Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability

Frans Adriaans[a)]
Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, the Netherlands

Daniel Swingley
Department of Psychology, University of Pennsylvania, 425 South University Avenue, Philadelphia, Pennsylvania 19104, USA

(Received 19 October 2016; revised 12 April 2017; accepted 12 April 2017; published online 4 May 2017)

Perceptual experiments with infants show that they adapt their perception of speech sounds toward the categories of the native language. How do infants learn these categories? For the most part, acoustic analyses of natural infant-directed speech have suggested that phonetic categories are not presented to learners as separable clusters of sounds in acoustic space. As a step toward explaining how infants begin to solve this problem, the current study proposes that the exaggerated prosody characteristic of infant-directed speech may highlight for infants certain speech-sound tokens that collectively form more readily identifiable categories. A database is presented, containing vowel measurements in a large sample of natural American English infant-directed speech. Analyses of the vowel space show that prosodic exaggeration in infant-directed speech has the potential to support distributional vowel learning by providing the learner with a subset of "high-quality" tokens that infants might attend to preferentially. Categorization models trained on prosodically exaggerated tokens outperformed models that were trained on tokens that were not exaggerated. Though focusing on more prominent, exaggerated tokens does not provide a solution to the categorization problem, it would make it easier to solve. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4982246]

[JFL] Pages: 3070–3078

I. INTRODUCTION

During the first year of life infants begin to discover the phonetic categories that define the consonants and vowels of their native language. Learning categories is a challenging task, as it requires grouping together complex acoustic tokens that show considerable variability along multiple acoustic dimensions. Nevertheless, infants evince knowledge of language-specific vowel categories around 6 months (Kuhl *et al.*, 1992; Polka and Werker, 1994), and language-specific consonant perception has been found around 10 months of age (Werker and Tees, 1984). This adaptation to the native language's speech sounds is viewed as a critical developmental step in language acquisition because of the role of phonetic categories in defining lexical and morphological distinctions.

It is widely accepted that infants' early category learning involves a distributional learning mechanism that detects clusters of tokens that are similar along relevant acoustic dimensions. Indeed, 6- and 8-month-old infants have been found to reveal enhanced discrimination of two categories (e.g., /d/ and /t/) after being exposed to tokens exemplifying a bimodal distribution along a distinguishing acoustic dimension, but not when exposed to a unimodal distribution (Maye *et al.*, 2002, 2008; see also Cristià *et al.*, 2011). Further evidence for the plausibility of distributional learning of phonetic category structure comes from analyses of infant-directed speech (IDS). Some analyses have shown that mothers appear to provide their infants with acoustic cues that could support distributional learning of phonetic categories (Werker *et al.*, 2007). In particular, IDS is characterized by expansion of the $F1$–$F2$ vowel formant space (Kuhl *et al.*, 1997), which could enhance the separability of vowel categories if this expansion is not compensated by increases in within-category variance.

The in-principle usefulness of distributional cues has been demonstrated in computer models of phonetic category learning. When categories are sufficiently separated in acoustic space, distributional learning models (often implemented as Gaussian Mixture Models) are able to learn category structure. Figure 1(a) shows an example of a set of categories that can be learned this way. The vowel formants of /i/, /ɑ/, and /u/ are clearly separated in acoustic space, and as expected, distributional models learn these categories with high accuracy (de Boer and Kuhl, 2003, see also Boersma *et al.*, 2003; McMurray *et al.*, 2009; Vallabha *et al.*, 2007). However, when considering distributional learning in a more realistic setting (e.g., when considering the full set of vowels that occur in a language), it becomes clear that phonetic categories are highly variable, and have overlapping distributions that pose a substantial problem for learning (Swingley, 2009). Figure 1(b) shows the problem of overlapping distributions, and illustrates that the detection of categories is far from trivial.

What is even more problematic for the distributional learning hypothesis is that several recent studies have argued

