Research Report

Psychological Science
1–7

DOI: 10.1177/0956797614562453
pss.sagepub.com

Mothers Speak Less Clearly to Infants Than to Adults: A Comprehensive Test of the Hyperarticulation Hypothesis

Andrew Martin[1], Thomas Schatz[2,3], Maarten Versteegh[2], Kouki Miyazawa[1], Reiko Mazuka[1,4], Emmanuel Dupoux[2], and Alejandrina Cristia[2]

[1]Laboratory for Language Development, RIKEN Brain Science Institute, Saitama, Japan; [2]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Études Cognitives de l'Ecole Normale Supérieure, École des Hautes Études en Sciences Sociales, Centre National de la Recherche Scientifique; [3]SIERRA Project-Team, Département d'Informatique de l'Ecole Normale Supérieure, Institut National de Recherche en Informatique et en Automatique, Centre National de la Recherche Scientifique; and [4]Department of Psychology and Neuroscience, Duke University

Abstract

Infants learn language at an incredible speed, and one of the first steps in this voyage is learning the basic sound units of their native languages. It is widely thought that caregivers facilitate this task by hyperarticulating when speaking to their infants. Using state-of-the-art speech technology, we addressed this key theoretical question: Are sound categories clearer in infant-directed speech than in adult-directed speech? A comprehensive examination of sound contrasts in a large corpus of recorded, spontaneous Japanese speech demonstrates that there is a small but significant tendency for contrasts in infant-directed speech to be less clear than those in adult-directed speech. This finding runs contrary to the idea that caregivers actively enhance phonetic categories in infant-directed speech. These results suggest that to be plausible, theories of infants' language acquisition must posit an ability to learn from noisy data.

Keywords

speech perception, psycholinguistics, language development, infant-directed speech, hyperarticulation

Received 6/21/14; Revision accepted 11/12/14

A great deal of research on infants' language acquisition has focused on the interplay between the nature of the input and the cognitive biases that inform the learning procedure. This relationship is often described in terms of a trade-off: The richer the input, the simpler the task required of the infant learner. It is therefore crucial that researchers understand the precise nature of infants' speech input when investigating language acquisition, particularly because the phonetic properties of infant-directed speech (IDS) differ substantially from those of adult-directed speech (ADS).

Some of these properties have been investigated in the context of this interplay (e.g., De Boer & Kuhl, 2003; Fernald, 2000). For example, in a highly influential study, Kuhl et al. (1997) showed that the corner vowels /i/, /u/, and /ɑ/ (as in *sheep, shoe*, and *shop*) were pronounced further apart in acoustic space in IDS compared with ADS. On the basis of these and similar findings (for a recent overview, see Cristia, 2013), Kuhl et al. argued that speakers (unconsciously) make the task of acquiring a language less formidable by enriching the input. We will call this the *hyperarticulation hypothesis*.

More recent findings, however, have begun to cast doubt on the idea that the differences found by Kuhl et al. (1997) do in fact contribute to learnability. Kirchhoff and Schimmel (2005), for example, found that the

Corresponding Author:
Alejandrina Cristia, Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS, DEC-ENS, CNRS, 29 rue d'Ulm, Paris 75005, France
E-mail: alecristia@gmail.com

phonetic distributions of /i/, /u/, and /ɑ/ overlap more in IDS than in ADS, which lowers performance when automatic speech-recognition algorithms are trained and tested on IDS input. Other researchers, also focusing on a handful of contrasts, found no clear improvement in IDS (e.g., McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013) and sometimes even a decrease in between-category distances (e.g., Benders, 2013). However, these few counterexamples do not prove that IDS never serves a pedagogical function. Indeed, speakers may hyperarticulate not all contrasts but only the ones that infants are learning at a given time (e.g., Sundberg, 1998). If so, on average, phonetic contrasts would be clearer in IDS after all.

