Emotion-related explanations of the vowel variability in infant-directed speech

Titia Benders

ARC Center for Excellence in Cognition and its Disorders
Department of Linguistics, Macquarie University
titia.benders@mq.edu.au

Abstract

Speech is inherently variable, and so is Infant-Directed Speech (IDS). IDS is also a highly emotional register. In two listener-rating studies of Dutch IDS, we explore emotion-related explanations for (Exp 1) and consequences of (Exp 2) the acoustic differences between IDS vowel tokens. Listeners rated IDS utterances on valence and energy (Exp 1) and on the perception of smiles and child-like speech (Exp 2). The predicted association between valence and formant frequencies was not found (Exp 1), but a higher second formant results in a more smiled and more child-like percept (Exp 2).

Index Terms: infant-directed speech, emotion, affect.

1. Introduction

Parents sound differently speaking to babies than to other adults. Easily perceived are the higher fundamental frequency (F0) and larger F0 range in infant-directed speech (IDS). Possibly less accessible to the naive ear are the acoustic changes to IDS vowels. The corner vowels (/i, u, a/) may be more distinct in IDS than in adult-directed speech (ADS), as measured in an acoustic space defined by the first and second formant (F1, F2) [1]. The corner vowels may also have overall higher formant frequencies in IDS than in ADS [2, 3, 4]. Lastly, the F1 and F2 of individual vowel tokens display larger variation around the mean in IDS than in ADS [1]. The present paper takes a first step towards affective explanations for (Experiment 1) and perceptual consequences of (Experiment 2) the variability in F1 and F2 between individual vowel tokens within IDS.

The acoustic characteristics of IDS can be explained in terms of the three functions that IDS serves: expressing affect, directing attention, and teaching language [5]. The affective function is associated with F0 height and the attentional function with F0 range [6]. Mothers may teach their baby language by enhancing the distances between the corner vowels in IDS and prepare them for between-speaker variability with high vowel variability within their IDS [1].

A different line of research attempts to explain all acoustic properties of IDS as the result of the generally stronger expression of emotion in this register [7]. As IDS primarily expresses a happy emotion, the F0 height and F0 range in IDS may both be consequences of the high energy in very happy speech [8].

IDS as highly emotional speech is difficult to reconcile with enhanced distances between the corner vowels. However, several recent studies have failed to replicate these enhanced distances and report a rise of the formant frequencies in IDS [2, 3, 4]. Raised formant frequencies in IDS could result from the expression of positive affect, as the (positive) facial expression of a smile raises formant frequencies [9]. The raised formants in IDS could also be due to mothers imitating the higher formants in child speech [10]. As adaptation to the interlocutor is a social process, also this latter explanation suggests that the raised formant frequencies in IDS are a consequence of the general expression of positive affect.

No connection has yet been drawn between the expression of emotion and formant variability in IDS. Such a connection would be in line with the idea that the expression of affect impacts on articulation, and thus on formant frequencies [8]. An emotion-related explanation of vowel variability could either replace the language-teaching account of this phenomenon or supplement it: even if vowel variability supports language development, the acoustic properties of each individual vowel token still need to be accounted for.

The present study tests whether the formant variability in IDS vowels is related to the expression of emotion in this register. Experiment 1 tests whether formant frequencies in IDS vowels are higher in utterances with more positive affect. Experiment 2 tests whether IDS utterances with raised vowel formant frequencies are perceived as smiled, child-like, or both.

2. Experiment 1

Experiment 1 tested emotion-related explanations of the IDS vowel formant variability by asking whether formant frequencies in IDS are higher in utterances with more positive affect.

To establish affect in IDS utterances, listeners were asked to rate the Valence and Energy of low-pass filtered IDS and ADS. Valence and Energy scales have been used extensively to study the acoustic correlates of affect [8]. Previous work on IDS has used scales such as Comfort and Direct Attention to understand the acoustic correlates of specific communicative intents [6]. Because the present study aims to understand acoustic variability in IDS in terms of general affective processes, the more general Valence and Energy scales were adopted.

The first sets of analyses established whether perceived Valence and Energy are higher in IDS than ADS, as well as related to the utterance F0 height and range. The final analyses tested whether F1 and F2 are higher in utterances with more positive Valence. The latter analyses only included utterances with the low-back vowel /ɑ/, because IDS causes the largest formant shifts in low [3], or low and back vowels [2, 4]. Predicting formant frequencies from rated affect is not circular, as listeners rated low-pass filtered speech with formants filtered out.

