ value of the first word in the given utterance. As a result, we obtained a set of word-level $F0_{word}$ measures for each utterance:

- $F0_{word}$ max, min, mean, standard deviation within an utterance (in semitones),
- $F0_{word}$ final tone - as the distance in $F0_{word}$ between the final and the penultimate word (in semitones),
- $F0_{word}$ declination - as the distance in $F0_{word}$ between the first and the final words (in semitones),
- $F0_{word}$ first – the first word F0 max value in Hz, which makes them speaker-dependent (in Hz).

Additionally, we measured the utterances length and speech rate to take into account their potential effect on melodic variability. Utterance length was determined by the number of orthographic words, including prepositions and articles. The speech rate was measured as mean duration of speech sounds in the utterance defined by automatic aligner per second.

2.3. Analysis of age-dependency

The whole dataset of 123 000 utterances was quantized into 3-month bins (12, 15, 18,..., 42 months) to ensure a proper number of samples per each bin[1]. There were samples from each speaker in each age bin. The age-dependency of each prosodic feature was tested by means of the Spearman's rank correlation coefficient between quantized child age and the feature values associated with the age bin [19]. The age-dependency of many linguistic properties of CDS is monotonic but non-linear, and this should show up in the Spearman rank correlation. Marginal 5 % percentiles of the samples in each age bin were discarded to remove potential outliers.

3. Results

Table 1 shows the melodic features of mother's CDS that were found to be significantly dependent on the child age. There is

[1]Feature data and analysis scripts are openly available at https://github.com/SPEECHCOG/CDS_pitch.

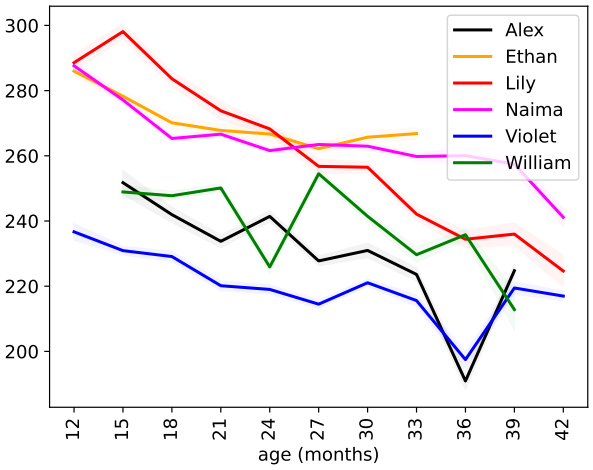

Figure 3: *The changes of max F0 of the first word in an utterance (in Hz) depending on the age of a child listener for each speaker in the Providence Corpus.*

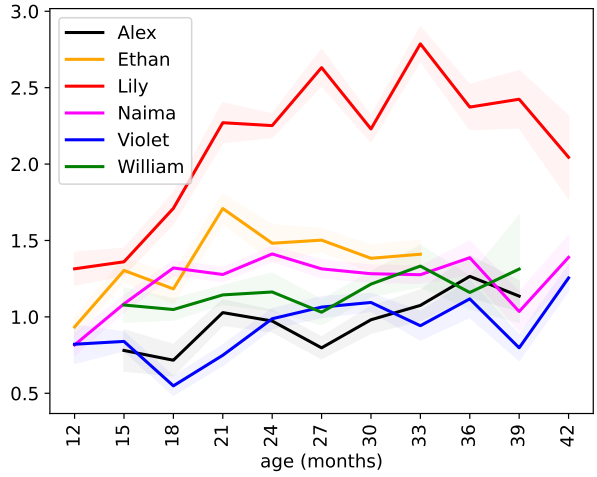

Figure 4: *The changes of the max F0$_{word}$ (in semitones) depending on the age of a child listener for each speaker in the Providence Corpus.*

Table 2: *The list of age-dependent CDS features depending on the utterance length (in words).*

Feature	**Utterance length**	**Spearman rho**	**p-value**
F0$_{utterance}$ mean	2	-0.92	< 0.001
	3	-0.96	< 0.001
	4	-0.95	< 0.001
	5	-0.95	< 0.001
	6	-0.94	< 0.001
	7	-0.83	0.002
F0$_{word}$ first	2	-0.98	< 0.001
	3	-0.95	< 0.001
	4	-0.96	< 0.001
	5	-0.96	< 0.001
	6	-0.96	< 0.001
	7	-0.93	< 0.001

a significant increase in both CDS utterance length and speech rate when a target child is growing up from age 1;0 to 3;6 (see Figures 1 and 2). In this period, the mean utterance length increases by 50 %, from 4.5 words up to 7 words per utterance. The speech rate increases by 15 %, from about 10.5 up to 12 sounds per second. The speed of changes is not constant—the changes over the second year of a child life are twice larger than those taking place over the third year of a child's life (60 % for utterance length and 70 % for speech rate). The consequence of the utterance length increase is that there are less and less one-tone-unit utterances in CDS. The precise analysis of CDS intonation would require high-quality automatic segmentation and labelling of tone units in utterances. Meanwhile we present general statistics on melodic features and the most easy-to-define accents: the first and the last accent in an utterance.