a)Electronic mail: f.w.adriaans@uu.nl

 0001-4966/2017/141(5)/3070/9/$30.00

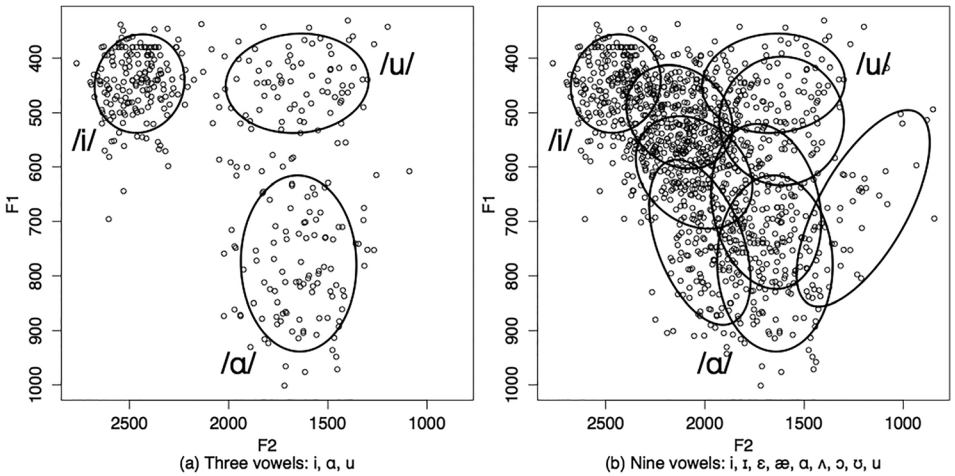

FIG. 1. Vowel distributions in IDS (based on data from Swingley, 2009).

that acoustic distributions in IDS are in fact *more* variable than distributions in adult-directed speech (ADS). In a comparison of IDS and ADS, Cristià and Seidl (2014) found that point vowels in IDS were indeed hyperarticulated, resulting in a stretching of the vowel space similar to the findings of Kuhl *et al.* (1997). However, the distance between specific contrasts (e.g., i-ɪ) was not enhanced. Moreover, the study found increased variability (larger within-category variance) in IDS as compared to ADS. Similarly, Martin *et al.* (2015) found that Japanese phonetic contrasts were less clear in IDS than in ADS. These findings question the relevance of IDS for language acquisition, and particularly question the hypothesis that speakers implicitly enhance contrasts to support phonological acquisition. Other studies have also questioned the benefit of acoustic distributions in IDS for language development (Benders, 2013; McMurray *et al.*, 2013).

Given the highly variable nature of IDS, how do infants manage to discover category structure? Perhaps infants' learning of categories is guided by additional sources of information. In particular, it has been proposed that the emerging lexicon might provide the infant with crucial guidance for the discovery of category structure (Swingley, 2009; Feldman *et al.*, 2013). Also the presence of neighboring segments might assist the infant in dealing with acoustically variable input (Dillon *et al.*, 2013).

The current study investigates another potentially useful source of information that might help infants in learning phonetic categories from IDS, namely the prosodic exaggeration parents use in IDS, speaking with a higher pitch register, wider pitch excursions, a limited set of intonation patterns, and greater duration of some sounds (Fernald and Simon, 1984; Fernald, 1985; Fernald *et al.*, 1989). It has been argued that the prosody that is typical of IDS serves to get the infant's attention and to express affection and other emotions to the child, a view supported by evidence that infants prefer prototypical IDS over ADS (Cooper and Aslin, 1990; Cooper *et al.*, 1997; Fernald, 1985; Pegg *et al.*, 1992). It has also been argued that IDS supports early language development (e.g., Werker *et al.*, 2007), possibly by enhancing infants' speech discrimination skills (Fernald and Kuhl, 1987; Karzon, 1985; Liu *et al.*, 2003; Trainor and Desjardins, 2002), or by facilitating word learning (Ma *et al.*, 2011; Thiessen *et al.*, 2005). For the acquisition of vowel categories there is evidence that exaggerated pitch contours might facilitate infants' vowel discrimination abilities (Trainor and Desjardins, 2002). However, given the large amount of variability found in vowel formant spaces in IDS, the role of IDS in language development is still debated (Cristià, 2013; Eaves *et al.*, 2016; Soderstrom, 2007), and it remains unclear whether the prosodic exaggeration that is typical of IDS is helpful or harmful for vowel category learning.

We hypothesize that prosodic exaggeration of the sort typical of IDS might be helpful for category learning, by guiding the infant's attention to a subset of relatively clear vowel tokens that improve distributional category learning. That is, while vowels of IDS are overall highly variable in their formants, the tokens that have exaggerated prosody might be relatively clear instances of their categories. If so, and if prosodic exaggeration leads infants to attend to these tokens, then distributional learning of phonetic categories should be enhanced. Previous computational models that aimed to explain category learning were typically fit to equally weighted vowel tokens (de Boer and Kuhl, 2003; Vallabha *et al.*, 2007). By overlooking prosodic properties that might make certain vowel tokens more influential than others, such models might therefore underestimate the learnability of vowel categories. We address this hypothesis by analyzing and comparing vowel distributions of tokens with different prosodic status (see also Mo *et al.*, 2009), and by simulating the distributional learning of phonetic categories. This allows us to examine whether prosodic focus could help infants discover category structure when the overall mass of instances exhibits extensive overlap.

