Measuring prosodic predictability in children's home language environments

Kyle MacDonald (kyle.macdonald@techlabs.mcd.com)
Speech and NLU Core Technology, McD Tech Labs

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University

Marisa Casillas (marisa.casillas@mpi.nl)
Language Development Department, Max Planck Institute for Psycholinguistics

Anne S. Warlaumont (warlaumont@ucla.edu)
Department of Communications, UCLA

Abstract

Children learn language from the speech in their home environment. Recent work shows that more infant-directed speech (IDS) leads to stronger lexical development. But what makes IDS a particularly useful learning signal? Here, we expand on an attention-based account first proposed by Räsänen et al. (2018): that prosodic modifications make IDS less predictable, and thus more interesting. First, we reproduce the critical finding from Räsänen et al.: that lab-recorded IDS pitch is less predictable compared to adult-directed speech (ADS). Next, we show that this result generalizes to the home language environment, finding that IDS in daylong recordings is also less predictable than ADS but that this pattern is much less robust than for IDS recorded in the lab. These results link experimental work on attention and prosodic modifications of IDS to real-world language-learning environments, highlighting some challenges of scaling up analyses of IDS to larger datasets that better capture children's actual input.

Keywords: prosody; infant-directed speech; language acquisition; computational reproducibility

Introduction

Children learn language by attending to the speech of those around them. Observational studies show that children who hear more language addressed to them show stronger lexical development (Weisleder & Fernald, 2013), and experimental work finds that infants preferentially attend to language-like signals (Vouloumanos & Werker, 2007) and infant-directed speech (IDS) in particular (Cooper & Aslin, 1990). However, real-world auditory environments are complex, containing a wide variety of sounds. How do infants figure out what parts of the acoustic input are relevant, and what causes them to attend to those signals preferentially?

One hypothesis is that speech directed to children has prosodic properties that are particularly good at getting and maintaining their attention (for a review, see Soderstrom, 2007). Descriptive work on IDS from targeted, lab-based recordings finds that, compared to adult-directed speech (ADS), IDS tends to be (a) higher in pitch, (b) more variable prosodically, and (c) contain hyper-articulated vowels and longer pauses (see Pretzer, Lopez, Walle, & Warlaumont (2019) for a discussion of the prosodic characteristics of IDS recorded in the home). IDS utterances also tend to be shorter, less complex syntactically, and contain more repetition as compared to utterances in ADS. Moreover, recent computational modeling work by Räsänen, Kakouros, & Soderstrom (2018) showed that IDS prosody is less predictable than ADS, even after normalizing the two registers such that differences in average pitch and pitch range were minimized.

Räsänen et al. (2018) propose that less predictable prosodic sequences could grab attention because they violate children's expectations of what acoustic information is likely to appear next. There has been much empirical and theoretical work on the link between predictability of an event and learners' attention. For example, eye-tracking work shows that when adults watch a movie, they tend to look at areas of the screen that most violate their expectations, as opposed to salient areas (e.g., high luminance) or random (e.g., television fuzz) (Itti & Baldi, 2009). Moreover, experimental work shows that 7-month-old infants attend more to stimuli of intermediate uncertainty for both auditory and visual inputs as opposed to completely random or entirely predictable sequences (Kidd, Piantadosi, & Aslin, 2012). And classic models of stimulus-driven learning are built on the principle that the probability of learning scales with the magnitude of surprise of an event (Rescorla & Wagner, 1972).

A key idea of this prior work is that the predictability of an event cannot be measured in isolation, and instead is linked to an individual's prior experiences and the immediately preceding context. Connecting these ideas to theories of early language acquisition, IDS should be particularly engaging to children when it deviates from the acoustics of the speech they have previously heard, including speech not directed towards them (ADS). Räsänen et al. (2018) demonstrated this for lab-recorded adult speech, showing that the predictability of IDS pitch trajectories to be lower than for ADS contours.

An important, open question for this predictability-based account is whether IDS prosody is less predictable than that of ADS in the home language environment. The present study asks whether these lab-based results will generalize to the IDS that infants hear over the course of a day in complex and messy real-world learning environments.

