Connecting stimulus-driven attention to the properties of infant-directed speech — Is exaggerated intonation also more surprising?

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, 00076, AALTO, Finland

Sofoklis Kakouros (sofoklis.kakouros@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, 00076, AALTO, Finland

Melanie Soderstrom (M_Soderstrom@umanitoba.ca)
Department of Psychology, University of Manitoba, P404 Duff Roblin Building, Winnipeg, MB R3T 2N2, Canada

Abstract

The exaggerated intonation and special rhythmic properties of infant-directed speech (IDS) have been hypothesized to attract infant's attention to the speech stream. However, studies investigating IDS in the context of models of attention are few. A number of such models suggest that surprising or novel perceptual inputs attract attention, where novelty can be operationalized as the statistical predictability of the stimulus in a context. Since prosodic patterns such as F0 contours are accessible to young infants who are also adept statistical learners, the present paper investigates a hypothesis that pitch contours in IDS are less predictable than those in adult-directed speech (ADS), thereby efficiently tapping into the basic attentional mechanisms of the listeners. Results from analyses with naturalistic IDS and ADS speech show that IDS has lower overall predictability of intonation across neighboring syllables even when the F0 contours in both speaking styles are normalized to the same frequency range.

Keywords: language acquisition; infant-directed speech; statistical learning; attention; stimulus predictability

Introduction

Infant-directed speech (IDS) is a speaking style that talkers often use when interacting with young infants. In contrast to adult-directed speech (ADS), IDS tends to have exaggerated intonational contours with higher fundamental frequency (F0) and larger frequency range (e.g., Grieser & Kuhl, 1988), hyperarticulated vowels (Kuhl et al., 1997; but see also Martin et al., 2015), and shorter utterances with a higher token/type ratio (Phillips, 1973). In addition to serving as language input tuned to the developmental stage of the listener (Snow, 1977), one hypothesized role of the exaggerated nature of IDS is that it may engage infants' attention to the speech stream more efficiently than ADS (e.g., Garnica, 1977; Fernald, 1989; see Soderstrom, 2007, for an overview), thereby facilitating language learning from speech.

Although the exaggerated intonation of IDS is often implicitly assumed to be the cause for higher attentional attractiveness, according to our knowledge, no study has systematically evaluated properties of IDS in the context of what is known about perceptual mechanisms for stimulus-driven attention. Instead, the evidence for higher attentional capture of IDS largely comes from behavioral studies that show that infants prefer to listen to IDS over ADS (Fernald, 1985; Cooper & Aslin, 1990; Pegg, Werker & McLeod, 1992). In addition, based on acoustic analyses and their perceptual correlates, IDS is often characterized as more salient or prominent than ADS, therefore also potentially being more interesting to the listeners (e.g., Garnica, 1977; Fernald, 1989). Since stimulus-driven attention and prominence of the perceived speech input seem both to be driven by unpredictability of the stimuli in the given context (see the next sub-section; but see also Kidd et al., 2012), the existing knowledge suggests that IDS might be more attractive to the listeners simply because it has different predictability properties over time than ADS. For instance, larger variability of F0 in IDS already implies, but *does not guarantee*[1], higher uncertainty regarding the realization of the intonation at any moment in time. However, no study has systematically compared the prosodic predictability of IDS and ADS from a statistical learning point of view, even though infants are known to be sensitive to statistical regularities in their perceptual experience (c.f., Saffran et al., 1996; Soderstrom et al., 2009, and references therein) and to the prosodic structure of their native language already from an early age (e.g., Nazzi et al., 1998).

In the present paper, a quantitative investigation is carried out in order to test whether IDS is indeed not just more variable, but also less predictable than ADS, thereby being in line with the recent predictability-based accounts of perceptual attention. Importantly, we assume that the listener is able to learn the typical behavior of intonational contours from speech experience and this creates the basis for prosodic expectations for new speech input. In order to do this, a straightforward computational model of statistical learning is applied to F0 trajectories of naturalistic IDS and ADS and tested in its ability to predict intonational contours on speech utterances from both speaking styles.