The only way to test this possibility is to simultaneously study all contrasts that are present in both registers (i.e., ADS and IDS). Naturally, this is virtually impossible using traditional phonetic analysis, which presupposes the measurement of different phonetic properties for different contrasts (e.g., formants for vowels, but burst spectra for stop consonants). It is therefore difficult to pool results across different types of contrasts.

In the present work, we adopted a comprehensive approach by using a novel procedure from speech technology, the minimal-pair ABX task of Schatz et al. (2013). In this procedure, an algorithm classifies a given phonetic token X as belonging to the same phoneme category as a one of two other tokens (A and B). If mothers enhance phonetic contrasts when speaking to infants, tokens should be overall easier to classify in IDS compared with ADS, and the algorithm's accuracy should thus be higher in IDS than in ADS. We tested this prediction using a corpus that contains spontaneous IDS and ADS, which ensured that our data were a realistic approximation of the actual input to which infants are exposed. The corpus is relatively large, so our results should be robust. We show that, contrary to the hyperarticulation hypothesis, contrasts are overall more difficult to discriminate in IDS than in ADS.

Method

The corpus

Our data were taken from the RIKEN Japanese Mother-Infant Conversation Corpus (for further details, see Igarashi, Nishikawa, Tanaka, & Mazuka, 2013; Mazuka, Igarashi, & Nishikawa, 2006), which contains speech from 22 mothers in two conditions: addressing their infants (between 18 and 24 months old; about 11 hr of speech in total) and speaking with an adult experimenter (about 3 hr in total). The IDS portion was gathered while each mother viewed picture books or engaged in free play with her child, and the ADS portion consisted of spontaneous conversation with an experimenter. The mothers' speech was captured using a headset dynamic microphone and was recorded on digital audiotapes at a sampling frequency of 41 kHz. The corpus has been carefully coded at several levels; the only levels relevant to the present study are the segmental (or phone) level and the prosodic level (i.e., where sentence boundaries are indicated). In previous work, we have documented that these IDS samples bear all the well-known attributes of IDS: significantly higher pitch, an expanded pitch range, and shorter utterance length compared with ADS (Igarashi et al., 2013).

Discriminability calculations

Previous researchers have relied on near-minimal pairs of words (e.g., *sheep* and *shoe*, or *sock* and *shoe*), elicited using toy objects (Kuhl et al., 1997), to study a selection of contrasts in relatively controlled contexts. Such an approach was not optimal for our goal of gaining a comprehensive view of the register differences in a corpus of spontaneous speech. Instead, we used syllabic minimal pairs—pairs of syllables that differ in only one segment.

Syllable boundaries were identified from the segmental descriptions by applying syllabification rules appropriate to Japanese. We allowed syllables with the shapes (C)(g)V(N), where C stands for a consonant, g for a glide, V for a long or short vowel, N for a moraic nasal, and parentheses indicate optional elements. Sequences of vowels were parsed into separate syllables.

Combining phonological and phonetic criteria, and considering the frequency distribution of syllables in the corpus, we decided to treat certain phones as variants of the same segment and to remove others from consideration. Geminate and singleton consonants were collapsed into a single category, and the preceding syllables were classified as open. Sounds that were too infrequent (e.g., long or devoiced vowels) or that did not occur in syllabic minimal pairs (e.g., the moraic nasal /N/) were not included in the comparisons, although they were considered as part of the context. Specifically, the following segments were included in the calculation of discriminability scores:

Onsets: /p/, /pʲ/, /t/, /tʲ/, /k/, /kʲ/, /kʷ/, /b/, /bʲ/, /d/, /dʲ/, /g/, /gʲ/, /gʷ/, /t͡s/, /t͡ʃ/, /s/, /ʃ/, /z/, /d͡ʒ/, /h/, /hʲ/, /Φ/, /Φʲ/, /v/, /m/, /mʲ/, /n/, /nʲ/, /r/, /rʲ/, /w/, /j/

Vowels: /a/, /e/, /i/, /o/, /ɯ/

We calculated the discriminability between each of the onsets and every other onset, keeping the rest of the syllable (the context), the register, and the sentence position constant. Likewise, we estimated the discriminability between each of the vowels and every other vowel

among those listed above, controlling for the context, the register, and the sentence position. To estimate discriminability, we first identified all the tokens of two categories (e.g., /ɯ/-/o/) in the speech of a given talker. We did this separately for each syllabic minimal pair (e.g., /ɯ/-/o/, /kɯ/-/ko/, and /sɯ/-/so/), register (IDS, ADS), and sentence position (initial, medial, final, or in isolation).