©2020 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

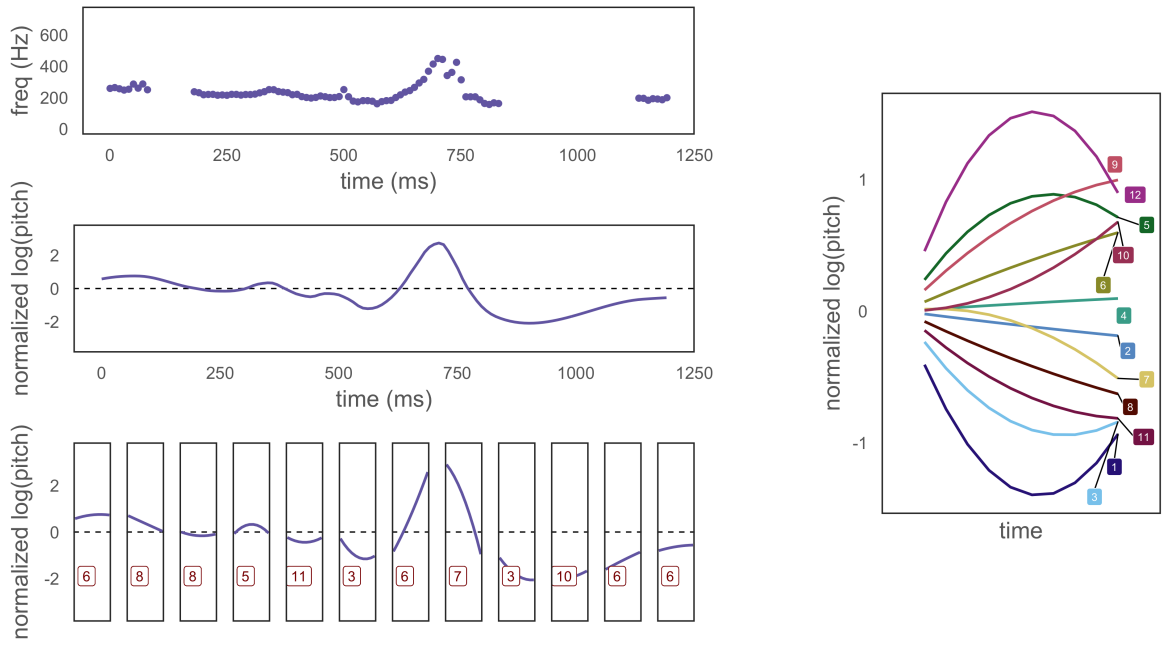

Figure 1: Pitch processing pipeline. The left panels show the primary steps for converting the raw acoustic signal to temporal sequences of pitch shapes that the neural network model will learn to predict. Each point in the top left panel represents a pitch estimate extracted from the utterance using the soundgen R package. The curve in the middle left panel shows the normalized pitch contour after interpolation via local polynomial regression. The bottom left panel shows the sequence of pitch shape categories for each 100 millisecond time bin for this utterance. The right panel shows the pitch shape corresponding to each category generated by k-means clustering on the polynomial coefficients in each time bin, for k = 12). Each curve represents the predicted pitch shape generated by using the polynomial coefficients capturing the central tendency of each cluster.

Current Study

Before analyzing speech in the home environment, we attempt to reproduce the key finding from Räsänen et al. (2018) – that lab-recorded IDS is less predictable than ADS. That is, we perform a computational reproducibility study by implementing our pitch extraction and computational modeling pipeline to provide evidence that our implementation is capable of estimating differences in prosodic predictability in clean, lab recordings before asking questions about the more complex home auditory environment.

Then we ask whether, and to what extent, these lab-based results would generalize to the language that children experience in their day-to-day lives. This is an important question because the lab recordings were focused on a particular communicative context where caregivers were given the explicit goal to talk about concrete objects. Thus, it is possible that the goal of directing children's attention would lead to overestimation of the differences between IDS and ADS since attentional control is the proposed functional role of pitch modification. Also, the lab recordings contain little background noise, distractions, and other-directed conversations. So while the lab-based results are interesting and have advantageous controls, we do not yet know the extent to which children might hear surprising (i.e., attention-grabbing) pitch contours in their naturalistic input.