Stimulus-driven attention and statistical learning

A number of models for stimulus-driven perceptual attention suggest that attention is drawn to stimuli that are low-probability, or *unpredictable*, in the given context (Itti & Baldi, 2009; Zhang et al., 2008; Tsuchida & Cottrell,

[1] Unless speech is assumed to be a normally distributed IID process without temporal contiguity, a larger F0 range does not guarantee lower temporal predictability (c.f., e.g., a sine wave).

2012; Zarcone et al., 2016), basically enabling the perceptual system to focus on aspects of the environment with the highest information content (Shannon, 1948), i.e., input that is not yet learned and thereby accurately predicted by the brain. However, infants are also known to prefer stimuli that are surprising or novel only as long as the input is not too unlikely in the given context, also known as the Goldilocks effect (Kidd et al., 2012). This suggests that the input should still be structured enough to support learning, thereby providing the basis for statistical expectations and evaluation of the relative information value of the inputs.

Earlier work with prosody perception suggests that low-probability intonation patterns in the context of otherwise predictable prosody are associated with higher perceptual prominence of the concurrent words (Kakouros & Räsänen, 2016a) and alter semantic processing of speech (e.g., Magne et al., 2005), having the same consequences as low-probability words in the given context (see Kakouros et al., submitted, for a discussion). Recent evidence also suggests that adult listeners are sensitive, and rapidly adapt, to changing statistical properties of the intonation patterns, leading to experience-based expectations for prosody whose violations give rise to the subjective impression of prominence (Kakouros & Räsänen, 2016b; Kakouros et al., submitted). Overall, the earlier research indicates that auditory attention and perceptual prominence are connected to the predictability of the prosodic patterns, and this may play a role also in the perception of IDS.

Importantly, the concept of predictability necessitates some type of mechanism for learning regularities from experience, thus connecting attention and prominence with the concept of statistical learning. The most parsimonious assumption would be that the prosodic learning utilizes the same statistical learning mechanisms hypothesized to play a role in other aspects of language acquisition, but now operating at the level of prosodic features such as F0 contours and energy envelopes instead of the phonemic units of the language. Since infants are known to be adept statistical learners, and since prosodic cues are perceptually accessible to them (e.g., Hawthorne, Mazuka & Gerken, 2016), it is likely that infants are sensitive to statistical regularities present at the prosodic level similarly to adults.

If predictability of the stimulus in a given context is a major factor in controlling stimulus-driven attentional orientation, as also exemplified by the widely used preferential head-turn or looking-time paradigms to probe infants' learning, we would expect IDS to have different predictability properties than ADS. In the present study, we will look into one specific aspect of IDS, namely, intonation, and test how well F0 contours can be predicted over time for the two speaking styles in question.

Data

The speech material used in the present experiments comes from the ManyBabies study that aims to replicate IDS preference across a large number of labs (The ManyBabies Consortium, 2017). In the context of that study, naturalistic speech from female caregivers to their infants or from caregivers to other adults was recorded in central Canada and Northeastern US. All caregivers had infants aged 122–250 days. The recordings were carried out in an infant-friendly greeting area/testing room using lapel clip-on microphones connected to smartphones. The task involved describing a closed set of labeled objects by asking the mother to take each object out of a bag one at a time and talk about it to her baby (IDS) or to an experimenter (ADS). In addition, there were two types of objects: those supposedly familiar to the infants (e.g., a ball or a block) and those considered as novel (e.g., a sieve or a whisk). After rough manual segmentation of the recordings into utterances, the utterances were also classified into three categories: utterances containing the familiar object word, those containing the unfamiliar object word, and utterances without naming of the object.

In the present study, we used the Canadian section of the recordings, containing speech from a total of 11 mothers. The US recordings (4 mothers) were excluded due the significant presence of room reverberations that could have impacted automatic F0 estimation. All utterances shorter than 1 s or with less than five syllables (see Methods) were discarded, leading to a total of $N = 882$ utterances (504 IDS, 378 ADS) with an average of 80.2 ± 29.9 utterances per talker. Average utterance length was 4.0 ± 2.5 seconds (3.1 ± 1.1 for IDS, 5.2 ± 3.1 s for ADS).

