1st International Conference on Tone and Intonation (TAI)
6-9 December 2021, Sonderborg, Denmark

The role of accommodation in expressing emotions to newborn babies

Katalin Mády, Uwe D. Reichel, Anna Kohári, Ádám Szalontai

Hungarian Research Centre for Linguistics, Budapest
{mady.katalin|uwe.reichel|kohari.anna|szalontai.adam}@nytud.hu

Abstract

Infant-directed speech (IDS) carries signals expressing strong positive emotions towards the baby. This has been shown in various studies for e.g. higher f0 and energy, lower speech rate etc. Longitudinal studies comparing adult-directed (AD) and ID speech revealed that the difference between acoustic features usually connected with strong and positive emotions were not present directly from birth in mothers' speech, but emerged only at a later time point. In this study, emotion expression in terms of arousal and valence was investigated through labellers' perception in two groups: primipara mothers giving birth to their first child, and multipara mothers who were already experienced in IDS with their older children. A set of fixed sentences taken from semi-spontaneous speech was scored by labellers. Both arousal and valence scores were higher in IDS. Arousal scores were further apart in the multipara group, while this distinction was not present for valence. A set of acoustic parameters were investigated according to low and high perceived arousal and valence in the same data set.

Keywords: infant-directed speech, emotion, arousal, valence, acoustics

1. Introduction

An important function of infant-directed speech (IDS) is to express positive emotions towards the baby, manifesting in prosodic features such as lower speech rate, higher f0, larger pitch range etc. These factors are subject to changes with the cognitive development of the baby (see [1] for a detailed overview). Parameters reflecting emotions in IDS were investigated in a longitudinal study ([2]) for Hungarian-speaking mothers in comparison to adult-directed speech (ADS). Recordings were carried out at 0, 4, 8 and 18 months of the baby, including a comparison between ADS and IDS at all timepoints. We were expecting a decrease in AD-ID differences towards 18 months of the baby assuming that the enhancement of language-specific differences overrides the role of emotion expression. However, no such tendency was found between 4 and 18 months. An unexpected finding was, however, that AD-ID differences were far less pronounced at the first recording that took place directly after the birth of the baby. One possible explanation was the fact that recordings took place in different environments for organisational reasons: mothers of newborn babies were accessible at a birth centre, while the subsequent recordings were carried out at a baby lab. Thus the question arised whether the shift towards more usage of IDS was due to external or emotional factors such as hospital environment, stress or fatigue after birth, or to linguistic factors such as missing experience in the communication with the baby, thus, accommodation accommodation. According to [3], this term is defined as the adjustments made in speech and communication (perceived or actual) resulting from attempts to take positive (or negative) account of an interlocutors behavior, group memberships, motivation, and needs, in an interpersonal (face-to-face or mediated) encounter, and their impact on the relationship and future encounters with the interlocutor or other members of his or her group.

In order to dissociate the possible influence of accommodation vs. environmental or language-external factors, we conducted a study in which mothers from the longitudinal data set, who gave birth to their first child (primipara group, PP), were compared to mothers of newborn who already had at least one child (multipara group, MP). We hypothesised that in case they use more IDS with their 0 month-old baby in the hospital, thus, identical environment than PP mothers, the role of accommodation (its presence in MP and absence in PP) has a stronger explanatory power on the longitudinal differences than environmental factors.

A previous study has shown differences between the utilisation of IDS-specific prosodic cues between multipara and the primipara mothers ([4]). With respect to f0 mean, findings could not be linked to emotion expression directly: both mother groups used higher f0 in IDS, but semitone values in the multipara group were higher in general, both in AD and ID speech. A different tendency was observed for energy level which was highest in multipara IDS compared to their ADS on the one hand and to primipara speech in both registers, on the other. Thus, multipara mothers differenciated their energy level according to the participant to whom they directed their speech, but not primipara mothers.