2.1. Methods

2.1.1. Participants

Participants were 15 volunteers from the SONA research participant database at the Radboud University Nijmegen, The Netherlands. All participants were adult monolingually raised native speakers of Dutch, with self-reported normal hearing and

(corrected to) normal vision. None of the participants were parents. Participants were compensated with Euro 10 for their participation. Data of 1 participant had to be excluded due to equipment failure. Demographic information from a second participant was missing due to experimenter error, but their data was included the analysis. We thus report data from 14 participants (11 female, 2 male, 1 unknown; 18-25 years old).

2.1.2. Stimuli

The stimuli came from a corpus of 1144 IDS and ADS utterances of 18 mothers who were recorded twice, at their infants' ages of 11 and 15 months [4]. For the IDS recordings, mothers were asked to play with their infants and some provided toys. ADS was elicited by experimenter questions about the infants' familiarity with and interest in the toys. The toy names contained the vowels /i/, /u/, /a/, or /ɑ/ in the stressed syllable (for more details: [4]).

Each utterance was manually re-annotated for its onset, offset, overlapping sounds, and voice quality. The 920 utterances with no sound overlap and a modal voice quality were selected for the present experiment (ADS=109; IDS at 11 months=473; IDS at 15 months = 338).

For presentation in the experiment, the utterances were extracted at zero crossings, low-pass filtered at 400Hz using the pass Hann filter with a smoothing of 100Hz as implemented in Praat [11], and scaled to the same loudness.

For later analyses, we measured the F0 of all utterances and the F1 and F2 of the vowel /ɑ/ as it appeared in the target words. The F0 curve of each (unfiltered) utterance was estimated in hertz using the cross-correlation method. The F0 range for the analysis was initially set at 120–400Hz. If the analysis of the median F0 failed, the F0 floor was updated to 75Hz, and if the analysis still failed, the criterion for the voicedness was lowered from 0.45 to 0.35. From these estimated curves, we computed the median F0 and the F0 range (maximum F0 minus minimum F0). Formants were measured in the central 40% of the vowel using the Burg-algorithm as implemented in Praat [11], which extracted 5 formants per frame with a ceiling of 5577 Hz. The median F1 and F2 were computed.

2.1.3. Procedure

Participants were randomly assigned to one of three experiment versions. Each version contained a third of the stimuli (version 1&2: 307; version 3: 306), equally taken from each speaker's ADS and IDS at both infant ages. Nine stimuli (three from each version) were also presented to all participants as practice trials.

The participants' task was to rate each utterance on two scales. The Valence scale ranged from "fully negative", to "neutral" to "fully positive". The Energy scale ranged from "fully calm", to "neutral", to "fully energetic". The scales were continuous — participants could indicate their rating by clicking anywhere on the scales. Each trial started by playing the stimulus and participants could replay once if desired. The next trial started as soon as the utterance had been rated on both scales.

Participants were instructed that they would listen to mothers speaking to their babies as if "listening through a wall": one cannot understand the words, but still recognise the speaker's affective state. Participants were told that they would rate each utterance for Valence and Energy, and were explicitly reminded that these are separate dimensions of affect.

The experiment started with the 9 practice trials, presented in randomised order. Participants could adjust the sound level during the practice trials and ask any additional questions afterwards. The experiment continued with the 306 or 307 experimental trials, which were randomised for each participant. Participants were required to take 2 breaks, during which they answered some on-paper questions, and filled out an exit questionnaire at the end. The procedure took at most 60 minutes.

The experiment was run in Praat's Demo window [11]

2.1.4. Analysis

The obtained Valence and Energy ratings ranged from 0 (the low end of the scale) to 100 (the high end of the scale). Each utterance was rated for Valence and Energy by 5 or 4 listeners, and these ratings were were averaged to obtain one Valence and one Energy score per utterance.

Data were analysed with linear mixed effects models using *R*'s *lmer* function [12]. Each analysis had one dependent variable and one or more independent variables. Analyses were conducted with different sets of dependent and independent variables, and on various subsets of the utterances. Both the variables and the subsets are addressed in more detail in the Results section. For each analysis, the dependent and the (continuous) independent variables were centred at 0 across all utterances in the subset, irrespective of speaker. All models were fit with by-speaker random intercepts and by-speaker slopes for each independent variable. Covariances between the random effects were not estimated due to convergence problems in some models. Statistical significance of the independent variables was evaluated by treating the t-statistic as a z-statistic, and thus interpreting $t > |1.96|$ as significant at alpha of 0.05.