Methods

The overall goal of the analysis was to compare predictability of F0 trajectories in the IDS and ADS utterances using a statistical model. This was done by first syllabifying and estimating F0 trajectories for all speech, parametrizing F0 trajectories during each syllable, clustering the syllable-specific parameters into a finite number of categories ("F0 shapes") in an unsupervised manner, and then modeling the temporal evolution of these F0 states across time. By training the predictive model from a set of utterances and then computing the likelihoods of F0 trajectories on a set of held-out utterances, measures of F0 predictability can be estimated from the data. Fig. 1 shows a schematic picture of the processing pipeline for an individual utterance. All experiments were conducted in MATLAB unless mentioned otherwise.

Pre-processing of F0 trajectories

F0 trajectories were estimated at a 100-Hz sampling rate with YAAPT-algorithm (Zahorian & Hu, 2008; version 4.0), constraining F0 estimates to the range of 100–600 Hz and using YAAPT's ptch_fix() tool for post-processing of the pitch tracks for potential estimation errors and for interpolation of the trajectories across unvoiced regions. For the predictability analysis, utterance-level F0 tracks were z-score normalized to zero mean and unit variance in order to focus on temporal behavior instead of the absolute mean or range of the pitch. In addition, the original non-normalized F0 contours were used as baseline features in the analyses.

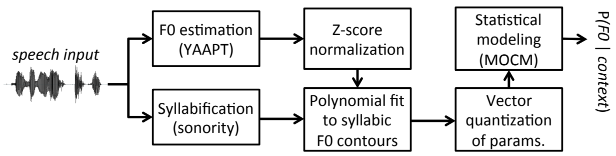
speech input
F0 estimation (YAAPT)
Syllabification (sonority)
Z-score normalization
Polynomial fit to syllabic F0 contours
Vector quantization of params.
Statistical modeling (MOCM)
P(F0 | context)

Figure 1: A schematic view of the F0 predictability analysis. The output is the probability of F0 in syllable s given the observed F0 in m preceding syllables (after training the statistical model on a number of training utterances).

All utterances were syllabified using a sonority envelope-based automatic syllabifier described in Räsänen, Doyle and Frank (submitted; see also Räsänen, Doyle & Frank, 2015, for an earlier but similar version). All syllables without any frames with reliable voicing (as determined by YAAPT) and syllables shorter than 50-ms were merged with the neighboring syllables, leading to a total of 8056 syllables in the data set. Note that although this type of acoustic syllabification is not perfectly accurate in terms of the phonological rules of the language, it still provides systematic chunking of speech into syllable-like units with each unit consisting of a sonorous peak surrounded by less-sonorous onsets and coda (see also, e.g., Villing, Ward & Timoney, 2006, and references therein). Importantly, such acoustic-signal based chunking can be argued to better match the syllabification capabilities of pre-linguistic infants that also must rely on non-phonological acoustic cues in their perception of speech before they master the sound system of their native language (Räsänen et al., submitted).

Following the syllabification, F0 trajectories during each syllable were parametrized by fitting a second order polynomial to the trajectory in time (Fig. 2) and using the polynomial coefficients without the constant term as a parametric description of the F0 during the syllable. Parameters across all syllables in the data were then vector quantized into Q discrete categories using standard k-means clustering with random initialization. In practice, these Q shapes correspond to different F0 patterns with varying curvature and rate of change as a function of time, larger Q simply meaning more fine-grained distinction between F0 patterns that occur during the syllables.