IDS is said to express stronger positive emotions towards the baby. However, there is a general difficulty when examining emotions via acoustic analysis. Single prosodic parameters like f0 or energy often cannot be directly related to specific emotions. For example, higher f0 can characterise both positive and negative emotion categories such as happiness and fear. Also, slow speech rate can express boredom or comfort. Thus, prosodic parameters need to be interpreted in a more complex manner. In the first part of this study the emotional load of IDS was investigated perceptualy, by using a two-dimensional arousal-valence space [5]. Subsequently, emotion scores for arousal and valence were analysed along with acoustic parameters usually connected to emotion expresson along these dimensions.

2. Percieved emotion in IDS and ADS

2.1. Methods and materials

22 primipara and 23 multipara mothers, all being native speakers of Hungarian, were recorded at the Birth Centre of the Military Hospital in Budapest one or two days after giving birth to their baby. Mothers of twins were not included in the study in order to have a uniform sample. No other infant than the mother's baby was present at the recording. The experimenter was a female research assistent in her 50'es. All mothers signed an informed consent form for the use of their recordings for sci-

10.21437/TAI.2021-51

entific purposes.

Speakers were asked to tell a fairytale first to the experimenter (ADS), then to their own baby (IDS) after familiarising with the story. The plot was based on human-like pixies playing hide-and-seek. They received a booklet with paintings some of which contained written utterances. Mothers were asked to tell the story in their own words but to incorporate the fixed utterances into their fairytale. This hybrid form was chosen in order to achieve a next-to spontaneous style but to dispose of a set of identical sentences allowing for a comparison between registers (ADS vs. IDS) and across speakers. An example is given in Figure 1. Mothers told the story first to the experimenter (ADS), then to their own infant (IDS) in each session. Speech was recorded to a digital recorder (Zoom H4n) and a supercardioid head-mounted microphon (Beyerdynamic TG H74c).

Figure 1: *Painting from the fairytale used in the experiment. Words:* "At last, we have found you, Szut! We thought you will never turn up again!"

Emotions were judged in a two-dimensional system. Arousal refers to the intensity of the affective state, valence to the level of pleasure [5, 6]. Four labellers (female students in their 20's without their own child) judged utterances for their arousal and valence on a scale between –4 and +4, using the MFC module of the Praat software. 0 was defined as neutral for both dimensions. The two endpoints of the scale for arousal were *unmoved* and *very strong* emotions, for valence *strongly negative* and *strongly positive*. Labellers could listen to sentences repeatedly before giving their judgement score, but they could not move back to previous utterances. Participants learned prior to the experiment that they were to listen to sentences taken from a fairytale. No other information was given, thus, they did not know that half of the sentences was uttered in infant-directed speech register. All participants went through a practice block first that contained 20 utterances from mothers different from those under investigation talking to an older baby and a different experimenter. The labellers listened to the sentences in four blocks, randomised within, bot not between blocks. All six utterances produced in the AD and ID register of the same mother were presented in the same block to allow for a direct comparability within one session. Primi- and multipara mothers were distributed between the blocks. Thus, within each block, labellers were presented with both registers and both parities (PP and MP), and the order of the blocks was varied between labellers. The data was structured as follows: 45 mothers $\times$ 2 parity groups (primi- vs. multipara) $\times$ 2 registers $\times$ 6 sentences $\times$ 4 labellers. Sentences that diverged from the target sentence, e.g. by added or missing words, were excluded from the analysis. 1908 utterances were analysed in total.

Krippendorff's alpha for inter-rater agreement was rather low: 0.4074 for arousal and 0.243 for valence. This was mainly due to the lack of reference scores for labellers: there is no obvious 0 point for arousal and valence to which they could anchor their judgements. On the other hand, the tendencies between their scores were parallel according to visual inspection on plots. As a consequence, we decided to use intercept-only linear mixed-effects models. Register (IDS vs. ADS) and parity (PP vs. MP) were included as fixed effects in the models, speaker, utterance and labeller as random effects. Analyses were run in R using the `lme4` [7] and `car` [8] packages.

