Infant-Directed Speech Enhances Temporal Rhythmic Structure in the Envelope

Victoria Leong\textsuperscript{1}, Marina Kalashnikova\textsuperscript{2}, Denis Burnham\textsuperscript{2} & Usha Goswami\textsuperscript{1}

\textsuperscript{1} Centre for Neuroscience in Education, Department of Psychology, University of Cambridge
\textsuperscript{2} MARCS Institute, University of Western Sydney

Abstract

Infant-directed speech (IDS) supports language learning via mechanisms that are still not well-understood. Here, we adopt a 'temporal sampling' perspective to investigate whether rhythmic enhancements in the temporal structure of IDS could support multi-timescale neuronal oscillatory sampling of the speech signal by the infant brain. We compare natural maternal speech directed to infants at the ages of 7-, 9-, 11- and 19-months, to adult-directed speech (ADS). Speech temporal structure is analysed using a novel multi-timescale Spectral-Amplitude Modulation Phase Hierarchy (S-AMPH) model, which extracts the Stress-rate, Syllable-rate and Phoneme-rate modulations (i.e. temporal patterns). Compared to ADS, we find that IDS shows a "stress-shifted" temporal profile. Stress-rate modulations dominate the modulation spectrum of IDS, whereas Syllable-rate modulations are dominant in ADS. Further, multi-timescale phase-synchronisation measures indicate that in IDS, Syllable-rate modulations are more synchronised to Stress-rate modulations and less synchronised to Phoneme-rate modulations. Thus, when speaking to infants, mothers pattern their syllables more regularly with prosodic stress, while allowing the phonemes within uttered syllables to vary more in timing. Accordingly, we conclude that the temporal structure of (Australian English) IDS is primarily stress-dominant, which could 'tune' the infant brain toward stress-based speech segmentation - an adaptive strategy for bootstrapping early language learning.

Index Terms: infant-directed speech, rhythm, temporal structure, amplitude envelope

1. Introduction

When speaking to infants, adults spontaneously adopt a prosodically-exaggerated speaking style called 'motherese' or infant-directed speech (IDS) [1]. The exaggerated pitch in IDS captures infants' attention and communicates affective warmth [1-3]. The acoustic adaptations in IDS are also thought to support infants' language learning directly. During the first year of life, infants develop knowledge of the phonology of their native language (the sounds in speech, e.g. phonemes, syllables, stress patterns) [4-6]. Well-formed phonology is essential for language development. Poor phonology is causally implicated in dyslexia [7] and characterizes some infants with acoustic landmarks to locate phrase and word boundaries, thereby 'bootstrapping' early language acquisition [12-13]. Accordingly, enhancements of rhythmic structure in IDS should be present and adaptive for infants' phonological development. Here, we investigate the acoustic modulation statistics of rhythm enhancement in IDS. Speech is richly patterned, with amplitude modulations at multiple timescales over a range of acoustic frequencies. The S-AMPH (Spectral Amplitude Modulation Phase Hierarchy) model is a low-dimensional representation of the statistically-dominant spectro-temporal modulation patterns in speech [14], and is particularly suited to the multi-timescale analysis of speech rhythm [15]. The model represents the amplitude envelope as an AM hierarchy, capturing the dominant temporal patterning of English at 3 phonological levels: prosodic stress (Stress AM), syllables (Syllable AM) and phonemes (Phoneme AM) respectively. Here, the S-AMPH model is used to identify differences between IDS and ADS at these 3 different phonological levels.

Our exploration of the temporal structure in IDS is also motivated by neural proposals concerning speech encoding. It is thought that endogenous oscillatory activity in the auditory cortex becomes rhythmically-entrained to speech modulation patterns [16], allowing the brain to perform simultaneous temporal sampling (parsing) of the speech signal at different timescales. For example, 'theta' (4-7 Hz) and 'gamma' (25-40 Hz) oscillations are implicated in the extraction of syllables and phonemes respectively. It follows that enhancement of speech modulation structure (as predicted for IDS) should help the infant brain to entrain more effectively to speech, facilitating the parsing of phonological structure from the acoustic signal. Further, auditory cortical oscillatory activity shows hierarchical phase-nesting or synchronisation across multiple timescales [17-18]. This phase synchronisation across different oscillatory rates is thought to stabilise auditory sensory representations [18], and facilitate multi-timescale integration of acoustic information [17,19]. Accordingly, in IDS, we expect the rhythmic speech modulations that drive infant neural oscillatory entrainment to be enhanced relative to ADS both in terms of (a) power and (b) multi-timescale phase-synchronisation. Further, it is of interest to examine the phonological timescale(s) at which rhythmic enhancement occurs, as the different timescales may not be equally bounded. It is also important to assess whether the rhythmic properties of IDS change as infants develop. Here, a longitudinal design is used to assess power and phase-synchronisation indices of temporal structure, utilising IDS collected over the first 1.5 years of life.