Table 1: *The scores for Valence and Energy (Experiment 1) and Face and Child (Experiment 2). Mean scores and standard deviations (between parentheses) are computed across all utterances included in the respective analyses.*

experiment	scale	IDS-11	IDS-15	ADS
1	Valence	57.241 (13.428)	57.815 (13.577)	41.421 (13.801)
1	Energy	51.524 (13.815)	53.237 (14.557)	42.52 (16.000)
2	Child	72.604 (13.053)	70.099 (13.749)	20.561 (14.919)
2	Mouth	61.838 (10.11)	60.493 (9.725)	44.747 (10.142)

2.2. Results

The mean Valence and Energy scores for the utterances in ADS, and in IDS to 15- and 11-month olds can be found in Table 1.

The Valence and Energy scores were the dependent variables in two separate analyses conducted on all 920 utterances with the independent variable Register (IDS=0.5 versus ADS=-0.5). Dutch IDS is more positive in Valence and higher in Energy than ADS (Valence: $\beta = 16.425, t = 9.339, p < 0.001$; Energy: $\beta = 10.111, t = 5.296, p < 0.001$). This difference between IDS and ADS in both Valence and Energy validates the scores.

The Valence and Energy scores of the 811 utterances from IDS -thus excluding the ADS utterances- were the dependent

variables in two separate analyses with the independent variable Infant Age (11-months=0.5 versus 15-months=-0.5). We found no evidence for differences in Valence or Energy between Dutch to 11- and 15-month-olds (Valence: $\beta = -1.0153, t = -1.062, p = 0.288$; Energy: $\beta = -1.640, t = -1.138, p = 0.255$). The absence of significant Valence and Energy differences between the IDS to 11- and 15-month-olds warrants collapsing these data in the subsequent analyses.

We then regressed the Valence and Energy scores in the 811 IDS utterances on the continuous independent variables F0 median and F0 range. A more positive utterance Valence can be predicted from a higher F0, whereas no association was observed between Valence and F0 range (F0 median: $\beta = 0.03, t = 2.582, p < 0.05$; F0 range: $\beta = 0.009, t = 1.356, p = 0.175$). A higher utterance Energy can be predicted from both a higher F0 and a larger F0 range (F0 median: $\beta = 0.043, t = 4.804, p < 0.001$; F0 range: $\beta = 0.037, t = 4.075, p < 0.001$).

The final, and for the purposes of this study most interesting, two analyses were conducted on the 255 utterances in IDS containing a target word with /ɑ/. The dependent variables were F1 and F2 of the vowel /ɑ/, and the continuous independent variables were the utterance Valence and Energy scores. A higher F1 of /ɑ/ can be predicted from a higher utterance Energy, whereas no association between F1 and utterance Valence was observed (Valence: $\beta = 0.001, t = 0.32, p < 0.749$; Energy: $\beta = 0.011, t = 2.435, p < 0.015$). A higher F2 in /ɑ/ is not reliably associated with either aspect of affect (Valence: $\beta = 0.005, t = 0.96, p < 0.337$; Energy: $\beta = -0.002, t = -0.381, p < 0.703$).

2.3. Conclusion and Discussion

The higher perceived positive valence in Dutch IDS replicates findings on Australian-English [6]. The present results are, to our knowledge, the first to directly show that IDS is perceived as more energetic than ADS. The emotion that is high in positive valence and energy is "happiness", confirming that IDS may be parsimoniously described as very happy speech [7].

A more positive valence is associated with a higher F0, here and in Australian-English IDS [6]. This casts doubt on the claim that utterance valence is primarily associated with articulation and not F0 [8]. The valence-F0 association may be specific to IDS, or result from a correlation between F0 and an unmeasured valence cue. Alternatively, rating valence in low-pass filtered speech may cause an atypically high reliance on F0.

A higher perceived energy is associated with a higher F0 and larger F0 range, which is in line with the claim that utterance energy primarily affects F0 [8]. Since mothers use both a high F0 and a large F0 range when they encourage their infants' attention [6], high-energy speech may contribute to the attentional function of IDS.

Contrary to our key predictions, the F1 and F2 variability between vowels is *not* explained by the utterance valence. The absence of the predicted valence-formant association may show that formant frequencies in IDS are *not* higher in utterances with higher positive valence. Specifically in IDS, utterances with positive valence may be produced with spread lips to express happiness or with pouted lips to express comfort. The association between valence and vowel formants might therefore be weaker or more complex than predicted.

Contrary to the theory that that utterance energy primarily affects F0, not articulation [8], we found that F1 variability *is* related to energy. A high F1 is associated with an opened mouth and a surprised open mouth is a frequent facial expression in IDS [13]. Possibly, mothers use high-energy speech in combination with a surprised facial expression to encourage attention. The energy-F1 association may thus arise from a common cause rather than a direct effect of energy on F1.

