Cognition 178 (2018) 193–206

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Is infant-directed speech interesting because it is surprising? – Linking properties of IDS to statistical learning and attention at the prosodic level

Okko Räsänen[a,*], Sofoklis Kakouros[a], Melanie Soderstrom[b]

[a] Dept. Signal Processing and Acoustics, Aalto University, P.O. Box 12200, 00076 AALTO, Finland
[b] Department of Psychology, University of Manitoba, P404 Duff Roblin Building, Winnipeg, MB R3T 2N2, Canada

ARTICLE INFO

Keywords:
Language acquisition
Infant-directed speech
Statistical learning
Attention
Stimulus predictability
Prosody

ABSTRACT

The exaggerated intonation and special rhythmic properties of infant-directed speech (IDS) have been hypothesized to attract infants' attention to the speech stream. However, there has been little work actually connecting the properties of IDS to models of attentional processing or perceptual learning. A number of such attention models suggest that surprising or novel perceptual inputs attract attention, where novelty can be operationalized as the statistical (un)predictability of the stimulus in the given context. Since prosodic patterns such as F0 contours are accessible to young infants who are also known to be adept statistical learners, the present paper investigates a hypothesis that F0 contours in IDS are less predictable than those in adult-directed speech (ADS), given previous exposure to both speaking styles, thereby potentially tapping into basic attentional mechanisms of the listeners in a similar manner that relative probabilities of other linguistic patterns are known to modulate attentional processing in infants and adults. Computational modeling analyses with naturalistic IDS and ADS speech from matched speakers and contexts show that IDS intonation has lower overall temporal predictability even when the F0 contours of both speaking styles are normalized to have equal means and variances. A closer analysis reveals that there is a tendency of IDS intonation to be less predictable at the end of short utterances, whereas ADS exhibits more stable average predictability patterns across the full extent of the utterances. The difference between IDS and ADS persists even when the proportion of IDS and ADS exposure is varied substantially, simulating different relative amounts of IDS heard in different family and cultural environments. Exposure to IDS is also found to be more efficient for predicting ADS intonation contours in new utterances than exposure to the equal amount of ADS speech. This indicates that the more variable prosodic contours of IDS also generalize to ADS, and may therefore enhance prosodic learning in infancy. Overall, the study suggests that one reason behind infant preference for IDS could be its higher information value at the prosodic level, as measured by the amount of surprisal in the F0 contours. This provides the first formal link between the properties of IDS and the models of attentional processing and statistical learning in the brain. However, this finding does not rule out the possibility that other differences between the IDS and ADS also play a role.

1. 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). Particular timbral characteristics have also been recently identified (Piazza, Iordan, & Lew-Williams, 2017). In addition to serving as language input tuned to the developmental stage of the listener (e.g., 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., Fernald et al., 1989; Garnica, 1977; Thiessen, Hill, & Saffran, 2005, see also Soderstrom, 2007, for an overview), thereby facilitating language learning from speech. In support of this, a systematically higher attentional preference and pre-linguistic and linguistic interactional engagement towards IDS speakers was found in a recent meta-analysis conducted by Spinelli, Fasolo, and Mesman (2017).

Although the exaggerated intonation of IDS is often implicitly assumed to be the cause for higher attentional attractiveness, according to

* Corresponding author.
E-mail addresses: okko.rasanen@aalto.fi (O. Räsänen), sofoklis.kakouros@aalto.fi (S. Kakouros), M_Soderstrom@umanitoba.ca (M. Soderstrom).

https://doi.org/10.1016/j.cognition.2018.05.015
Received 26 October 2017; Received in revised form 15 May 2018; Accepted 21 May 2018

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 (Cooper & Aslin, 1990; Fernald, 1985; 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., Fernald et al., 1989; Garnica, 1977; Soderstrom, 2007). However, the existing work has not explicated the concept of saliency or the nature of the link between saliency and attention in a manner that would enable concrete predictions of what counts as attention-capturing speech and why. In addition, it appears that larger magnitudes or variations in standard acoustic correlates of prosody do not necessarily map to higher attentional capture of the speech. For instance, Singh, Morgan, and Best (2002) found that IDS with higher and more variable pitch was not enough to capture infants' attention over emotional speech.