The only utterance level F0 measure of CDS which significantly changes over the studied age range is the mean F0. There is no significant change in within-utterance variability of F0, as expressed by F0$_{utterance}$ standard deviation or by F0$_{utterance}$ range. The mean F0 gradually decreases with age for children between 1 and 3 years of age. This might be affected by an increasing number of both prosodic words and tone units in the utterances.

A closer analysis of the first accent in an utterance, which is the least affected by intonational structure of the utterance and is expressed by F0$_{word}$ of the first word in the utterance, reveals the same behaviour as mean F0$_{utterance}$ measure (see Figure 3). Figure 3 shows changes of F0$_{word}$ of the first word in an utterance for each of the mothers and as a function of their children's age. The range of changes within 2.5 years is up to 4 semitones (see, e.g., Lily's mother's F0 as it decreases by 60 Hz—from 290 Hz to 230 Hz).

There are no significant changes in within-utterance F0 variability. Although there is a statistically significant increase of F0$_{word}$ max measure, the increase is barely perceivable—less than 0.5 semitone in 2.5 years. Still, the finding indicates that there is an increase in the proportion of utterances where the most prominent accent is placed on other than the first word of an utterance. There is also a statistically significant increase of F0$_{word}$ sd measure. Most probably the reason for this increase is the same as for increase of F0$_{word}$ max, but the magnitude of the change is even less than for the F0$_{word}$ max measure. We speculate that the reason for that is an increased syntactical complexity of utterances in CDS addressed to older children. The increased number of syntactic constituents in longer utterances would require more intermediate phonological phrases to be pronounced and some proportion of them will have non-final rising tones. A larger number of non-final rising tones would increase F0$_{word}$ max.

All the other investigated melodic features of CDS didn't show any significant dependency with age. This is especially interesting for F0$_{utterance}$ min, as it should be influenced by utterance lengthening in terms of word count, as the length of the utterance has been shown to influence F0 min values due to melodic declination (see [20]).

We conducted another analysis of the same data controlling the utterance length, i.e. we compared prosodic features of the same-length utterances over different age bin. It was done for utterances of the range from 2 to 7 words in an utterance. The results showed that there are no anymore significant age-dependency of F0$_{word}$ max and F0$_{word}$ sd for any of utterance lengths, see Table 2. The significant age-dependency

of F0 mean and the first F0 accent preserved as high for all of utterance lengths. The speech rate also showed significant age-dependency in this setup.

4. Discussion

The obtained results complement the existing literature on age-dependent changes in CDS, replicating a decrease in mean F0 with increasing child age. The measures of intonation variability, including F0 SD and range and measures for F0 accents and movements, did not show change as a function of child age.

The generalizability of our results is limited by two major factors. First, Providence Corpus contains no recordings during the first year of child's life, thus potential changes in pitch variability could happen during that period, especially since the role of CDS gradually changes from affect-conveying and attention-orienting towards dyadic linguistic exchanges. To study this, a longitudinal speech data carefully collected from similar communicative contexts but young age groups would be required. Second, the number of speakers in the corpus is limited to six and the obtained results show inter-speaker variability of melodic features. From one hand, our finding does not support previously published results on mothers' pitch variability changing during the first years of child life. On the other hand, the results do not contradict any previously published findings that CDS has a larger pitch variability than ADS.

We argue that the observed age-dependency of $F0_{word}$ max and $F0_{word}$ SD might be influenced by the increasing number of words in utterances addressed to older children, as the age-dependency disappears when we control utterance length. The mean pitch and therefore its general variability are dependent on the utterance length due to melodic declination [21]. The correlation of declination factor and general pitch properties of an utterance is not straightforward, as the declination factor may itself depend on utterance length and be lower for longer utterances [22, 20]. As for CDS, the lower mean F0 of longer utterances was noted by McRoberts and Best [4], although they didn't presented experimental results on this finding and didn't reveal any correlation between utterance length and pitch variability.

Both mean F0 and F0 of the first word accent in an utterance show a significant decrease over age of a target child. The effect of age on initial pitch of the utterance in our results is a joint effect of child age and longer utterances, as it was shown previously that longer utterances have higher initial pitch to compensate for melodic declination [22]. Thus, if we would normalize the effect of the utterance length, the lowering of initial F0 with child age would be even larger. The age-effect is smaller for some speakers (Violet's and Ethan's mothers), while it is large for others. These findings support the results published in [4] and [8] for American English and contradict those published in [12] for Dutch. They do not contradict the findings on the increase of F0 during the first year of child's life, as there are no recordings in Providence Corpus for this age. As our results show the monotonic decrease of F0 in mother's speech with child age, an open question is by what child age does the intonation and F0 height converge to that of normal ADS. The answer to this question would require another longitudinal speech recordings beyond 4 years of age.