3. Experiment 2

Experiment 2 assessed the perceptual consequences of the variability in IDS vowel formant frequencies by asking whether utterances with raised vowel formant frequencies are perceived as smiled, child-like, or both.

Listeners were asked to rate the perceived Mouth shape and Child-likeliness of unfiltered IDS and ADS. These new scales were developed for the purposes of the present study. The first sets of analyses established whether perceived Mouth shapes and Child likeliness are higher in IDS than ADS. The critical analyses tested whether utterances with higher formant frequencies are perceived to be produced with a more retracted Mouth shape –and thus with larger smiles– as well as perceived to be more Child like. For reasons described in experiment 1, the latter analyses only included utterances with/ɑ/.

3.1. Methods

Only the differences with Experiment 1 are indicated below.

3.1.1. Participants

Participants in Experiment 2 were 15 volunteers who had not participated in Experiment 1. Data of 1 participant had to be excluded due to equipment failure. We thus report data from 14 participants (10 female, 4 male; 19-30 years old).

3.1.2. Stimuli

Stimuli were the same 920 utterances, but not low-pass filtered.

3.1.3. Procedure

The two scales in this experiment were Mouth and Child. The Mouth scale ran from "pouted lips", to "neutral mouth", to "retracted corners of the mouth". The experimenters provided examples of pouting and smiling lips during their explanation of the Mouth scale. The Child scale ran from "adult-like", to "intermediate", to "child-like".

The instruction did not mention that the speech would sound as if heard through a wall. Participants were asked to base their answer on the tone of voice and not on the content.

3.2. Results

The mean Mouth and Child scores to the utterances from ADS, and IDS to 15- and 11-month olds can be found in Table 1. All analyses reported here have Mouth and Child as the dependent variables in two separate models that are otherwise identical.

Two analyses on all 920 utterances with the independent variable Register (IDS=0.5; ADS=-0.5)] showed that Dutch IDS is perceived to be produced with more retracted lips and in a more child-like manner than ADS (Mouth: $\beta = 16.562, t = 12.276, p < 0.001$); Child: $\beta = 50.624, t = 20.479, p < 0.001$). The difference between IDS and ADS in both Mouth and Child scores validates the scales.

Two analyses on the 811 utterances from IDS with the independent variable Infant Age (11 months=0.5; 15 months=-0.5) showed that IDS to 11-month olds is perceived to be

possibly produced with more retracted lips and certainly in a more child-like manner than IDS to 15-month olds (Mouth: $\beta = 1.546, t = 1.872, p < 0.061$; Child: $\beta = 2.866, t = 2.539, p < 0.011$). The (marginally) significant effects of Infant Age warrant maintaining Infant Age as a factor in the subsequent analyses.

The final two analyses were conduced on the 255 utterances from IDS containing a target word with /ɑ/. The continuous independent variables were the F1 and F2 of /ɑ/, the F0 median and F0 range of the utterance, and Infant Age. Each of the acoustic predictors was also allowed to interact with Infant Age. The by-speaker slopes for Infant Age and the interactions were excluded from the models because of convergence problems.

IDS utterances are perceived to be produced with more retracted lips when the vowel /ɑ/ has a higher F2 and when the utterance F0 is higher (F2: $\beta = 1.76, t = 2.797, p- = 0.005$; F0 median: $\beta = 0.026, t = 1.753, p = 0.08$). The Mouth scores were not significantly predicted from either vowel F1 or utterance F0 range (F1: $\beta = 0.221, t = 0.383, p = 0.702$; F0 range: $\beta = 0.014, t = 1.559, p = 0.119$).

IDS utterances are perceived to be produced in a more child-like manner when the vowel /ɑ/ has a higher F2 and when the utterance F0 is higher. (F2: $\beta = 2.923, t = 3.326, p = 0.001$; F0 median: $\beta = 0.044, t = 2.317, p = 0.02$). The Child scores in IDS were not significantly predicted from either vowel F1 or utterance F0 range (F1: $\beta = 0.914, t = 1.022, p < 0.307$; F0 range: $\beta = 0.016, t = 1.575, p = 0.115$). The effects of Infant Age and the interactions between the acoustic predictors and Infant Age were not significant in either model.

3.3. Conclusion and Discussion

These results provide the first evidence that IDS sounds more smiled and more child-like than ADS. IDS is perceived to be more smiled and child-like to 11- than to 15-months-old, which mirrors the higher F2 in IDS to 11-month-olds [4].