To estimate differences in lab-based and home IDS vs. ADS, we leveraged two existing, open datasets. Following Räsänen et al. (2018), we used the ManyBabies dataset, which contains speech stimuli of North American English created for a replication of chidren's IDS preferences across a large number of labs (The ManyBabies Consortium, 2017). For the home recordings, we used the IDSLabel dataset (Bergelson et al., 2019), which consists of subsets of utterances from daylong audio recordings of children learning North American English, most of which are openly available in HomeBank (VanDam et al., 2016). The IDSLabel utterances were automatically identified by LENA software as adult vocalizations taking place close in time to infant vocalizations. Then human annotators filtered out cases where the audio did not appear to represent adult vocalization and labeled the adult utterances as IDS vs. ADS.

We hypothesized that home-recorded IDS pitch would be less predictable as compared to ADS, showing similar patterns to the lab recordings. However, we thought that the size of the IDS-ADS difference might be reduced since home recordings capture a broader set of communicative contexts.

Measuring Prosodic Predictability

Pitch Estimation The behavioral data analyzed in this study are time series of measurements of the estimated fundamental frequency (f_0) of caregiver utterances (see Figure 1 for an overview). Throughout the paper, we use $\log(f_0)$ and refer to this measurement as pitch because psychophysical tasks show that changes in $\log(f_0)$ track with subjective judgments of perceived pitch.

We first extracted pitch estimates for voiced regions of the utterance using the soundgen R package (Anikin, 2019). We then normalized the log-transformed pitch estimates such that each speech register had a mean of zero and a standard deviation of one. This allowed for a stronger test of IDS vs. ADS predictability by controlling for IDS having higher average pitch and often higher pitch variability. Finally, we interpolated across unvoiced regions of the utterances by fitting a local polynomial regression, which generates a complex, non-linear curve by fitting separate, simple models to subsets of the data. We chose parameters for the loess such that the interpolated curve would not vary wildly between voiced regions of the utterance (a second degree polynomial with a smoothing factor of 0.2). We also removed any unvoiced regions so as not to interpolate beyond measured aspects of the speech signal.

Temporal Segementation and Quantization We converted the continuous pitch contours to a quantized form by fitting a second degree polynomial to each 100-millisecond segment of audio and extracting the linear and quadratic coefficients for each fit. One hundred milliseconds was the bin width used in Räsänen et al. (2018), a choice based on keeping a similar temporal resolution as syllable-based segmentation algorithms.

We then used k-means clustering to divide the two-dimensional space of the polynomial coefficients into k categories of pitch shapes. This clustering step allowed us to represent changes in pitch using a single number that mapped to the centroid of each cluster, capturing the slope and degree of curvature within each time bin. The choice of k controls the number of different pitch shapes that can be represented, with higher values allowing for more fine-grained and complex shapes. Figure 1 shows an example of the range of pitch shapes that can be represented using this approach (e.g., rising/falling and hills/valleys). To ensure that our results were not sensitive to the choice of k, we selected a range of values (6, 12, and 24) and averaged the results across each.

Quantifying Prosodic Predictability We modeled temporal changes in pitch using deep neural networks (DNN) with a specific architecture designed for learning sequences: Long Short Term Memory Networks (LSTM) (Hochreiter & Schmidhuber, 1997). At a high level of description, the model can leverage prior context by using loops to pass information from previous states of the model to the next state. Here, we measure the accuracy of the LSTM in predicting the next pitch shape when processing IDS vs. ADS.

We trained and tested the LSTM using a variant of k-fold cross-validation. For each fold, we sampled a different held-out test dataset of 10% of the utterances, with the other 90% being used for model training. To ensure the stability of our findings, we ran the entire analysis pipeline five times and averaged the results, and we trained the model across five different values for the proportion of IDS in the input: 0, 0.25, 0.5, 0.75, and 1. The value of 0.5 means that the durations of IDS and ADS in the training data were equivalent to zero being all ADS and one being all IDS.

We also directly compared the LSTM against a baseline time series model. This baseline predicts that the world does not change and that the next pitch shape will be the same as the immediately preceding value (i.e., a persistence model). This comparison allowed us to quantify the importance of using a DNN that can learn more complex relationships between prior prosodic context and upcoming pitch trajectories.