2. Method

2.1 Participants

Nine mother-infant dyads (5F and 4M infants) participated in the study. The sample was a subset of a larger cohort of infants participating in an on-going longitudinal study investigating word segmentation. Prosody provides naive listeners like infants with acoustic landmarks to locate phrase and word boundaries, thereby 'bootstrapping' early language acquisition [12-13]. Accordingly, enhancements of rhythmic structure in IDS should be present and adaptive for infants' phonological development. Here, we investigate the acoustic modulation statistics of rhythm enhancement in IDS. Speech is richly patterned, with amplitude modulations at multiple timescales over a range of acoustic frequencies. The S-AMPH (Spectral Amplitude Modulation Phase Hierarchy) model is a low-dimensional representation of the statistically-dominant spectro-temporal modulation patterns in speech [14], and is particularly suited to the multi-timescale analysis of speech rhythm [15]. The model represents the amplitude envelope as an AM hierarchy, capturing the dominant temporal patterning of English at 3 phonological levels: prosodic stress (Stress AM), syllables (Syllable AM) and phonemes (Phoneme AM) respectively. Here, the S-AMPH model is used to identify differences between IDS and ADS at these 3 different phonological levels.

Our exploration of the temporal structure in IDS is also motivated by neural proposals concerning speech encoding. It is thought that endogenous oscillatory activity in the auditory cortex becomes rhythmically-entrained to speech modulation patterns [16], allowing the brain to perform simultaneous temporal sampling (parsing) of the speech signal at different timescales. For example, 'theta' (4-7 Hz) and 'gamma' (25-40 Hz) oscillations are implicated in the extraction of syllables and phonemes respectively. It follows that enhancement of speech modulation structure (as predicted for IDS) should help the infant brain to entrain more effectively to speech, facilitating the parsing of phonological structure from the acoustic signal. Further, auditory cortical oscillatory activity shows hierarchical phase-nesting or synchronisation across multiple timescales [17-18]. This phase synchronisation across different oscillatory rates is thought to stabilise auditory sensory representations [18], and facilitate multi-timescale integration of acoustic information [17,19]. Accordingly, in IDS, we expect the rhythmic speech modulations that drive infant neural oscillatory entrainment to be enhanced relative to ADS both in terms of (a) power and (b) multi-timescale phase-synchronisation. Further, it is of interest to examine the phonological timescale(s) at which rhythmic enhancement occurs, as the different timescales may not be equally bounded. It is also important to assess whether the rhythmic properties of IDS change as infants develop. Here, a longitudinal design is used to assess power and phase-synchronisation indices of temporal structure, utilising IDS collected over the first 1.5 years of life.

2. Method

2.1 Participants

Nine mother-infant dyads (5F and 4M infants) participated in the study. The sample was a subset of a larger cohort of infants participating in an on-going longitudinal study investigating
the acoustic determinants of language acquisition. Selected families had no history of language or cognitive disorders. All mothers spoke Australian English as their native language and obtained average non-verbal IQ scores for block design, matrix reasoning and digit span subscales of the Wechsler Adult Intelligence Scale IV [20] (mean = 12.22, SD = 1.25).

2.2 Design

Each mother-infant dyad was assessed longitudinally between the (infant) ages of 7 months and 19 months. IDS samples were collected at 5 time-points, when the infants were 7 (mean age = 31 weeks, SD = 1.5), 9 (mean age = 40 weeks, SD = 1.7), 11 (mean age = 48 weeks, SD = 0.9), 15 (mean age = 66 weeks, SD = 2.3), and 19 (mean age = 79 weeks, SD = 11.1) months old respectively. Unfortunately there was excessive background noise in the samples collected at 15 months, hence this time-point was excluded from the current analysis. Additionally, each mother contributed samples of ADS when their infant was approximately 12 months old (mean age: 54.11 weeks, SD = 12.6 weeks).

2.3 Experimental Set-up

For the IDS recordings, mother and infant interacted alone in a laboratory room. The mother sat facing the infant who was sitting in a high chair. A video camera was mounted in each corner of the room to allow for monitoring and video recording of the session. The mother wore a head mounted microphone (AudioTechnica AT892), connected to Adobe Audition CS6 software via an audio input/output device (MOTU Ultralite MK3). The speech samples were digitally recorded at 16 kHz (or digitally recorded at 44.1kHz and re-sampled to 16 kHz). The experimenter monitored the video and audio for each session from an adjoining room. For the ADS recordings, each mother interacted with a female experimenter in the laboratory room using the same recording apparatus, and the infant was not present. The IDS and ADS recording sessions lasted between 5 to 10 minutes each.