For our analyses we created a large database of vowels taken from recordings of natural mother-infant interactions (Brent and Siskind, 2001). Most earlier studies have been based on vowel tokens that were elicited in a laboratory setting, and that occurred in a relatively small number of words or nonwords. For example, the classic study by Hillenbrand *et al.* (1995) covers the entire set of American English vowels and is based on a large number of speakers. However, vowels are produced in one particular phonological context (/hVd/) and only one token per vowel was analyzed for each speaker, thereby obscuring within-speaker variability. A

more recent study on learning vowel categories from IDS (Vallabha *et al.*, 2007) focused on four vowel categories (/i/, /ɪ/, /e/, /ɛ/) placed in minimal-pair nonce words and read to children from books; it is possible, indeed likely, that parents would hyperarticulate under such conditions. In what follows, we used a new vowel database that we developed based on audio recordings of IDS taken from the Brent corpus (Brent and Siskind, 2001). We present acoustic measurements based on ≈4400 vowel tokens encompassing the entire set of non-schwa monophthongs of American English. Because the vowel tokens were taken from recordings of natural (unscripted) mother-infant interactions, the analyses and simulations are thus based on realistic data, acknowledging the variability and complexities that are found in everyday speech. The database presents ecologically valid training materials for the current study as well as for future studies.[1]

II. VOWEL DATABASE

Vowel productions were examined for three different speakers in the Brent corpus (Brent and Siskind, 2001), available through CHILDES (MacWhinney, 2000). A total of four recording sessions of around 75 min each were selected for use in the current study (speaker "f1": sessions "f10jan97" and "m20jan97"; speaker "d1": session "m6jan97"; speaker "w1": session "f21jun96"). These sessions were selected based on recording quality. The ages of the infants at the time of recording range from 10 months and 3 days to 10 months and 26 days. The recordings consist of natural, unscripted IDS and therefore have no restrictions on the words or vowel types that may occur. The resulting data set contains a total of 4435 tokens covering the monophthongs of American English (/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /ɔ/, /ʊ/, /u/, see Table I).

Several acoustic measurements were obtained through a combination of automatic and manual procedures. Vowel formants (*F*1, *F*2) were measured at midpoint. A first-pass measurement was done automatically using Praat (Boersma and Weenink, 2012), with optimized settings for each speaker. Tokens that were more than 1.5 standard deviations away from their vowel category's mean (on either the *F*1 or *F*2 dimension, around 25% of the total dataset) were manually checked by a phonetically trained research assistant who corrected the formants if necessary based on spectrographic analysis. Each vowel token's duration was measured in milliseconds. In order to assess the amount of prosodic exaggeration in each vowel, several pitch measures were obtained:

- Mean *F*0, measured between 20% and 80% of the vowel's duration;
- Minimum *F*0, measured between 20% and 80% of the vowel's duration;
- Maximum *F*0, measured between 20% and 80% of the vowel's duration;
- *F*0 movement, which was calculated as the difference between the minimum and maximum *F*0.

TABLE I. Frequency of occurrence of vowel categories per speaker.

Speaker	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ʌ/	/ɔ/	/ʊ/	/u/	Total
*f*1	386	560	279	229	222	211	49	68	170	2174
*d*1	171	211	75	90	109	123	46	68	149	1042
*w*1	244	228	141	119	75	142	80	57	133	1219
Total	801	999	495	438	406	476	175	193	452	4435

To minimize the potential effects of measurement or labeling errors that may have remained after hand-checking the data, tokens that were more than 2.5 standard deviations away from a vowel category's *F*1 or *F*2 mean for a particular speaker were labeled as outliers, and were removed from the data set. This criterion removed 3.4% of all data points (leaving a total of 4435 tokens in the final database).