We used the `lme4` package (Bates, Mächler, Bolker, & Walker, 2014) to fit mixed-effects regression models. The mixed-effects approach allowed us to model the nested structure of our data – multiple utterances for each caregiver and speech register – by including random intercepts for each speaker, and a random slope for each speech register. We used Bayesian estimation to quantify uncertainty in our point estimates, which we communicate using a 95% Highest Density Interval (HDI). The HDI provides a range of credible values given the data and model. Since our key dependent measure is the probability of a correct pitch prediction (a value bound between [0, 1]), we used Beta regression. Our key prediction is that the LSTM will make fewer correct predictions for IDS compared to ADS. We interpret the model's lower accuracy as an index of surprisal, a feature known to attract attention in behavioral experiments with infants (Kidd et al., 2012) and adults (Itti & Baldi, 2009).

Participants and procedure

Lab recordings Following Räsänen et al. (2018), we used the ManyBabies dataset. Recordings were collected in a lab observation setting during a 20-minute session. In each session, caregivers were given a bag containing a set of familiar (e.g., ball) and novel objects (e.g., a whisk) and were asked to talk about each object, one at a time, either to their child (IDS) or to the experimenter (ADS). The ManyBabies data can be downloaded from: {`https://osf.io/re95x/`}.

The dataset contained only English recordings taken from mothers in Canada (n = 11). We included all utterance types (familiar, novel, and no label) but filtered utterances with fewer than 10 valid pitch estimates after pitch extraction. After this filtering, there were a total of 1074 utterances (677 IDS, 397 ADS). The average length of each utterance was 2.43 seconds for IDS and 4.92 seconds for ADS.

Home recordings We used the IDSLabel Homebank dataset (available to Homebank members at `https://homebank.talkbank.org/access/Password/IDSLabel.html`). This corpus contains data from the day-

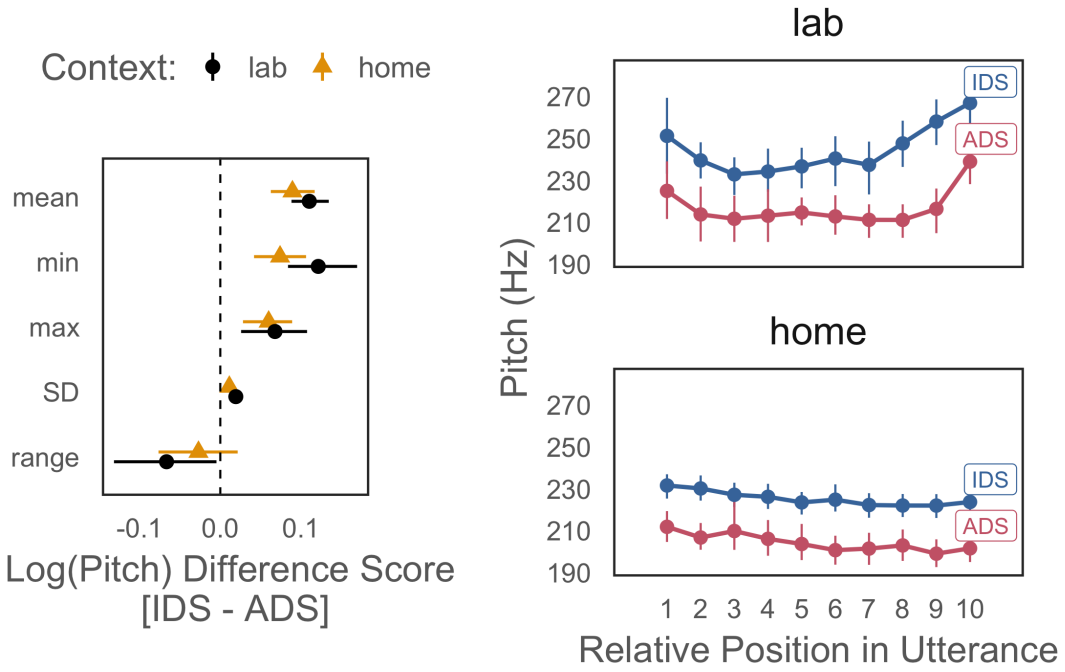

Figure 2: Pitch analyses. The left panel shows differences in log-transformed pitch between IDS and ADS for five summary statistics in both the lab and home recording environments. Each point represents the average difference (IDS - ADS) for that statistic. Larger difference scores indicate a higher value for IDS. The right panel shows changes in the average pitch for IDS (blue) and ADS (red) as a function of position within utterances for both the lab (top) and home (bottom) recording contexts. Error bars are 95% confidence intervals computed via non-parametric bootstrap.