2.4 Task

For the IDS recordings, the mothers were told that the purpose of the experiment was to capture their interaction with their infant during a brief play session. The experimenter instructed mothers to use three target words containing the corner vowels /i/, /u/ and /a/ when interacting with their infants: "sheep", "shoe", and "shark" respectively. This ensured that the mother-infant interactions focused on similar semantic content, and that the recordings contained similar vocalic exemplars. To facilitate the interaction, three toys (sheep, shoe, and shark) and pictures depicting the target objects were provided. Mothers were asked to talk to their infant for as long as they felt appropriate. All mothers and infants were observed to be motivated and engaged in the task. For the ADS recordings, mothers were interviewed about the experimental play sessions. The experimenter asked questions that would elicit responses containing the target words, such as, "Do you remember what the toys were?"

2.5 Speech Segments Used for Analysis

The raw recordings of IDS and ADS were manually divided into shorter segments for analysis using Praat software [21]. Each segment contained a complete phrase(s) from the original utterance. Portions of speech that were interrupted by

the infant or adult addressee, or that contained excessive background noise were not used. In total across the 5 speaking conditions (4 IDS, 1 ADS), the 9 speakers contributed 301 speech segments that were used in this analysis. Table 1 provides a breakdown of the number of segments in each speaking condition. On average, these segments were 10.13s in length (range: 5.48s-17.08s), and each speaker contributed an average of 6-7 segments in each condition.

<table>
<thead>
<tr>
<th>Speaking condition</th>
<th>IDS</th>
<th>ADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of segments analysed</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>Average segments per speaker</td>
<td>7.44</td>
<td>7.11</td>
</tr>
<tr>
<td>6.78</td>
<td>6.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Summary of IDS and ADS speech segments used

2.6 Analysis Methods

2.6.1 Signal Processing: The S-AMPH Representation

The S-AMPH representation of the amplitude envelope for each speech segment was extracted using a 2-stage filtering process, as shown in Figure 1. First, the raw acoustic signal was band-pass filtered into 5 frequency bands using a series of adjacent finite impulse response (FIR) filters. The 5 frequency bands were: (1) 100-300 Hz; (2) 300-700 Hz; (3) 700-1750 Hz; (4) 1750-3900 Hz; and (5) 3900-7250 Hz. The Hilbert envelope was extracted from each band-filtered signal. These 5 Hilbert envelopes were then down-sampled to 1050 Hz and passed through a second series of band-pass filters in order to isolate the 3 different AM rates within the envelope modulation spectrum. These 3 AM rates correspond to the typical durations of major phonological units and were the 'Stress' rate (prosodic feet, 0.9-2.5 Hz), 'Syllable' rate (syllables, 2.5-12 Hz) and 'Phoneme' rate (phonemes, 12-40 Hz). For details of the derivation of bands for spectral and modulation rate filtering, please see [14]. The result of this two-step filtering process was a 5 (frequency) x 3 (rate) spectro-temporal representation of the speech envelope, comprising 'Stress', 'Syllable', and 'Phoneme' AMs for each of the 5 acoustic frequency bands.

2.6.2 Multi-Timescale Synchronisation Measure: Phase Synchronisation Index (PSI)

To measure the degree of temporal synchronisation (phase-locking) between pairs of speech AMs at different timescales (i.e. Stress:Syllable and Syllable:Phoneme), the n:m phase synchronisation index (PSI) was computed. The n:m PSI was
originally conceptualised to quantify phase-synchronisation between two oscillators of different frequencies (e.g. muscle activity) [22]. This measure was subsequently adapted for use in neural analyses of oscillatory phase-locking (e.g. [23]), and we apply this adaptation to speech AMs here. The PSI was computed as:

\[ \text{PSI} = \frac{1}{n} \sum_{i=1}^{n} |\theta_i - \theta_2| \]

In Equation (1), \( n \) and \( m \) are integers describing the frequency relationship between the two AMs being compared. Following a previous study [15], for the Stress:Syllable AM comparison, an \( n:m \) ratio of 2:1 was used, while for the Syllable:Phoneme AM comparison, an \( n:m \) ratio of 3:1 was used. 01 and 02 refer to the instantaneous phase of the two AMs at each point in time. Therefore, \((\theta_1 - \theta_2)\) is the generalised phase difference between the two AMs, which was computed by taking the circular distance (modulus 2\( \pi \)) between the two instantaneous phase angles. The angled brackets denote averaging of this phase difference over all time-points. The PSI is the absolute value of this average, and can take values between 0 (no synchronisation) and 1 (perfect synchronisation), as illustrated in Figure 2.