Since stimulus-driven attention and prominence of sensory input seem both to be connected to the unpredictability of the stimuli in the given context (see the next sub-section; see also Kidd, Piantadosi, & Aslin, 2012, 2014), it would be parsimonious to assume that the greater attention-capturing properties of IDS over ADS would be also based on probabilistic expectations. In other words, existing knowledge suggests that one reason why IDS might be more attractive to the listeners is simply because it exhibits different predictability properties than ADS at the level of prosodic features, given earlier exposure to speech in both speaking styles. 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. This is despite the fact that infants are known to be sensitive to statistical regularities in their perceptual experience (c.f., Saffran, 2003; Saffran, Aslin, & Newport, 1996; Soderstrom, Conwell, Feldman, & Morgan, 2009, and references therein) and to the prosodic structure of their native language already from an early age (e.g., Nazzi, Bertoncini, & Mehler, 1998).

In the present paper, a quantitative investigation is carried out in order to test whether IDS intonation is indeed not just more variable, but also less predictable than ADS, thereby being in line with the 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.

1.1. 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; Tsuchida & Cottrell, 2012; Zarcone, van Schijndel, Vogels, & Demberg, 2016; Zhang, Tong, Marks, Shan, & Cottrell, 2008). From this viewpoint, attentional processing can be seen as a mechanism that enables the perceptual system to focus on aspects of the environment with high information content (Shannon, 1948), that is, input that is not yet learned and thereby accurately predicted by the brain (cf., Clark, 2013; Friston, 2010; Helmholtz, 1860; Ranganath & Rainer, 2003). In addition, all behavioral paradigms for testing infant learning in experimental settings are based to a greater or lesser extent on some kind of preferential sampling of the environment based on the current knowledge of the infant. This inherently couples familiarity with external stimuli to the resulting attentional behavior. However, infants are also known to prefer visual and auditory stimuli that are surprising or novel only as long as the input is not too unlikely in the given context. This phenomenon is also known as the Goldilocks effect (Kidd et al., 2012, 2014) where the preferred degree of predictability is not too much or too little. This suggests that instead of just preferring the most unlikely inputs in each situation, the input stream should still be structured enough to support learning—a prerequisite for experience-based evaluation of the relative information value of different competing inputs.

Predictability-based accounts also have a notable role in several aspects of human language processing (e.g., Jurafsky, 1996; Jurafsky, Bell, Gregory, & Raymond, 2001; Watson, Arnold, & Tanenhaus, 2008). In the context of prosody, earlier work has demonstrated the role of prosodic prominence in the regulation of information rate in speech (e.g., Aylett & Turk, 2004) and, e.g., tight coupling of lexical and syntactic analysis and prosodic parsing of the speech stream (e.g., Buxó-Lugo & Watson, 2016; Steinhauer, Alter, & Friederici, 1999). In addition, recent work in adult speech 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, 2016) similarly to low-probability lexical items in a predictive context (e.g., Cole, Mo, & Hasegawa-Johnson, 2010). Recent preliminary findings also suggest that adult listeners are sensitive, and rapidly adapt, to changing statistical properties of the intonation patterns, and this leads to experience-based expectations for prosody (Kakouros, Salminen, & Räsänen, 2018). Violations to experimentally induced prosodic expectations also lead to increased subjective impression of prominence and seem to alter the semantic processing of the speech stream (Kakouros et al., 2018; see also, e.g., Magne et al., 2005). Overall, the earlier research indicates that auditory attention and perceptual prominence are connected to the predictability of the prosodic patterns, and this may also play a role 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 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 of various linguistic and non-linguistic regularities, since it has long been known that prosodic cues are perceptually accessible to them (e.g., Hirsh-Pasek et al., 1987; Kuhl & Miller, 1982), and since they also seem to be sensitive to distributional properties of non-native tonal patterns at least at the age of 11–12 months (Liu & Kager, 2014), it is reasonable to assume that infants are also sensitive to the general statistical structure of speech prosody.

If predictability of the stimulus is a major factor in controlling stimulus-driven attention in infants, as exemplified by the widely used preferential head-turn or looking-time paradigms to probe infants' learning, and if predictability also plays a role in infants' behavioral preferences for IDS, 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.

2. 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

[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 (cf., e.g., a simple sine wave, which is equally predictable regardless of amplitude or frequency).