Temporal modeling of F0 state sequences

As a result of the pre-processing, the F0 trajectory of each utterance was described as a sequence of discrete states $q_s \in Q$, one state per syllable s. In order to quantify the predictability of F0, a mixed-order Markov chain model, or MOCM, was trained for the sequences (Saul & Pereira, 1997). Instead of computing n-gram statistics for different n-gram orders and then choosing and/or merging the models with best predictive capability, MOCM allows modeling of varying order Markov chains with a single set of model parameters. In MOCM, the probability of an F0 shape q_s in syllable s, given the preceding m syllables, is calculated as

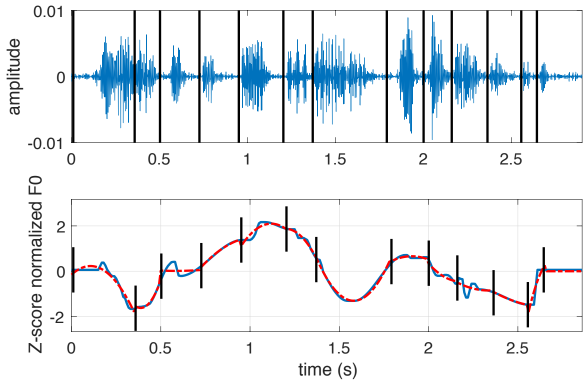

Figure 2: An example of syllable-wise 2nd order polynomial approximation of the F0 trajectory. Top: The original speech waveform. Bottom: YAAPT-estimated and z-score normalized F0 trajectory with interpolation across unvoiced segments (blue solid line) and the corresponding 2nd order polynomial least-squares fit for F0 during each syllable. Syllable boundaries are shown with vertical lines.

$$P(q_s \mid q_{s-1},\ldots,q_{s-m}) = \sum_{k=1}^{m} \lambda_k(q_{s-k})\mathbf{M}_k(q_{s-k},q_s)\prod_{j=1}^{k-1}[1-\lambda_j(q_{s-j})] \tag{1}$$

where lag-specific transition matrices $\mathbf{M}$ and transition weights $\boldsymbol{\lambda}$ are estimated from training data using the Expectation Maximization (EM) algorithm (Saul & Pereira, 1997). In the context of the present study, $\mathbf{M}_k$ describes the transition probabilities between syllabic F0 contours at different lags k while $\boldsymbol{\lambda}$ weighs these probabilities from different distances based on the reliability of the probability estimates in the context of the observed shapes.

In the experiments, a third order (m = 3) MOCM model was trained using the syllabic F0 sequences from 90% of the combined pool of IDS and ADS utterances. This was followed by syllable-by-syllable estimation of F0 likelihoods on the remaining held-out utterances using Eq. (1). The procedure was repeated in a 10-fold manner until all utterances had been used in the training and test sets. The division of utterances into training and testing sets was purely random, and therefore both contained speech from the same 11 unique talkers. We decided not to use speaker-specific models for F0 due to the modest number of utterances per talker that would have caused data sparsity issues in the model estimation. As a result, the obtained probability estimates describe *how expected is the F0 behavior in the given context given a preceding exposure to a large number of F0 trajectories,* low probability reflecting unexpected and thereby attention capturing intonation.

Note that the choice of Q, the number of quantization levels for the F0 shapes, contains an inherent tradeoff between the resolution of the F0 trajectory modeling and the amount of data required for model estimation. Although there is no a priori reason to consider any Q specifically favoring IDS or ADS due to the z-score normalization of all F0 values, we wanted to minimize the impact of Q in our findings. Therefore the simulation was conducted for $Q = 6$, 12, and 24 with syllable-specific likelihood estimates averaged across all these runs. In addition, all likelihoods were averaged across five runs of the entire experiment to diminish variation caused by random initialization of the k-means clustering process, even though the k-means clustering results for the two dimensional features were found to be highly consistent across individual runs.

Data analysis

Five utterance-level statistical descriptors, namely, the mean, SD, min, max, and range (max−min) were calculated for the F0 likelihoods across all syllables and for the original F0 trajectories in Hz in each utterance. Talker and style-specific (IDS vs ADS) means for the descriptors were then averaged across all the utterances from the given talkers. Before any statistical analysis, the statistical descriptors for F0 likelihoods were corrected for the variable amount of matching training data for the speaker and speaking style in question. This was done by first fitting a speaker-independent linear regression model from the number of matching training samples to the statistical descriptors, and then subtracting the prediction from the original values, basically decorrelating the measures with respect to the amount of training data.