![Hypothetical AM pairs that yield PSI scores of 1 (left) and 0.66 (right) respectively.](image)

**2.6.3 Methodological Control: Speech-Shaped Noise (SSN) Surrogates**

As a control, speech-shaped noise (SSN) surrogates were created for each speech segment in all 5 speaking conditions. Each SSN surrogate had the same frequency spectrum as the original speech segment, but comprised white noise modulated with a random amplitude envelope rather than speech. The S-AMPH hierarchy was extracted and PSI values for these surrogates were computed using identical signal processing steps to the actual speech samples. Accordingly, the PSI values computed using these surrogates indicate the amount of 'baseline' synchronisation that would be present, by chance, in the amplitude envelope of speech-shaped noise which does not contain deliberate rhythmic patterning. These SSN surrogates also control for 'rhythm artefacts' that might be introduced as a result of the signal processing procedure (e.g. 'filter ringing').

**3. Results**

**3.1 Power: Envelope Modulation Spectrum**

Figure 3 plots the envelope modulation spectrum averaged across frequency bands and speakers. The x-axis shows the modulation rate, the y-axis shows the power in dB (normalised with respect to the mean power of each frequency band in each sample). Vertical dotted lines indicate the peak modulation rate (i.e. highest power) for each condition.

![Grand mean modulation spectrum averaged across frequency bands and speakers.](image)

**3.2 Multi-Timescale Synchronisation: PSI Scores**

Figure 4 shows the average PSI scores obtained with regard to Stress-Syllable AM synchronisation (top left subplot), and Syllable-Phoneme AM synchronisation (top right subplot) for each age group and acoustic frequency respectively. The equivalent PSI scores obtained for the SSN surrogates in each condition are shown in the bottom panels of Figure 4. From visual inspection, it is apparent that all 4 IDS timepoints showed markedly greater synchronisation than ADS at low acoustic frequencies for the Stress-Syllable AM pair (top left subplot). By contrast, all 4 IDS timepoints showed markedly reduced synchronisation compared to ADS at low acoustic frequencies for the Syllable-Phoneme AM pair (top right subplot). Thus, in IDS, there appears to be a selective shift toward rhythmic synchronisation at slow temporal rates (Stress-Syllable), which occurs at the expense of less synchronisation at faster (Syllable-Phoneme) temporal rates.

To assess whether the IDS and ADS PSI scores were statistically different, two separate repeated measures ANOVAs were conducted, taking PSI scores for the Stress-Syllable AM pair and the Syllable-Phoneme AM pair as dependent variables respectively. In both ANOVAs, Age [5 levels] and Frequency [5 levels] were entered as the within-subjects variables. The results of the ANOVAs confirmed our initial conclusions. For the Stress-Syllable AM pair, there was a significant main effect of Frequency (F(4,32) = 11.0, p<.0001), but no main effect of Age (F(4,32) = 2.05, p=.11). Importantly, however, the predicted interaction between Age and Frequency was significant (F(16,128) = 5.67, p<.0001). This confirmed that the predicted interaction between Age and Frequency was significant (F(16,128) = 5.67, p<.0001). This confirmed that there was indeed a different pattern of PSI scores for IDS versus ADS. Tukey HSD post-hoc analysis of this interaction revealed that for the two lowest frequency bands (100-300 Hz and 300-700 Hz), all 4 IDS timepoints (1-9, 11 & 19 months) showed significantly higher PSI scores than corresponding ADS (p<.05[*] for all comparisons). Thus, IDS was characterised by a significant boost in low-frequency
rhythmic synchronisation between Stress-rate and Syllable-rate AMs in the utterances. For the Syllable-Phoneme AM pair, there was a significant main effect of Age (F(4,32)=13.6, p<.0001), as well as of Frequency (F(4,32)=68.7, p<.0001). Importantly, the interaction between Age and Frequency was again significant (F(16,128)=7.38, p<.0001). Tukey HSD post-hoc analysis of this interaction revealed that for the two lowest frequency bands, all 4 IDS pointsw showed significantly lower PSI scores than corresponding ADS (p<.0001 [**] for all comparisons). Thus, in IDS, there is a boost in low-frequency Stress-Syllable synchronisation and a concurrent drop in low-frequency Syllable-Phoneme synchronisation. Finally, identical ANOVA analyses conducted on the SSN surrogate PSI scores (which yielded values of -0.05) revealed no significant main effects, or interaction effects, confirming that our findings did not arise from methodological artifacts.