long recordings of 61 participants, which was was created by sampling from four larger corpora in the HomeBank database (VanDam et al., 2016; `homebank.talkbank.org`). All data were collected using the LENA audio recorder, which was worn by children in specialized clothing and recorded a full day of audio. Caregivers were given minimal instructions about what tasks to do on the day of the recording.

Bergelson et al. (2019) used LENA's utterance segmentation and speaker diarization algorithms to identify audio segments that were likely to be caregiver speech. Next, trained, human annotators listened to each clip and coded the speaker gender (male/female) and addressee (child/adult) using primarily acoustic-phonetic information. Coders were instructed to code whether the IDS register was being used (i.e., whether it "sounds like" IDS) even if the addressee was not a target child. After sampling, preprocessing, and human annotation, the final dataset included 10822 utterances produced by adults (6624 IDS, 4198 ADS). If the clip appeared to have been incorrectly labelled as adult speech, that was also indicated by the annotators and those clips are excluded from the present analyses. For more details about sampling plan, preprocessing, and guidelines for human annotations, see Bergelson et al. (2019). The average length of an utterance in seconds was 1.31 for IDS and 1.51 for ADS.

Results and Discussion

First, we present a set of standard pitch analyses, comparing IDS to ADS in both the lab and home contexts using five summary statistics: average, minimum, maximum, range, and standard deviation. Then, we compare the output of the computational models – the tendency to predict the correct pitch shape – to ask whether IDS pitch is more predictable than ADS in lab-based recordings and whether this difference generalizes to recordings made in the home environment. Specifically, we compare the LSTM to random guessing and a naive time series model to quantify the value using more prior prosodic context to predict upcoming pitch changes.

Standard Pitch Analyses To quantify differences in pitch across contexts, we fit the following model: $log(pitch metric) \sim speech register * recording context + (1 + speech register | participant id)$. IDS had a higher average, minimum, maximum, and standard deviation in both the lab and home environments (Figure 2; all $p < .001$).