In addition, a phonetically trained research assistant listened to each utterance in the database and identified prosodically exaggerated parts of the utterance. Specifically, for each vowel in the utterance a label was added indicating whether the vowel occurred in a syllable that the mother was judged to be emphasizing. This resulted in 1041 tokens (23.5%) being labeled as having "focus" (i.e., the token was prosodically exaggerated) and 3394 tokens (76.5%) with the label "no focus" (i.e., the token was not prosodically exaggerated). This allowed us to select vowel tokens based on their prosodic status, and then characterize the formant distributions of prosodically exaggerated and non-exaggerated instances. We then trained and tested distributional learning models on the prosodically exaggerated and non-exaggerated subsets, addressing the issue of whether or not it is easier to cluster tokens that show exaggerated prosody. Finally, we quantified and tested the degree of prosodic exaggeration in the focus tokens along several acoustic prosodic dimensions.

III. ANALYSES OF THE VOWEL SPACE

As noted above, measurements of IDS have found it to be hyperarticulated relative to ADS, in the sense that the point vowels are more distant from one another in IDS than ADS. Here, following Kuhl *et al.* (1997), we analyzed the area described by the triangle whose vertices are the mean *F*1 and *F*2 values of the vowels /i/, /ɑ/, and /u/. Figure 2 shows mean formant values for vowels with prosodic focus and without prosodic focus for each individual speaker in our database. For all three speakers, the triangle formed by the prosodically exaggerated vowels is larger than the triangle formed by unexaggerated vowels.

There were some individual differences in the phonetic direction of hyperarticulation. For example, while speaker *d*1 showed systematic expansion in both *F*1 and *F*2 in all three corners of the vowel triangle, speaker *w*1's high vowels showed expansion only in *F*2. To get a better picture of the overall degree of expansion, each speaker's tokens were converted to *z*-scores and then averaged. Figure 3 shows the averaged vowel space across different speakers. The area of the triangle formed by point vowels with prosodic focus is 57% larger than the area of the triangle formed by vowels without prosodic focus, confirming that point vowels in focused position are systematically hyperarticulated as compared to vowels in unfocused position.

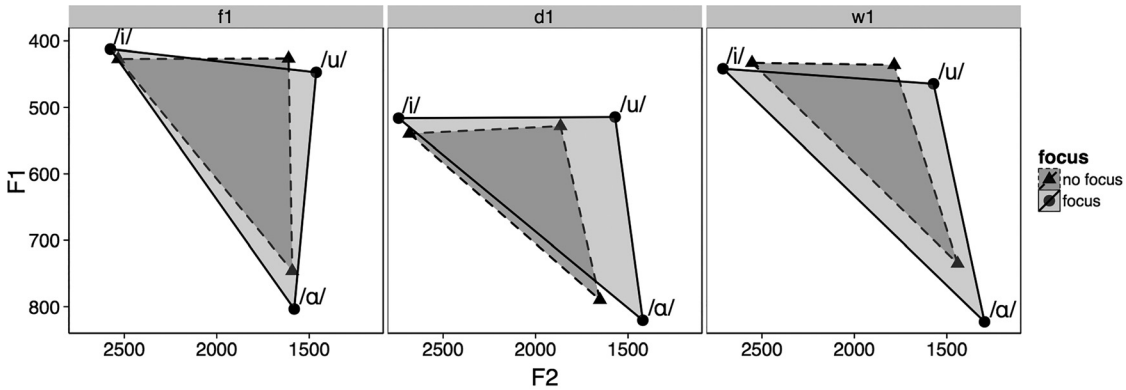

FIG. 2. Point vowels for each speaker (*f*1, *d*1, *w*1) in focused (focus) and unfocused (no focus) position.

What about non-point vowels? As can be seen in Table I, point vowel tokens make up a relatively small percentage of the total set of vowels found in everyday infant-directed English (around 37% in our set of nine monophthongs). When considering consequences for vowel learnability, it is therefore important to consider the entire vowel space more broadly. Figure 4 illustrates that hyperarticulation was not restricted to point vowels. To assess the degree of hyperarticulation across the set of monophthongs, the Euclidean distance between each vowel's mean and the center of the vowel space was calculated (e.g., Bradlow *et al.*, 1996). Table II shows that for most vowels, the prosodically focused vowels' means were further away from the center, indicating a general tendency toward hyperarticulation in focused vowels across the vowel space.

Another way to assess the degree of hyperarticulation is by measuring distances between vowel categories that are adjacent in acoustic space, such as /i/ and /ɪ/. Table III shows the Euclidean distances between adjacent categories in focused and unfocused position. The distance between adjacent categories is consistently larger in focused position, across the entire vowel space. The only exception is the distance between vowels /ɔ/ and /ʊ/, which is 9.3% smaller in focused position. However, these categories are relatively distant to begin with, and their occurrence in our database is relatively rare (see Table I).