Second, we converted each token into a sequence of spectral frames on a Mel frequency scale (O'Shaughnessy, 1987). These frames were computed by running the speech through 13 filter banks centered between 100 and 6855 Hz and applying a cubic root compression on the dynamic range of the output of each of these filters. Dynamic time warping (DTW) was used to compute the optimal frame alignment between each pair of tokens. Given a specific pair of tokens, the algorithm looked for the optimal alignment path to discover frame correspondences between the two tokens so as to optimize how the spectral contrast was captured. The distance between two frames was computed as the inverse cosine of the normalized scalar product between the two vectors, each containing 13 values. The sum of the frame-wise distances along the optimal path was used as the distance between that pair of tokens. DTW has been used in a variety of previous phonetic research (e.g., Cummins, 2009; for a clear explanation of DTW for nonspecialists, see Kirchner, Moore, & Chen, 2010).

This analysis pipeline is one of many that have been developed in speech technology. This specific pipeline was chosen before data analysis on the basis of independent performance evidence (e.g., the 13-channel, compressed-Mel-spectrogram representation fares well in direct comparisons with several other representations; Schatz et al., 2013) and psychological and physiological plausibility (i.e., all the operations involved—short-term spectral analysis, Mel warping of the frequency scale, and compression of the dynamic range—can be related to documented aspects of human auditory processing; Moore & Moore, 2003, Chapters 3–6; Schnupp, Nelken, & King, 2011, Chapter 2).[1] However, we do not have conclusive experimental data showing that these parameters represent the way that infants encode speech. This is a widespread issue—no study comparing the acoustic properties of IDS and ADS has justified the researchers' choice of parameters through experiments with infants. This uncertainty is an inevitable consequence of the difficulty of collecting perceptual data from infants, particularly at the large scale required for our study. Indeed, our approach provides a unique perspective on infants' spoken input and is more systematic than is possible in experiments with humans.

Third, the algorithm compiles all possible triplets of tokens. Imagine that there are two tokens for /sɯ/ (/$sɯ_1$/ and /$sɯ_2$/) and one for /so/ in sentence-medial position in the IDS of a given talker. The algorithm can form two ABX triplets: /$sɯ_1$/-/so/-/$sɯ_2$/ and /$sɯ_2$/-/so/-/$sɯ_1$/. In each triplet, the algorithm compares the distance between X and A and that between X and B (e.g., in the first triplet, /$sɯ_1$/-/$sɯ_2$/ versus /so/-/$sɯ_2$/). If the distance between X and A is smaller than that between X and B, the algorithm returns a response of A; otherwise, the algorithm gives a response of B. Because the identity of the tokens is known, a response can be correct (if the X token was indeed drawn from the A category, as in the example) or incorrect (if it was not). The final score was the proportion of responses that were correct among all the triplets that had been compiled, and the compilation was done separately for each syllabic minimal pair, speaker, register, and sentence position.

A total of 28,629 unique scores were computed. Because the corpus consists of spontaneous speech, the frequency of the syllabic minimal pairs varies considerably from speaker to speaker, and some pairs are not available for all speakers. Indeed, each score was based on comparisons of between two triplets and nearly six million triplets (median = 42). To determine which estimations were reliable, we randomly split the speakers into two groups, and then we calculated the correlation in scores separately for each minimal pair (but collapsed the data across register and sentence position) across the two arbitrary groups. We used visual inspection to determine the minimum number of occurrences needed to achieve good correlations (further details can be found at Open Science Framework, https://osf.io/it8ab/). After exclusion of such cases, there were 21,109 unique scores.