2.2. Results

Mean arousal scores differred both along registers and parity: MP AD: 0.37, MP ID: 0.92, PP AD: 0.34, PP ID: 0.53. Thus means expressed emotions scored right of zero meaning neutral. Raw scores before calculating the mean show the same tendency: ID is connected with a higher amount of positive values, and the difference is larger for the multipara group (see Figure 2, top). This is also reflected by the model output: arousal is significantly higher in IDS in both the PP and MP group ($t = 7.50, p < 0.001$), not for parity ($t = 0.023, p = 0.47$), but the interaction of the two ($t = -2.94, p = 0.003$).

Effects on valence scores, on the other hand, were much less pronounced between the two registers and the two speaker groups. All mean scores were close to zero: MP AD: -0.17, MP ID: 0.06, PP AD: -0.05, PP ID: 0.06. Despite of the small mean differences between registers scoring around 0.1 valence mean scores, register received significantly higher scores according to the linear mixed-effects models ($t = 2.54, p = 0.004$), but not parity or its interaction with register.

3. Acoustics of valence and arousal

Findings on the acoustics of valence and arousal are summarized in [9] (for further details please see the references therein): high as opposed to low arousal is characterized amongst others by higher speech rate, higher intensity mean and variability, higher f0 mean and variability, higher spectral balance indicating increased vocal effort, and a higher first formant (F1). The latter is likely due to an increased mouth opening in high arousal.

As found e.g. by [10, 11, 12] the relation between acoustics and valence is less straightforward than for arousal. For example, higher f0 characterises both elation joy (positive) and fear (negative), while comfort (positive) and boredom (negative) are both reflected by lower f0 [11]. As summarized in [9], positive valence is amongst others characterised by higher speech rate, by lower intensity and a lower intensity variability. In addition, valence is positively correlated with the second formant (F2) due to more lip spreading caused by smiling [13] which leads to a relatively more fronted articulation.

Based on these findings we expected for high as opposed to low arousal:

- an increase in energy mean and variation,
- an increase in f0 mean and variation,
- increased speaking rate,
- higher F1 values, and
- a higher spectral balance

For positive as opposed to negative valence, we expected

- lower energy mean and variation,

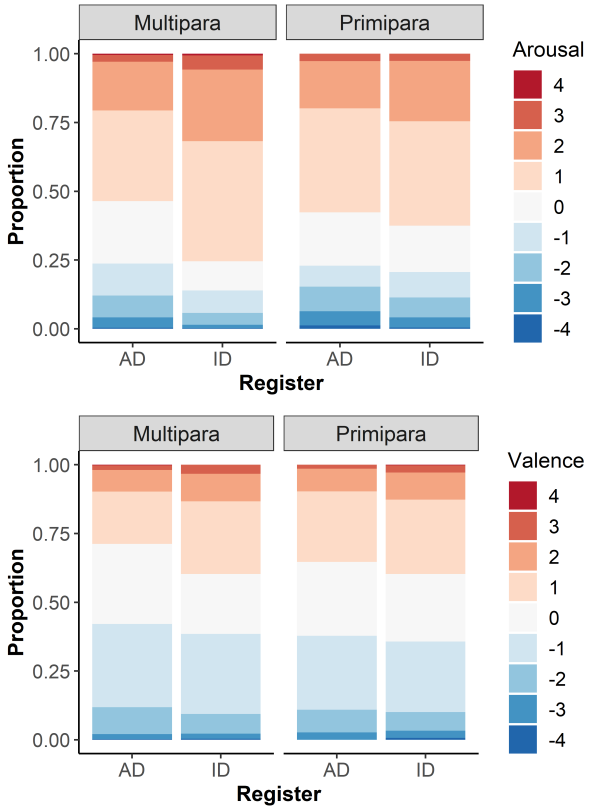

Figure 2: *Scores for arousal (top) and valence (bottom).*

- higher f0 variation,
- changes of f0 mean with any polarity,
- increased speaking rate, and
- higher F2 values.