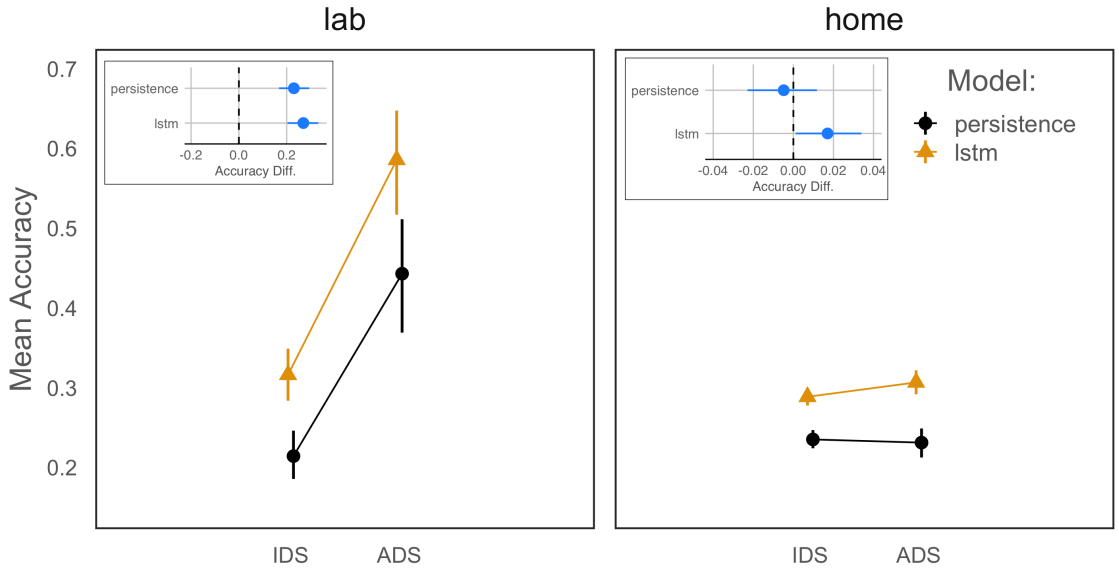

Figure 3: Accuracy results for the naive baseline (persistence) and LSTM models. Each point represents the mean accuracy in predicting the correct subsequent pitch shape. Color and shape represent the different modeling approaches. The two inset plots show the difference in accuracy between IDS and ADS with higher values indicating more predictable pitch in ADS. Error bars are 95% confidence intervals computed via nonparametric bootstrap.

Pitch range, however, was only reliably different in the lab recordings and not in the home context. Moreover, the average, min, max, and standard deviation was higher for the lab recordings (all $p < .001$), which could be due at least in part to the fact that the home recordings include both male and female caregiver utterances whereas the lab recordings included only female voices; other possible reasons could have to do with differences in the content, function, and context of the home-recorded utterances.

The right panel of Figure 2 shows the time course of pitch across the utterances where each time bin represents 10% of the utterance. IDS was higher in pitch at all points in the utterance across both recording contexts. The lab recordings, however, showed a different temporal pattern. Visual inspection suggests that pitch tended to increase towards the end of the lab-recorded utterances especially for IDS (see positions 8 through 10 in Figure 3). In contrast, the pitch of both IDS and ADS in the home tended to decrease over the course of utterances. This difference in temporal patterns across the lab and home provides some indication that the lab recordings were capturing a specific communicative context.

Taken together, these results suggest that we were able to successfully extract and estimate pitch in the more complex and noisier home recording context – a useful result for future work using automated pitch estimation with daylong audio data – and that patterns of IDS vs. ADS differences in basic pitch measures discovered in a controlled lab context generalize to real-world input experienced by infants.

Predictability of IDS vs. ADS We turn now to the question of whether normalized pitch trajectories are less predictable for IDS than for ADS, controlling for differences in mean pitch and pitch variability. To better understand the value of using a neural network modeling framework for measuring predictability, we compared the accuracy of the LSTM's predictions to a baseline time series model. To quantify differences in predictability we fit a linear mixed-effects regression predicting the accuracy pitch predictions as a function of speech register, context, and model type: $accuracy \sim speech\,register * context * model + (1 + speech\,register|participant\,id)$.

Overall, the LSTM outperformed the persistence model (β = -18.41, p = .001). Moreover, there was an interaction between model type and context such that the LSTM was critical for estimating predictability differences between IDS and ADS in the home recordings (β_{int} = -5.12, $p < .001$). This result means that considering more than just the prior 100 ms improves the model's ability to predict upcoming trajectories.

Next, we address our primary question of interest – whether IDS would be less predictable than ADS, and whether this effect would generalize to the home language environment. Pitch trajectories of both lab- and home-recorded IDS were more challenging to predict than pitch trajectories of lab-recorded ADS (β = -2.73, p = .006). In addition to the qualitative difference, our quantitative results are similar to those reported in Räsänen et al. (2018) for both speech registers. This finding indicates a successful replication of the key result from Räsänen et al. (2018) and provides evidence

that our implementation of the pitch estimation and modeling pipeline can detect differences in prosodic predictability across speech registers. These results also confirm our key hypothesis – that IDS in the home has prosodic contours that are likely more surprising to infant learners, providing evidence for the functional role of prosodic modifications present in real-world language input.

We also found an interaction between speech register and context such that the extent of the IDS vs. ADS difference was smaller in the home recordings (β_{int} = -4.96, p = .001). This interaction was driven by the ADS in the home recordings being harder to predict relative to those in the lab recordings (see Figure 3). One plausible explanation for the reduced difference is that the home recordings capture a much wider range of contexts, with a variety of communicative goals. For example, we might not expect a large difference in predictability if the caregiver's goal was to soothe the child, a context that was not captured in the lab recordings. In addition, we might expect a large difference in predictability for emotionally-charged ADS, which may be more likely to occur at home than in the lab.