We sought to avoid confounding register with either prosodic position or the prevalence of certain contexts or contrasts. We therefore performed separate analyses on sentence-initial, sentence-medial, and sentence-final syllables (there were not enough triplets in isolation to warrant analyses). Sentence-medial syllables were the most prevalent (13,629 syllables, of which 6,061 were IDS and 7,568 were ADS), and only in this data set did we find minimal pairs present in both registers for all 22 speakers. Therefore, we report in full only the results for the sentence-medial data set.[2] This resulted in the selection of 118 syllabic minimal pairs. Thus, the analyses reported here included 5,192 unique scores (22 speakers × 2 registers × 118 minimal pairs), achieving an unprecedented coverage of 10 vowel contrasts and 36 onset contrasts.

Results

Our null hypothesis states that ABX scores do not differ across the two registers. To evaluate whether the null hypothesis can be rejected, we first calculated a difference score (ABX score for IDS minus ABX score for ADS) for each speaker and minimal pair. These scores can be

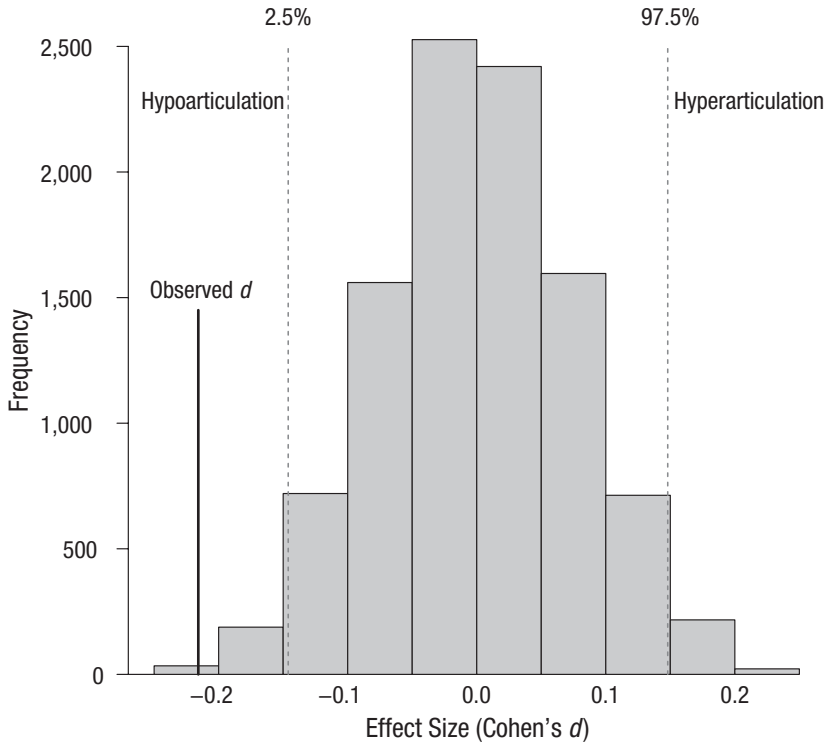

Fig. 1. The distribution of effect sizes obtained in the permutation test. Also shown are the observed effect size in the data set and the cutoff points for evidence of hypoarticulation and hyperarticulation.

viewed as forming a matrix in which the 22 speakers are the rows and the 118 minimal pairs are the columns. For each minimal pair, we calculated the effect size (Cohen's *d*) as the mean difference score divided by the standard deviation (over speakers). Positive effect sizes indicated that IDS scores were higher than ADS scores, and negative effect sizes indicated that ADS scores were higher than IDS scores. The median effect size across the 118 minimal pairs was −0.216.