3.1. Feature extraction

All features were extracted using the Praat 5.3 [14] and CoPaSul 1.0.1 [15] software.

Energy The energy contour is simply represented in terms of the root mean squared deviation within Hamming windows of length 50 ms and window shift 10 ms. From this contour we calculate the median and the interquartile range rather than the mean and standard deviation. These robust non-parametric distribution parameters were chosen in order not to rely on the assumption that energy follows an unimodal Gaussian distribution.

Spectral balance We calculate spectral balance in terms of SPLH–SPL, i.e. the signal's sound pressure level subtracted from the level after pre-emphasis. Pre-emphasis was carried out in the time domain as follows: $s'[i] = s[i] - \alpha \cdot s[i-1]$, where $s[i]$ and $s'[i]$ refer to the original amplitude and the amplitude after pre-emphasis at time index i. α determines the lower frequency boundary for pre-emphasis by 6dB per octave, and it was set to 0.97. Higher spectral balance values indicate more energy in the high-frequency bands which in turn indicates increased vocal effort.

F0 Fundamental frequency (f0) was extracted with Praat by autocorrelation (window shift 10 ms). Voiceless utterance parts and f0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a speaker-dependent base value b. b was defined as the median below the 5th percentile of all f0 values of the speaker's utterances. F0 was finally transformed to semitones (ST) relative to this base value as follows $F0_{st} = 12 \cdot log_2(\frac{F0_{Hz}}{b})$. Similarly to energy, from the f0 contour we calculate the median and interquartile range.

Formants We extracted formants with Praat from the LPC coefficients using the Burg algorithm and applying joint Viterbi smoothing. Window length was set to 25 ms, window shift to 2 ms, the formant ceiling frequency to 5500 Hz, and the lower frequency for pre-emphasis to 50 Hz. From the F1 and F2 contours we measured the arithmetic means of F1 and F2 values within 30 ms windows centered on the vowel segment midpoints in Bark. Vowel segments were identified by automatic signal-text alignment on the phoneme level using WebMAUS [16].

Duration We measure the utterance durations in seconds. Since all speakers produce the same sentences, duration can be seen as a proxy for (inverse) speaking rate.

Statistical analyses In order to compare the acoustic feature values across low and high valence and arousal, the four labellers' scores were averaged for each sentence. These mean scores were divided from the 9-level scale into 2 classes, < 0 classified as *low* and ≥ 0 as *high*. Trends are presented together with the effect size in terms of Cohen's d. d values below 0.2 indicate a *very small* effect size, values between 0.2 and 0.5 indicate a *small* effect size, and values between 0.5 and 0.8 indicate a *medium* effect size.

3.2. Results

In Figures 3 and 4 we present the feature distributions across emotion levels for arousal, and valence, respectively. Furthermore, we give the effect sizes for all features in both dimensions in Table 1.

Table 1: *Cohen's d values and the trend of acoustic features for high vs low valence and high vs low arousal. A $>$ trend indicates that high values on the respective emotion dimension co-occur with high feature values. A $<$ trend indicates an inverse emotion-feature relationship. For very small effect sizes well below .2 we mark the absence of a trend as $-$.*

Dimension	Feature	Cohen's d	Trend
Arousal	energy median	.30	$>$
	energy IQR	.39	$>$
	f0 median	.15	$-$
	f0 IQR	.29	$>$
	F1 mean	.16	$-$
	spectral balance	.59	$>$
	duration	.49	$<$
Valence	energy median	.05	$-$
	energy IQR	.02	$-$
	f0 median	.09	$-$
	f0 IQR	.05	$-$
	F2 mean	.24	$>$
	duration	.19	$(<)$