General Discussion

In this work, we successfully replicated the primary result from Räsänen et al. (2018): that IDS pitch trajectories were less predictable than ADS trajectories. Moreover, we showed that this result generalized to the home language environment, finding that the predictability difference held even after normalizing the pitch trajectories to reduce differences in the average and standard deviation of pitch. This result provides support for attention-based accounts of the prosodic modifications of IDS – that caregivers' modifications to their speech lead to less predictability and consequently higher surprisal of the auditory stimulus, a feature that has been demonstrated to attract infants' attention in lab experiments (Kidd et al., 2012) and that forms the basis of models of learning in the brain (Rescorla & Wagner, 1972).

Second, while we found that while IDS in children's home language environment is indeed less predictable compared to ADS, the extent of the IDS vs. ADS predictability difference was substantially smaller in children's actual language input. One possible explanation for this difference is that the daylong home recording data contain a wide variety of contexts, some of which might be expected to show differing effects on pitch predictability. For example, in the home recordings, we might expect a good deal of IDS to be focused on soothing a fussy infant. Those soothing-oriented IDS utterances would likely show a very different pattern of pitch predictability (Fernald, 1989), perhaps even showing higher predictability, i.e., less surprisal, compared to the average ADS. Similarly, some home-recorded ADS might be much higher arousal than those recorded in the lab; we might expect ADS produced during heated adult arguments, excited greetings, or even excited statements about what an infant just did to have less predictable pitch contours. Higher surprisal of adult pitch contours might help adults to obtain other adults' attention.

A third significant difference is that the duration of the naturalistic utterances, especially ADS, was much shorter than in the lab recordings, and Räsänen et al. (2018) showed that longer utterances were easier for an LSTM to predict. Another possibility could be that both background noise and variability in speaker properties may be greater in the home recordings. All of these factors could lead to lower predictability of home-recorded adult speech overall, and to more sources of variability, thus reducing the effect size for differences between IDS and ADS.

This work has several important limitations. First, both datasets (lab and home) use recordings of English-speaking caregivers in a Western cultural context. It would be interesting to use these methods to ask whether IDS pitch is less predictable in cultural contexts where prior research reports lower amounts of child-directed speech from caregivers. Second, our analysis only focused on a single prosodic variable – pitch – and within-utterance predictability. There are likely other components of speech, such as intensity or the timing of utterances, that caregivers use to direct infant's attention. Third, we only measured the average predictability across the utterance. It could be that IDS and ADS may behave differently in terms of how predictability is distributed in time with IDS being especially unpredictable at the start or end of utterances. Finally, we trained the DNN on group-level data and performed aggregated analysis. Future work could explore training/analyzing model performance on individual children's language input to ask questions about individual differences in children's experiences of IDS vs. ADS.

This work relates lab-based studies on the effects of prosodic modifications to children's home language-learning environments. The results suggest that IDS in naturalistic, infant-centered daylong recordings is less predictable than the ADS, but that this pattern is less robust than for lab-based recordings of caregiver speech. Overall, we are optimistic about the use of novel computational tools (e.g., neural networks) to measure theoretically-relevant features of children's home language environment. And we hope that our exploration here highlights some of the challenges and key future directions for scaling up lab-based analyses to larger datasets that better capture children's actual language input.

Data/code available at
`https://github.com/kemacdonald/lena-pred`
Preregistration at
`https://osf.io/esv8z`

Acknowledgements

We are grateful to the families who participated in this research. Thank you to Curt Chang for helpful discussion. This work was supported by a NSF RIDR grant to AW.

References

Anikin, A. (2019). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, *51*(2), 778–792.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north american babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 1497–1510.

Hochreiter, S., & Schmidhuber, J. (1997). LSTM can solve hard long time lag problems. In *Advances in neural information processing systems* (pp. 473–479).

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399.

Pretzer, G. M., Lopez, L. D., Walle, E. A., & Warlaumont, A. S. (2019). Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe. *Infant Behavior and Development*, *57*, 101325.

Räsänen, O., Kakouros, S., & Soderstrom, M. (2018). Is infant-directed speech interesting because it is surprising?–linking properties of ids to statistical learning and attention at the prosodic level. *Cognition*, *178*, 193–206.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, pp. 128–142). Thieme Medical Publishers.

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–164.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.