To assess whether this effect size could arise by chance, we used a permutation test (Fisher, 1935). Specifically, we constructed 10,000 matrices; in each one, we flipped the signs of the difference scores for all the minimal pairs associated with a randomly drawn number and selection of speakers. We estimated the median effect size for each of these 10,000 matrices and used the resulting distribution as that corresponding to our null hypothesis. As is evident in the histogram in Figure 1, it is highly unlikely that we would have observed an effect size of –0.216 if the null hypothesis had been true. Indeed, –0.216 is below the 2.5th percentile of the distribution of effect sizes under the null hypothesis (from our permutation test, the exact $p = .0023$). The finding that IDS contrasts were significantly deteriorated compared with ADS contrasts was replicated in a range of analyses, except analyses of syllable-final minimal pairs, for which the difference was not significant (results, as well as the data and scripts used to generate them, are available at Open Science Framework; see note 2).

We illustrate these results further with two figures in which we have collapsed minimal pairs across contexts (e.g., we calculated the median score among /ɯ/-/o/, /kɯ/-/ko/, and /sɯ/-/so/ separately for each register and speaker, to represent the contrast between /ɯ/ and /o/). Most of the contrasts (34 of 46) had negative effect sizes, and negative effect sizes tended to be larger than positive ones, as is evident in Figure 2. In Figure 3, average ABX score for IDS is plotted as a function of average ABX score for ADS. Note that most of the 95% confidence intervals (ellipses) fall below the diagonal.

Conclusions

In a comprehensive comparison of many types of contrasts in a corpus of spontaneous Japanese speech, we found no evidence that phonetic contrasts are enhanced overall in IDS. To the contrary, we found a small but significant advantage for ADS contrasts, which suggests that, if anything, the system of phonetic contrasts is deteriorated in IDS compared with ADS. These results are inconsistent with a strong version of the hyperarticulation hypothesis (which would posit that IDS is characterized

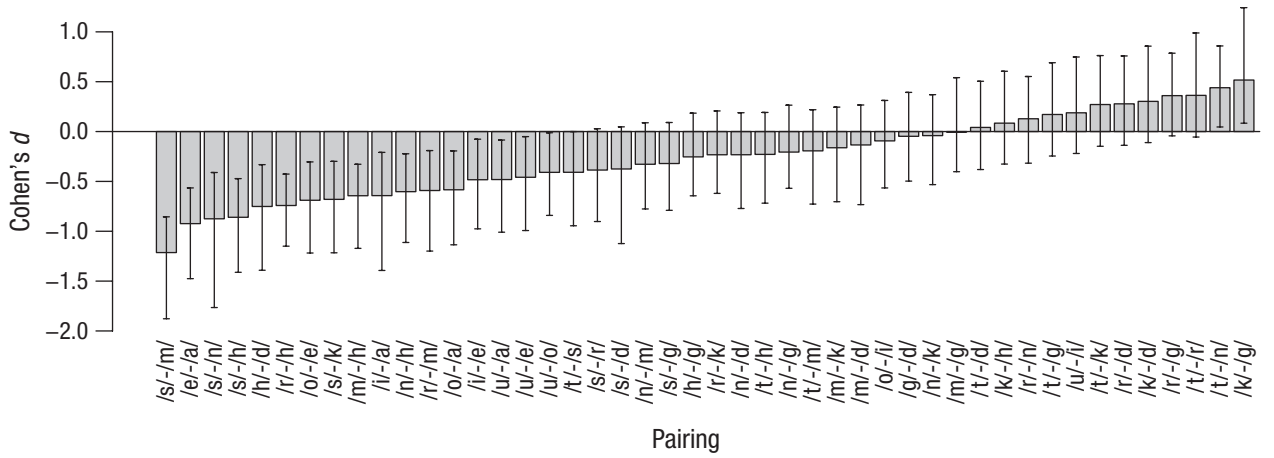

Fig. 2. Effect size as a function of contrast (collapsed across contexts). The error bars represent 95% confidence intervals from bootstrap resampling (N = 10,000 samples). /u/ stands for /ɯ/.