In line with [10, 11, 12] the connection between the exam-

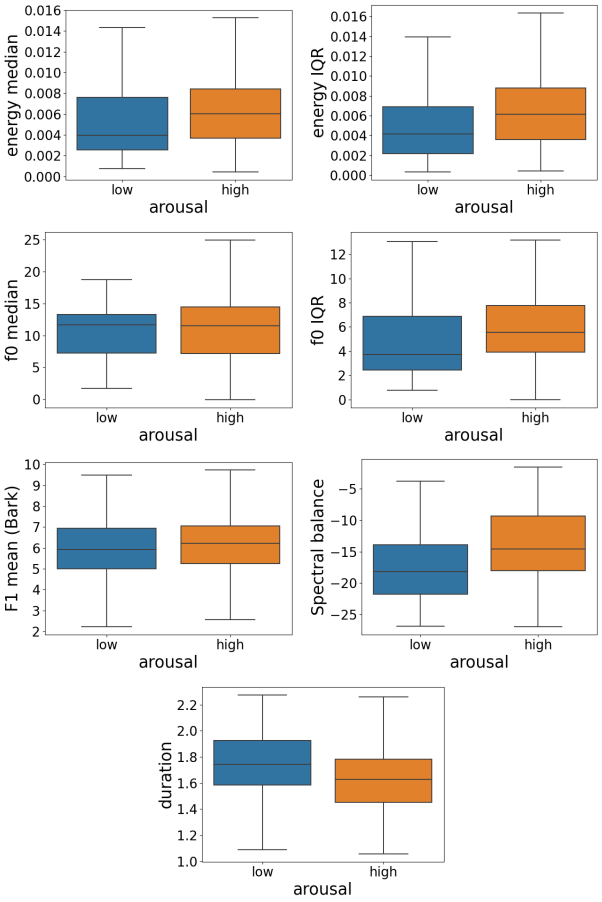

Figure 3: *Acoustics of perceived arousal binned into* low *and* high. *From top to bottom row: energy median and IQR, f0 median and IQR, F1 mean, spectral balance, and duration.*

ined acoustic features and perceived valence is much weaker than for arousal. Only two of the six features showed a (close to) small effect size: F2 increased and duration decreased with high valence, which both confirm the production studies summarized in [9]. Lower duration indicates a slightly increased speaking rate for high valence. A raised F2 might be caused by and perceived as smiling which is itself an expression of positive valence. Against our expectations we did not find more differences in f0 or energy variation, nor changes in median values for perceived high valence.

For arousal, five of the seven features showed an at least small effect size. As expected, energy level and variability as well as f0 variability increase with perceived high arousal. Duration got shorter, which as for valence indicates a slightly increased speaking rate. A higher spectral balance indicates increased vocal effort. Against our expectations the f0 median did not contribute to arousal perception, and neither did the mean of the first formant.

4. Dicussion

Results show that mothers express stronger emotions when talking to their infant already from birth. Multipara mothers who are already experienced in IDS communication with their other children adapt their speech behaviour to the newborn baby,