Regarding a speech rate, there is a difference between our results and the ones obtained by Ko [10] previously on the same corpus. We used a number of pronounced sounds per second as a measure of speech rate, while Ko used a number of pronounced word per second. Both methods are valid, having their are own pros and cons [23] and a general conclusions about age-dependency of the speech rate and about non-linearity of this dependency supported by both measures.

5. Conclusions

Our analyses show an age-dependent trend in CDS for an increasing speech rate and utterance length as children get older. The increase in utterance length is almost monotonic, whereas speech rate data show strong fluctuations over child age, potentially due to a small number of mother-infant dyads in the present corpus.

As for pitch, the mothers' pitch in CDS lowers significantly during the second and the third years of their children's life, thereby being largely in line with earlier reports. However, the results show no significant dependency between pitch variability of mother's speech and child's age. We observe a mild age-dependency of the max F0 word accent. However this dependency is not observed if we control for the number of words in the utterances, e.g., taking into account only 3-word or 4-word utterances. The pitch height changes up to 4 semitones for some speakers, indicating substantial changes in terms of absolute F0 in Hz.

6. Acknowledgements

This research was supported by L-SCALE project funded by Kone Foundation and by Academy of Finland project no. 314602.

7. References

[1] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants," *Developmental Review*, vol. 27, no. 4, pp. 501–532, 2007.

[2] D. D. Albin and C. H. Echols, "Stressed and word-final syllables in infant-directed speech," *Infant Behavior and Development*, vol. 19, no. 4, pp. 401–418, 1996.

[3] 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.

[4] G. W. McRoberts and C. T. Best, "Accommodation in mean f0 during mother–infant and father–infant vocal interactions: A longitudinal case study," *Journal of Child Language*, vol. 24, no. 3, pp. 719–736, 1997.

[5] N. B. Ratner, "Durational cues which mark clause boundaries in mother–child speech," *Journal of Phonetics*, vol. 14, no. 2, pp. 303–309, 1986.

[6] K. Mády, U. D. Reichel, A. Kohári, and A. Szalontai, "The role of accommodation in expressing emotions to newborn babies," in *Proc. 1. International Conference on Tone and Intonation*, 2021.

[7] M. Soderstrom, M. Blossom, R. Foygel, and J. L. Morgan, "Acoustical cues and grammatical units in speech to two preverbal infants," *Journal of Child Language*, vol. 35, no. 4, pp. 869–902, 2008.

[8] D. N. Stern, S. Spieker, R. Barnett, and K. MacKain, "The prosody of maternal speech: Infant age and context related changes," *Journal of Child Language*, vol. 10, no. 1, pp. 1–15, 1983.

[9] D. Raneri, K. Von Holzen, R. Newman, and N. B. Ratner, "Change in maternal speech rate to preverbal infants over the first two years of life," *Journal of Child Language*, vol. 47, no. 6, pp. 1263–1275, 2020.

[10] E.-S. Ko, "Nonlinear development of speaking rate in child-directed speech," *Lingua*, vol. 122, no. 8, pp. 841–857, 2012.

[11] C. Kitamura and D. Burnham, "Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year," *Infancy*, vol. 4, no. 1, pp. 85–110, 2003.

[12] T. Benders, "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent," *Infant Behavior and Development*, vol. 36, no. 4, pp. 847–862, 2013.

[13] K. Demuth, J. Culbertson, and J. Alter, "Word-minimality, epenthesis and coda licensing in the early acquisition of English," *Language & Speech*, vol. 49, no. 2, pp. 137–173, 2006.

[14] M. Lavechin, Y. Sy, H. Titeux, M. A. C. Blandón, O. Räsänen, H. Bredin, E. Dupoux, and A. Cristia, "BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models," in *Proc. INTERSPEECH 2023*, 2023, pp. 4588–4592.

[15] B. MacWhinney, "The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database," 2000.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[17] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[18] V. J. Van Heuven and J. Haan, "Phonetic correlates of statement versus question intonation in Dutch," in *Intonation: Analysis, modelling and technology*. Springer, 2000, pp. 119–143.

[19] C. Spearman, "The proof and measurement of association between two things," *American Journal of Psychology*, vol. 15, pp. 72–101, 1904.

[20] D. Kocharov, N. Volskaya, and P. Skrelin, "F0 declination in Russian revisited," in *18th International Congress of Phonetic Sciences, ICPhS 2015*, 2015.

[21] J. Vaissière, "Language-independent prosodic features," in *Prosody: Models and measurements*. Springer, 1983, pp. 53–66.

[22] J. Yuan and M. Liberman, "F0 declination in English and Mandarin broadcast news speech," *Speech Communication*, vol. 65, pp. 67–74, 2014.

[23] J. Trouvain, "Tempo variation in speech production: Implications for speech synthesis," PhD thesis, Universität des Saarlandes, 2004.