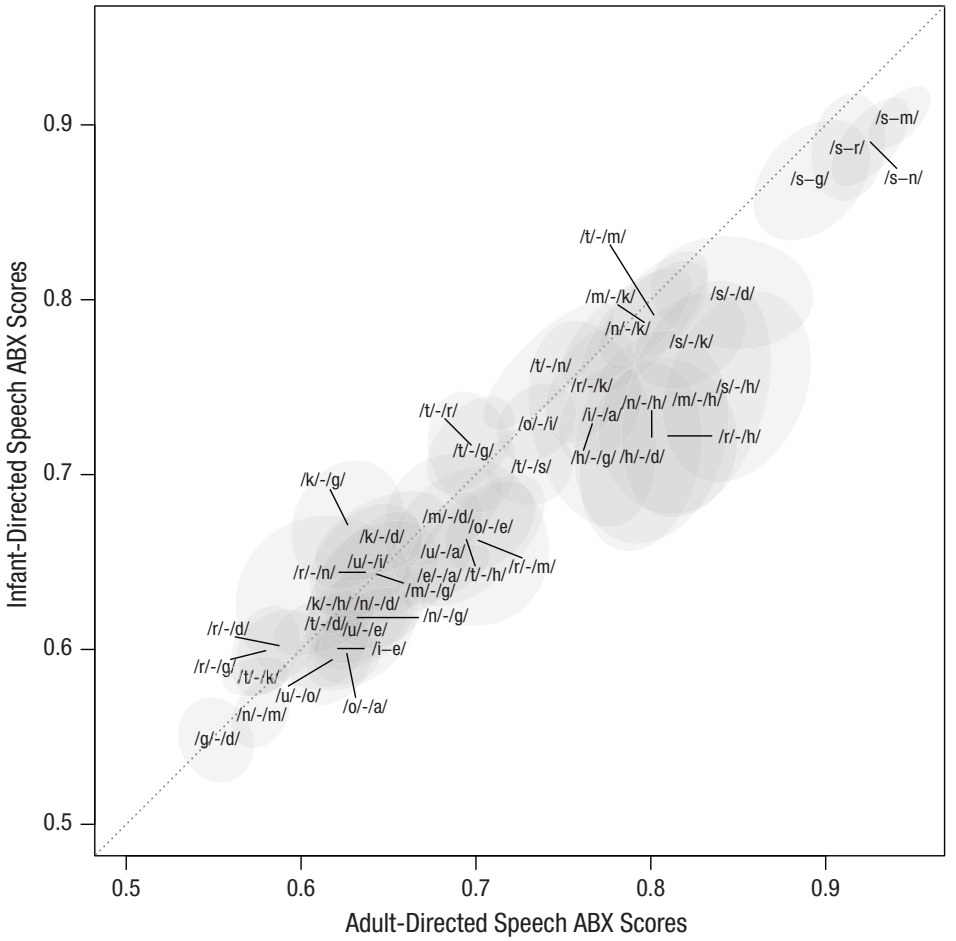

Fig. 3. Average ABX score for infant-directed speech as a function of average ABX score for adult-directed speech for each of the contrasts (collapsed across contexts). For each contrast, the average across speakers is either at the center of the contrast label or at the endpoint of the line leading from the contrast label. The ellipses indicate 95% confidence intervals over the 22 speakers. The dotted gray line indicates where points would fall if the averages were equal. /u/ stands for /ɯ/.

by widespread enhancement of pronunciation) as well as with weaker versions (in which a loss of clarity in some contrasts would be balanced by enhancements in others). However, our results are perfectly consistent with both recent empirical work (McMurray et al., 2013) and novel theoretical perspectives that highlight the communicative function of IDS and how it shapes infant development by boosting cognitive processing (Benders, 2013; Ramirez-Esparza, Garcia-Sierra, & Kuhl, 2014; Weisleder & Fernald, 2013).