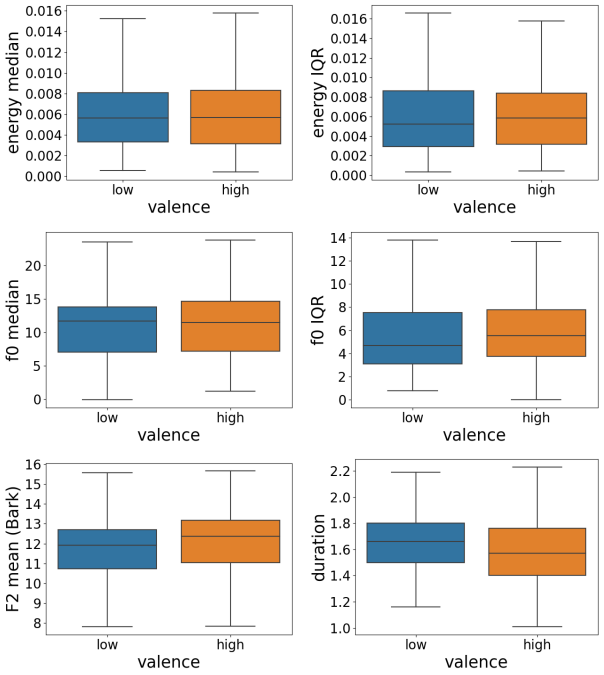

Figure 4: *Acoustics of perceived valence binned into* low *and* high. *From top to bottom row: energy median and IQR, f0 median and IQR, F2 mean, and duration.*

while primipara mothers still develop this skill with growing age of their baby. Thus it seems that environmental factors (stress, fatigue or the hospital room as a location) cannot explain the longitudinal results [2] alone. On the other hand, it is possible that the smaller amount of ID usage both in terms of perceived emotion and acoustic features connected to them might go back to accommodation in more general terms: mothers with their first baby might be less experienced with their role as a mother in general. It is possible that they were shy when communicating with their child, especially in the presence of an experimenter.

In perceptual rating experiments (e.g. [17]), valence was less well recognized by speech acoustics alone than arousal. This finding is generally confirmed by cross-modal comparisons of acoustic and visual features in automatic prediction of these emotion dimensions (e.g. [18]): while arousal is well manifested in the acoustic speech signal, valence is better predicted from facial expression and spoken content. This is reflected by the results between high vs. low arousal and valence for various acoustic parameters in the present study. The comparison showed that acoustic information does contribute to arousal perception with respect to various features to a much higher extent than to valence judgements. An important finding is that f0 height is not in a straightforward relationship with perceived strong and positive emotions, although this link is often mentioned in connection with infant-directed speech.

5. Acknowledgements

This study was funded by the National Research, Development and Innovation Office, grants K 115385, K 135038 and PD 134775.

6. References

[1] C. Saint-Georges, M. Chetouani, R. Cassel, F. Apicella, A. Mahdhaoui, F. Muratori, M.-C. Laznik, and D. Cohen, "Motherese in interaction: at the cross-road of emotion and cognition? (a systematic review)," *PLoS ONE*, vol. 8, no. 10, pp. 1–17, 2013.

[2] K. Mády, U. D. Reichel, A. Kohári, A. Deme, and A. Szalontai, "Primary functions in infant-directed speech and their longitudinal development," in *Proc. Beszédkutatás – Speech Research, Budapest*, Budapest, 2020, pp. 62–64.

[3] C. Gallois and H. Giles, *Communication Accommodation Theory*. John Wiley & Sons, 2015.

[4] K. Mády, U. D. Reichel, A. Szalontai, A. Kohári, and A. Deme, "Prosodic characteristics of infant-directed speech as a function of maternal parity," in *Proc. 9th Speech Prosody, Poznań*, Poznań, 2018, pp. 294–298.

[5] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2009, pp. 65–68.

[6] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," in *Proc. Interspeech*, Hyderabad, 2018.

[7] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[8] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[9] D. P. Szameitat, C. J. Darwin, D. Wildgruber, K. Alter, and A. J. Szameitat, "Acoustic correlates of emotional dimensions in laughter: arousal, dominance, and valence," *Cognition and emotion*, vol. 25, no. 4, pp. 599–611, 2011.

[10] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[11] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and emotion*, vol. 15, no. 2, pp. 123–148, 1991.

[12] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," *Handbook of emotions*, vol. 2, pp. 220–235, 2000.

[13] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception & Psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[14] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Tech. Rep., 1999, 132–182.

[15] U. D. Reichel, "Copasul manual–contour-based parametric and superpositional intonation stylization," *arXiv preprint arXiv:1612.04765*, 2016.

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

[17] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The measurement of meaning*. University of Illinois press, 1957, no. 47.

[18] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.