Interestingly, while /ɪ/ is closer to the center of the vowel space in focused position, and thus appears not to be hyperarticulated, it is in fact at a larger distance from its immediate neighbors /i/ and /ɛ/. This is an important finding because Cristià and Seidl (2014) found that the distance between neighboring categories (such as /i/ and /ɪ/) was decreased in IDS as compared to ADS, a potentially puzzling finding if one entertains the hypothesis that IDS aids language learning by making the signal clearer. Though we do not measure ADS here, it is possible that IDS is not consistently clearer over all tokens, but does present an advantageous learning signal if one considers the most prominent vowel tokens.

Of course, from the perspective of distributional category learning, increased separation of the category means could be of no use if this separation were accompanied by a concomitant increase in variability. We therefore evaluate the issue of category learnability (which depends on both between-category distance and within-category variability) more directly by simulating the distributional learning of the vowel categories in our dataset. In particular, we examined whether prosodic focus could help in discovering category structure in cases of large overlap between categories. If distributional models of vowel learning show improved performance when trained on prosodically defined subsets of vowel data, then this would constitute evidence that the

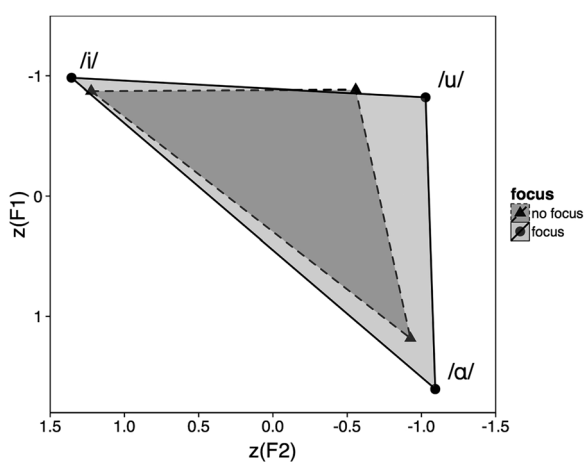

FIG. 3. Point vowels pooled across speakers. Formants were normalized to *z*-scores before averaging.

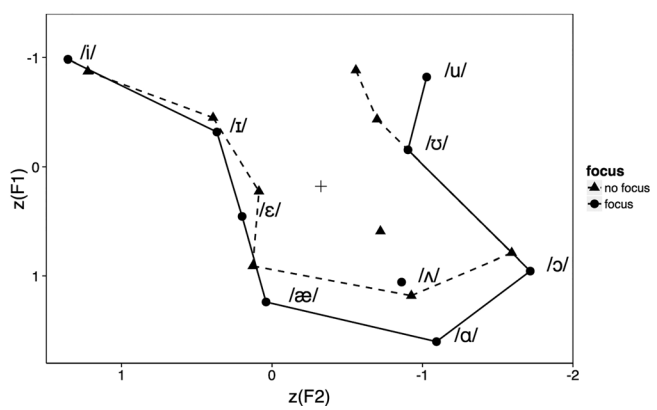

FIG. 4. The entire set of monophthongs. The center of the vowel space is indicated as "+." Formants were normalized to *z*-scores before averaging across speakers.

TABLE II. Euclidean distance d from the center of the vowel space c to the category means of vowels without prosodic focus (v_{nofoc}) and vowels with prosodic focus (v_{foc}). Calculations were based on z-transformed $F1$ and $F2$ measurements.

Vowel	$d(c, v_{nofoc})$	$d(c, v_{foc})$	% change
/i/	1.87	2.04	+9.1%
/ɪ/	0.95	0.85	−10.7%
/ɛ/	0.41	0.59	+43.1%
/æ/	0.86	1.12	+30.7%
/ɑ/	1.17	1.62	+38.5%
/ʌ/	0.57	1.03	+79.7%
/ɔ/	1.41	1.59	+13.2%
/ʊ/	0.72	0.67	−7.0%
/u/	1.09	1.22	+12.3%

prosodic exaggeration in IDS can support phonetic category learning. It is important to note that the point of these simulations is not to provide the most realistic learning model. Such a model would need additional cues (Swingley, 2009; Feldman *et al.*, 2013). Rather, we use computational modeling to test different input representations, by quantifying the benefits of prosodic highlighting for learning vowel categories based on formant distributions.