Our study is the first to attack the key question of the input fueling infants' acquisition of spoken language by examining a broad range of spectral contrasts in a language. Our approach allowed us to identify large-scale patterns common to many contrasts. Moreover, the discrimination measure we used takes into account the full complexity of the category structure and not just isolated features (e.g., the mean value along some acoustic dimension).

Future work can improve on this research in two main ways. First, because we modeled each sound with a high-dimensional set of parameters, it was difficult to directly interpret the proximal and distal factors that render two sound categories harder or easier to discriminate (decreased acoustic distance vs. increased variance, or the degree to which these phenomena are triggered by greater amounts of smiling or whispering in IDS, etc.). Establishing the cause of the deterioration we found will require a different approach (e.g., eliciting speech under specific conditions).

Second, this study should be extended by examining other contrasts (e.g., those that depend mainly on temporal cues, such as consonant gemination) and other populations (e.g., infants in other age groups or with other linguistic backgrounds). Although we have no reason to suppose that our robust and stable findings would not be replicated in additional conditions, our conclusions would be bolstered by independent corroborating evidence.

In sum, our findings invite three key conclusions. First, descriptions of IDS that specify enhancement as a necessary feature are simply not appropriate. Second, it is important to study real IDS, because it is significantly different from ADS. Finally, the hypothesized learner must be prepared to face even more variability than is found in typical ADS. The study of IDS can thus play a crucial role in constraining theories and models of language acquisition.

Author Contributions

R. Mazuka oversaw the collection and coding of the corpus. K. Miyazawa provided key analyses. A. Martin wrote the syllabification algorithms. M. Versteegh wrote the feature-extraction algorithms. T. Schatz designed the ABX task. E. Dupoux carried out the acoustical analyses. A. Cristia carried out the statistical analyses, A. Martin and A. Cristia produced the first draft, and all authors contributed to the definition of the research question, the methodological approach, and the writing of this manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by the European Research Council (Grant ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (Grants ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-IDEX-0001-02 PSL*, and ANR-10-LABX-0087 IEC), the Fondation de France, and the Japan Society for the Promotion of Science (Kakenhi Grant 24520446, to A. Martin).

Notes

1. We additionally inspected Mel-frequency Cepstral coefficients, which yielded the same pattern of results (data available upon request).
2. Results for the sentence-initial and sentence-final pairs, along with the full list of minimal pairs, can be found at Open Science Framework, https://osf.io/it8ab/.

References

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior & Development*, *36*, 847–862.

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Language & Linguistics Compass*, *7*, 157–170.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*, 16–28.

De Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*, 129–134.

Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. *Phonetica*, *57*, 242–254.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Igarashi, Y., Nishikawa, K., Tanaka, K., & Mazuka, R. (2013). Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech. *The Journal of the Acoustical Society of America*, *134*, 1283–1294.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*, 2238–2246.

Kirchner, R., Moore, R. K., & Chen, T. Y. (2010). Computing phonological generalization over real speech exemplars. *Journal of Phonetics*, *38*, 540–547.

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.

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for learning Japanese: Riken Japanese mother-infant conversation corpus. *Technical Report of the Institute of Electronics, Information, and Communication Engineers (IEICE), TL2006-16*, *106*(165), 11–15.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*, 362–378.

Moore, B. C., & Moore, B. C. (2003). *An introduction to the psychology of hearing* (5th ed.). Bingley, England: Emerald Press.

O'Shaughnessy, D. (1987). *Speech communication: Human and machine*. Reading, MA: Addison-Wesley.

Ramirez-Esparza, N., Garcia-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants is linked to concurrent and future speech development. *Developmental Science*, *17*, 880–891.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (Eds.), *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association* (pp. 1781–1785). Baixas, France: International Speech Communication Association.

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*. Cambridge, MA: MIT Press.

Sundberg, U. (1998). *Mother tongue – Phonetic aspects of infant-directed speech* (Doctoral dissertation, University of Stockholm). Retrieved from http://www.ling.su.se/perilus/perilus-xxi/perilus-xxi-1.84573

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152.