Critically, utterances in IDS are perceived as more smiled or more child-like when they contain a vowel with a relatively high F2 or have a higher utterance F0. The present study study thus goes beyond the observation that listeners perceive smiles and frowns from speech with, respectively, a raised and lowered F2 [9], and demonstrates listeners' sensitivity to vowel formant differences in a speech register that is 'happy' across the board.

Interestingly, the aforementioned group effect of infant age on the perception of smiles and child imitation disappears once the vowel F2 and F0 of each individual utterance is considered. The perceived difference between IDS to 11- and 15-month-olds may thus be a direct consequence of the acoustics.

Although smiles and child speech are generally associated with a higher F2 as well as F1 [9, 10], no associations were observed between vowel F1 and the perceptual scores. Note that F1 was not higher in IDS than in ADS in the corpus from which the stimuli were taken [4]. Combined with the present results, this suggests that F2 may be a stronger cue than F1 to smiles and the imitation of child speech, at least in IDS.

4. General Discussion

Two experiments, which constitute the first study on listener ratings of Dutch IDS, explored emotion-related explanations for and perceptual consequences of the acoustic differences between individual vowel tokens within IDS. Experiment 1 failed to confirm the prediction that a more positive utterance valence can explain higher vowel formant frequencies. However, a higher utterance energy can explain a higher vowel first formant (F1). Experiment 2 confirmed the prediction that a consequence of a higher second formant (F2) is the percept of more smiled and more child-like speech. Variability in the production of vowels in IDS is thus associated with some aspects of (perceived) speaker emotion.

These results can inform future studies which directly manipulates speaker affect [14]. Direction manipulation of speaker affect is required to establish that the emotion a speaker wishes to express directly explains her vowel formant frequencies. The present findings would be confirmed if F1 is higher in utterances with high-energy emotions, such as anger or surprise, than in utterances with low-energy emotions, such as sleepiness or calmness, and if F2 is higher when the parent is instructed to smile or imitate their infant. A controlled study will also shed further light on the presence or absence of an association between speaker valence and the vowel formant frequencies in IDS.

5. References

[1] Kuhl, P.K. and Andruski, J.E. and Chistovich, I.A. and Chistovich, L.A. and Kozhevnikova, E.V. and Ruskina, V.L. and Stolyarova, E.I and Sundberg, U. and Lacerda, F., "Cross-Language Analysis of Phonetic Units in Language Addressed to Infants", Science, 277(5044):684–686, 1997.

[2] Englund, K. and Behne, D., "Infant Directed speech in Natural Interaction – Norwegian vowel Quantity and Quality", Journal of Psycholinguistic Research, 34:3: 259–280, 2005.

[3] Green, J.R. and Nip, I.S.B. and Wilson, E.M. and Mefferd, A.S. and Yunusova, Y., "Lip Movement Exaggerations During Infant-Directed Speech", Journal of Speech, Language, and Hearing Research, 53:1529–1542, 2010.

[4] Benders, T., "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion,not didactic intent", Infant Behavior and Development, 36:847–862, 2013.

[5] Soderstrom, M., "Beyond Babytalk: Re-evaluating the Nature and Content of Speech Input to Preverbal Infants", Developmental Review, 27:501–532, 2007

[6] Kitamura, C. and Burnham, D., "Pitch and Communicative Intent in Mothers' Speech: Adjustments for Age and Sex in the First Year", Infancy, 4(1):85–100, 2003.

[7] Trainor, L.J. and Austin, C.M. and Desjardins, R.N., "Is Infant-Directed Speech Prosody a Result of the Vocal Expression of Emotion?", Psychological Science, 11(3):188–195, 2000.

[8] Scherer, K.R., "Vocal Affect Expression: A Review and a Model for Future Research, Psychological Bulletin,99:143–165, 1986.

[9] Tartter, V.C. and Braun, D., "Hearing Smiles and Frowns in Normal and Whisper Registers", Journal of the Acoustical Society of America, 96(4):2101–2107, 1994.

[10] Vorperian, Houri K and Kent, Ray D, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data", Journal of Speech, Language, and Hearing Research, 50(6), 1510–1545, 2007.

[11] Boersma, P. and Weenink, D., "Praat, Doing Phonetics By Computer", [Computer Program], 2014, http://www.praat.org/.

[12] Bates, D. and Maechler, M. and Bolker, B., "lme4: Linear mixed-effects models using S4 classes", [Computer Program], 2012, http://CRAN.R-project.org/package=lme4.

[13] Chong, S.C.F. and Werker, J. and Russell, J.A. and Caroll, J.M., "Three Facial Expressions Mothers Direct to Their Infants", Infant and Child Development, 12:211–232, 2003.

[14] Fernald, A., "Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message?", Child Development, 60(6):1497:1510, 1989.