In order to test differences between IDS and ADS, the normalized descriptors for F0 predictability and descriptors for the original F0 values were then compared between the IDS and ADS conditions using the paired t-test with significance level of $p < 0.05$ (Holm-Bonferroni corrected for the ten comparisons and $df = 10$ for all reported stats).

Results

Fig. 3 shows a summary of the results together with markers and t-statistics for significant differences between IDS and ADS. As expected, the mean frequency of F0 in the utterances is higher in IDS (210.9 Hz ± 29.0 Hz) than in ADS (189.9 ± 23.9 Hz). In addition, the average utterance-level maximum and minimum F0 are significantly higher in IDS, but the overall variability and absolute range (in Hz) are not different between the speaking styles.

As for the predictability, the mean predictability of F0 in IDS was significantly lower than in ADS ($t(10) = 4.82$, Cohen's $d = 1.93$). In addition, maximum predictability during each utterance was also lower ($t(10) = 5.46$, $d = 2.10$) and so was the range of predictability values across the syllables in the utterances ($t(10) = 5.19$, $d = 1.88$). In contrast, variability of predictability across the utterances was not different between IDS and ADS. Notably, the average F0 probabilities are within a similar range to what was found to be optimal stimulus complexity for attentional capture in the visual perception experiments of Kidd et al. (2012) and significantly above chance-level ($p = 0.0972$). This suggests that the F0 trajectories might be in a suitable complexity region for triggering novelty preference, enabling predictive learning but also leaving room for unpredictable patterns and events.

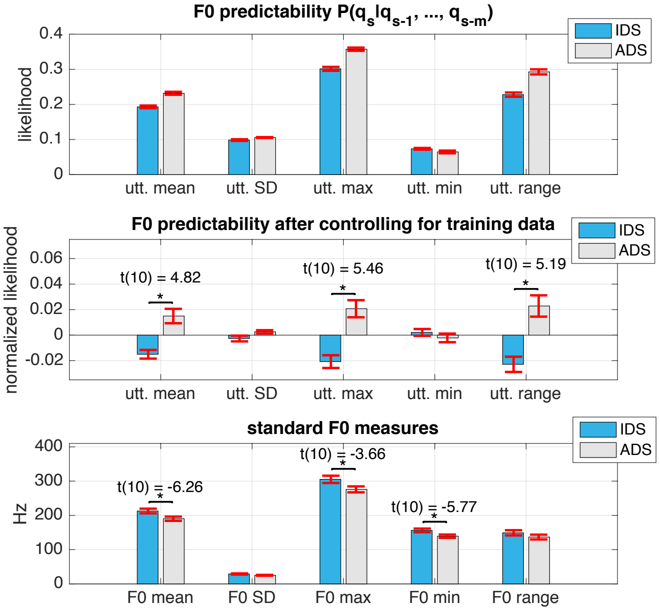

Figure 3: Top: Utterance-level statistical descriptors of F0 predictability, averaged across all ADS/IDS utterances. Middle: F0 predictability after controlling for the amount of matching training data (speaker & style) for each utterance. Bottom: Utterance-level descriptors of original F0 in Hz, averaged across all utterances. Error bars denote ±1 SE across all talkers. Significant differences between IDS and ADS are denoted with asterisks and related t-values (paired t-test, $df = 10$, and using significance level $p < 0.05$ with Holm-Bonferroni correction for the ten comparisons).

We also repeated the entire analysis but now using linear instead of the 2nd order model for the syllabic F0 contours (i.e., encoding only the average direction and rate of change in F0 during the syllable). This replicated all the main findings (significantly lower mean, max, and range for the predictability of F0 in IDS; not shown separately). We also tested whether there were differences in the predictability descriptors between the three sentence types (familiar object, unfamiliar object, no labeling) but none of the tests were significant after controlling for multiple comparisons. In addition, the predictability difference is not simply due to a larger quantization error for IDS parameters, since the reported pattern of results persists also if only the IDS data are used for the k-means codebook creation leading to lower quantization errors (RMSE) for the IDS F0 trajectories.

Overall, the main result confirms the hypothesis that the intonation contours in IDS are less predictable than in ADS, at least for the present data set in question.