Figure 4. PSI values computed for the Stress-Syllable AM pair (left column) and the Syllable-Phoneme AM pair (right column), for IDS and ADS conditions (top panel) and SSN surrogates (bottom panel) respectively. The x-axis indicates the frequency (spectral) band from which the AMs originate. The y-axis shows the PSI value. Error bars indicate the standard error of the mean. [*p<.05, ***p<.0001].

4. Discussion & Conclusions

We predicted that the temporal structure (rhythm) of IDS would be enhanced relative to ADS, conferring a phonological benefit for the infant listener. Accordingly, we assessed the rhythmic properties of IDS by measuring the (a) power and (b) degree of phase-synchronisation present within a multi-timescale (S-AMPH) representation of the speech amplitude envelope. We also explored whether certain speech modulation timescale(s) were selectively enhanced in IDS, and whether these acoustic enhancements changed as infants developed. As predicted, enhancements in the temporal structure of IDS were found. Both power and phase-synchronisation, measures revealed a 'stress-shift' in the temporal structure of IDS as compared to ADS. With regard to power, the modulation spectrum of IDS directed to the youngest infants peaked at stress-related modulation rates, rather than at syllable-related modulation rates (which is the case for ADS). Thus, the dominant temporal rhythmic patterning of IDS occurs at a larger phonological grain-size (prosodic stress feet) than for ADS (syllables). Further, the peak modulation rate of IDS systematically shifted upward with the increasing age of the infant, changing from the stress rate (~2.5 Hz) to the syllable rate (~4 Hz) over the first 1.5 years. With regard to phase synchronisation, Syllable-rate modulations in all IDS conditions were more synchronised to Stress-rate modulations and less synchronised to Phoneme-rate modulations. We interpret this as indicating that syllables in IDS are emphasised more regularly with prosodic stress but that their phonemic constituents are more variable in timing. For example, in IDS, stress may be placed regularly every 2-3 syllables rather than stochastically over a stretch of 5-10 syllables. Conversely, individual phonemes within uttered IDS syllables may be more variable in timing, perhaps due to the greater 'stretching' of vowels as compared to consonants. Accordingly, we conclude that IDS is characterised by a stress-dominant temporal rhythmic structure.

Since the AM patterns in speech entrain cortical oscillatory activity at equivalent timescales [16], the 'stress-shift' observed here should elicit stronger 'delta' (1-3 Hz) entrained oscillatory activity. This would bias infants' neural processing toward the recovery of larger phonological units that are cued by prosodic stress (e.g. 'Strong-weak' trochaic words like "MU-mmy"), serving to 'boot-up' early language systems. Indeed, this mechanism is consistent with experimental evidence that during the first year, English-learning infants initially use a stress-based strategy for segmenting words from continuous speech [6]. For languages with other rhythm typologies, the temporal structure of IDS could differ from ADS at the most rhythmically-important timescale for that particular language (e.g. Syllable-rate for Spanish, Phoneme-rate for Japanese). As a result, infants hearing 'syllable-enhanced' or 'mora-enhanced' IDS would be biased to parse speech at the most appropriate phonological grain size for their native language. Indeed, our results also have implications for the debate surrounding the acoustic basis of language rhythm typologies [25]. Our findings suggest that listeners' intuitions regarding English 'stress-timing' (i.e. perceived rhythmic regularity between stressed syllables) may originate in early life as a result of exposure to IDS that does have high stress-syllable phase-synchronicity. Phase-synchronicity is distinct from durational isochrony because two oscillators that are phase-locked (synchronised) will continue to produce high PSIs even when both oscillators are stretched or compressed in duration (e.g. as speakers slow down or speed up). When the phase-relationship between Stress- and Syllable-rate modulations is regular, the rhythmic patterning of stressed and unstressed syllables is also regular, as in poems and rhymes [14], conferring high transitional predictability to IDS. During infancy, perceptual systems may become hard-wired to extract and expect regularities in speech at these stress-syllable timescales. Consequently, during adulthood, prosodic stress may still be perceived as predictable even though the ADS acoustic input is now less regular. In summary, the stress-dominant acoustic temporal structure of (English) IDS could tune the infant brain toward stress-based speech perception. This would be adaptive for early language acquisition, and may persist into adult life.

5. Acknowledgements

We thank Maria Christou-Ergos for preparation of data for analysis. This work was supported by a Lucy Cavendish College Junior Research Fellowship to VL, and Australian Research Council funding to DB and UG.
6. References