As a follow-up validation of the findings, we also ran binary logistic regression to classify all the individual utterances into IDS or ADS classes using the utterance-level descriptors for probabilities and raw F0 values as features and using likelihood ratio as the criterion for forward stepwise feature selection (using SPSS version 23.0, IBM Corp., Armonk, NY). The resulting model achieved IDS/ADS utterance classification rate of 74.8% using a final set of four features: SD of likelihood (Wald statistic = 31.28, $p < 0.001$; $df = 1$ for all features), mean likelihood ($W = 23.08$, $p < 0.001$), likelihood range ($W = 88.34$, $p < 0.001$), together with max of original F0 in Hz ($W = 66.24$, $p < 0.001$). This further shows that the predictability differences of intonation in IDS and ADS do not simply appear as aggregate measures across a large number of utterances, but can be also used to effectively classify individual utterances into ADS or IDS.

Finally, a subset of the utterances used in the present study had been previously rated for their IDS-likeness using a low-pass version of the recordings as part of the ManyBabies project (see The ManyBabies Consortium, 2017, for details). These utterances were rated on a 7-point Likert scale by several naïve raters recruited from Amazon's mechanical Turk. We therefore calculated the correlation between all the utterance-level F0 descriptors and the human IDS-likeness ratings for all the IDS utterances with ratings ($N = 442$). The human judgments of IDS-likeness correlated with the mean (Spearman's $r = 0.25$, $p < 0.001$), SD ($r = 0.31$, $p < 0.001$), min ($r = 0.154$, $p = 0.002$), max ($r = 0.35$, $p < 0.001$), and range ($r = 0.32$, $p < 0.001$) of the original F0 values, i.e., with all of them. Surprisingly, all the descriptors of F0 trajectory likelihoods were uncorrelated with the human ratings ($p > 0.05$ for all comparisons).

Since predictability was nonetheless a reliable cue in our classification of utterances into IDS and ADS based on the original study labels, the finding with the naïve ratings data suggests a dissociation between perceptual correlates of IDS-like speech in naïve listeners (e.g., high and variable pitch) and the lower predictability of intonation in IDS as a potential attractor of listeners' attention. Notably, an earlier study by Singh, Morgan & Best (2002) has also reported that higher and more variable pitch alone was not sufficient to capture infants' attention when pitted against affective speech. This suggests that the properties that make an utterance sound IDS-like to a naïve listener may be unrelated to those that lower the predictability of IDS. How those properties relate to the attentional attractiveness of IDS is currently unclear and requires further investigation.

Discussion and conclusions

This study aimed to test whether the exaggerated intonation in IDS also translates into less predictable prosody over time. The results support this idea, even when the actual mean and range of F0 values in the predictive analysis was normalized between the IDS and ADS utterances. In addition, while IDS intonation is less predictable than ADS, it is still relatively structured as indicated by the mean predictability that is significantly above the chance-level given the analyzed quantization levels. These findings provide initial support to the idea that IDS may be more attentionally attractive simply because it is more surprising without being too chaotic (c.f., Kidd et al., 2012), thereby tapping to the basic attentional mechanisms causing orientation towards unfamiliar inputs.

In addition, some evidence for a dissociation between human ratings of IDS-likeness and predictability of the utterances was also discovered, warranting further research in the issue. In fact, a dissociation between F0 variability and F0 predictability is expected on the basis of predictability-based accounts to prominence and attention in speech. More specifically, it has been argued that the perceptual system should allocate processing resources to the aspects of the input that are not yet predicted by the brain independently of the physical magnitude or other absolute property of the input. In contrast, highly predictable inputs, by definition, have low information value and are therefore low priority targets for sensing and learning even if they have large magnitude on some scale such as loudness or pitch (e.g., Kakouros & Räsänen, 2016b; Kakouros et al., submitted; see also, e.g., Friston & Kiebel, 2009). In the context of speech, the talker can control the listener's attention by freely using non-canonical prosodic forms on any word or words of choice without changing the literal meaning of the utterance (Kakouros et al., 2016b; Kakouros et al., submitted). The present study suggests that caregivers may (implicitly) utilize a similar strategy to maintain infants' attention on the speech stream or highlighting certain segments of speech.

However, the present work only provides an initial investigation into the predictability aspects of IDS using a certain modeling approach. Much more work is needed to understand the underpinnings of IDS and how it relates to learning and attention mechanisms of the human cognition. This also includes the need to replicate the present investigation on different speech data and also preferably with alternative approaches to quantifying suprasegmental statistical structure. In addition, prosody is much more than F0 trajectories, and therefore aspects such as timing, utterance duration, and intensity should be investigated from the predictability point of view independently and in conjunction with F0.

Acknowledgments

O.R. was funded by Academy of Finland project titled "Computational modeling of language acquisition" (274479). All MATLAB scripts for syllabification, mixed-order Markov chains, and reported analyses can be found at https://git.io/vDT0N. The original recordings are available from The ManyBabies Consortium at https://osf.io/re95x/. The authors would like to thank Michael C. Frank for his help with the ManyBabies data.

References

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

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, 8, 181–195.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 60, 1497–1510.

Friston, K. & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks*, 22, 1093–1104.

Garnica, O. K. (1977). Some prosodic and paralinguistic features of speech to young children. In C. E. Snow & C. A. Ferguson (Eds.): *Talking to children: Language input and acquisition* (pp. 63–88). Cambridge, UK: Cambridge University Press.

Grieser, D., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: support for universal prosodic features in motherese. *Developmental Psychology*, 24, 14–20.

Hawthorne, K., Mazuka, R., & Gerken, L. (2015). The acoustic salience of prosody trumps infants' acquired knowledge of language-specific prosodic patterns. *Journal of Memory and Language*, 82, 105–117.

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

Kakouros, S., & Räsänen, O. (2016a). Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features. *Cognitive Science*, 40, 1739–1774.

Kakouros S. & Räsänen O. (2016b). Statistical Learning of Prosodic Patterns and Reversal of Perceptual Cues for Sentence Prominence. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Philadelphia, PA, pp. 2489–2494

Kakouros, S., Salminen, N., & Räsänen, O. (submitted). Making predictable with style – Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. Submitted for publication.

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.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277, 684–686.

Magne, C., Astèsano, C., Lacheret-Dujour, A., Morel, M., Alter, K., & Besson, M. (2005). On-line processing of "pop-out" words in spoken French dialogues. *Journal of Cognitive Neuroscience*, 15, 740–756.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, 26, 341-347.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology*, 24(3), 756–766.

Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development*, 15, 325–345.

Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Development*, 44, 182–185.

Räsänen O., Doyle G. & Frank M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *Proc. Interspeech-2015*, Dresden, Germany, pp. 3204–3208.

Räsänen, O., Doyle, G., & Frank, M. C. (submitted). Pre-linguistic rhythmic segmentation of speech into syllable-like units. Submitted for publication.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.

Saul, L. & Pereira, F. (1997). Aggregate and mixed-order Markov models for statistical language processing. In *Proc. 2nd Conf. Empirical Methods Natural Language Processing*, Providence, RI, USA, Aug. 1997, pp. 81–89.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423.

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy*, 3, 365–394.

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, 4, 1–22.

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

Soderstrom, M., Conwell, E., Feldman, N., & Morgan, J. (2009). The learner as statistician: three principles of computational success in language acquisition. *Developmental Science*, 12, 409–411.

The ManyBabies Consortium (2017). Quantifying the sources of variability in infancy research using the infant-directed speech preference. Manuscript under review. https://osf.io/re95x/

Tsuchida, T., & Cottrell, G. W. (2012). Auditory saliency using natural statistics. *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (CogSci-2012), Sapporo, August 1–4 (pp. 1048–1053).

Villing, R., Ward, T., & Timoney, J. (2006). Performance limits for envelope-based automatic syllable segmentation. *Proc. ISSC*-2006, Dublin, Ireland, pp. 521–526.

Zahorian, S. & Hu H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of The Acoustical Society of America*, 123, 4559–4571.

Zarcone, A., van Schijndel, M., Vogels, J., & Demberg, V. (2016). Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, 7, article no. 844.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8, 1–20.