IV. COMPUTER SIMULATIONS

The learnability of vowel categories was examined in two different simulations, each involving three vowel categories. One simulation assessed the learnability of the point vowels /i/, /ɑ/, /u/, which are relatively distant from one another in the $F1$–$F2$ formant space. The other simulation had three vowels that are close in formant space: /i/, /ɪ/, /ɛ/. These front vowels were selected because they were more numerous in the dataset than other comparable adjacent groups such as the back vowels. In line with earlier work on computational modeling of phonetic category learning (e.g., de Boer and Kuhl, 2003; McMurray *et al.*, 2009; Vallabha *et al.*, 2007), we treated categories as multivariate Gaussian distributions. The learning problem was characterized as estimating the parameters (means, covariances, and mixing proportions) for these distributions. Categories were defined as two-dimensional distributions (the z-scores of the first and second formants). Data points were assigned to the category that had the maximum likelihood for that point. Parameters of the Gaussian distributions were estimated using the EM algorithm (Dempster *et al.*, 1977) as implemented in the *MCLUST for R* software package (Fraley and Raftery, 2006). All models reported below were trained to discover three ellipse categories. Since vowel ellipses are known to vary in volume, shape, and orientation (e.g., Hillenbrand *et al.*, 1995), the models were given no information or constraints with respect to volume, shape, or orientation.

For each simulation, three different models were trained, which were each based on a different input representation. A baseline model was trained on the entire set of tokens for the three categories under inspection (the "ALL TOKENS" set). A second model was trained on a subset of tokens, namely those tokens that were labeled as having prosodic focus (the "FOCUS" set). Finally, a third model was trained on the complementary subset of tokens without prosodic focus, as a control (the "NO FOCUS" set). In order to balance the number of tokens taken from each vowel category (i.e., we want to focus on qualitative differences, not on quantitative differences), 2000 training points were sampled from each vowel category's multivariate normal distribution in the appropriate subset (e.g., Vallabha *et al.*, 2007). After clustering, each model's classification performance was tested using 2000 newly sampled data points. One hundred repeated runs were done for each model. We predicted that Gaussian mixture models trained on prosodically prominent vowel tokens (the FOCUS set) would provide a better classification of the data than Gaussian mixture models that were trained on the complete set of vowel tokens (ALL TOKENS), or models that were trained on tokens without prosodic focus (NO FOCUS).

Table IV shows the mean classification accuracies [and 95% confidence intervals (CIs), which were calculated using arcsine-square-root transformations] in the /i/-/ɑ/-/u/ and /i/-/ɪ/-/ɛ/ simulations. The classification accuracies confirm the difference in between-category distances in the two simulations, with overall classification scores being much lower for /i/-/ɪ/-/ɛ/ (as expected). Importantly, the model trained on the FOCUS set outperformed the ALL TOKENS and NO FOCUS sets in both simulations. The FOCUS training set thus presents the learner with category distributions that show less overlap in the $F1$, $F2$ space (see Fig. 5). Note that the superior performance on the FOCUS set was not due to the smaller number of tokens that generated the distribution. This explanation is ruled out by the NO FOCUS set, which performs worse than the ALL TOKENS baseline, even though its distribution is based on fewer tokens.

TABLE III. Euclidean distance d between adjacent category means for vowels without prosodic focus (v_{nofoc}) and vowels with prosodic focus (v_{foc}). Calculations were based on z-transformed $F1$ and $F2$ measurements.

Vowel pair	$d(v1_{nofoc}, v2_{nofoc})$	$d(v1_{foc}, v2_{foc})$	% change
/i/ - /ɪ/	0.93	1.19	+27.7%
/ɪ/ - /ɛ/	0.74	0.79	+6.9%
/ɛ/ - /æ/	0.68	0.80	+16.9%
/æ/ - /ɑ/	1.09	1.19	+9.4%
/ɑ/ - /ɔ/	0.77	0.90	+16.2%
/ɔ/ - /ʊ/	1.52	1.37	−9.3%
/ʊ/ - /u/	0.47	0.68	+44.0%

TABLE IV. Classification accuracies for point vowels i-ɑ-u and front vowels i-ɪ-ɛ. The displayed scores are the means obtained through 100 repeated runs, along with the 95% CI.

	i-ɑ-u			i-ɪ-ɛ		
		95% CI			95% CI	
Model	Mean	Lower	Upper	Mean	Lower	Upper
ALL TOKENS	0.9148	0.9143	0.9153	0.6556	0.6490	0.6630
NO FOCUS	0.8993	0.8988	0.8999	0.6483	0.6418	0.6555
FOCUS	0.9673	0.9670	0.9676	0.7139	0.7060	0.7236

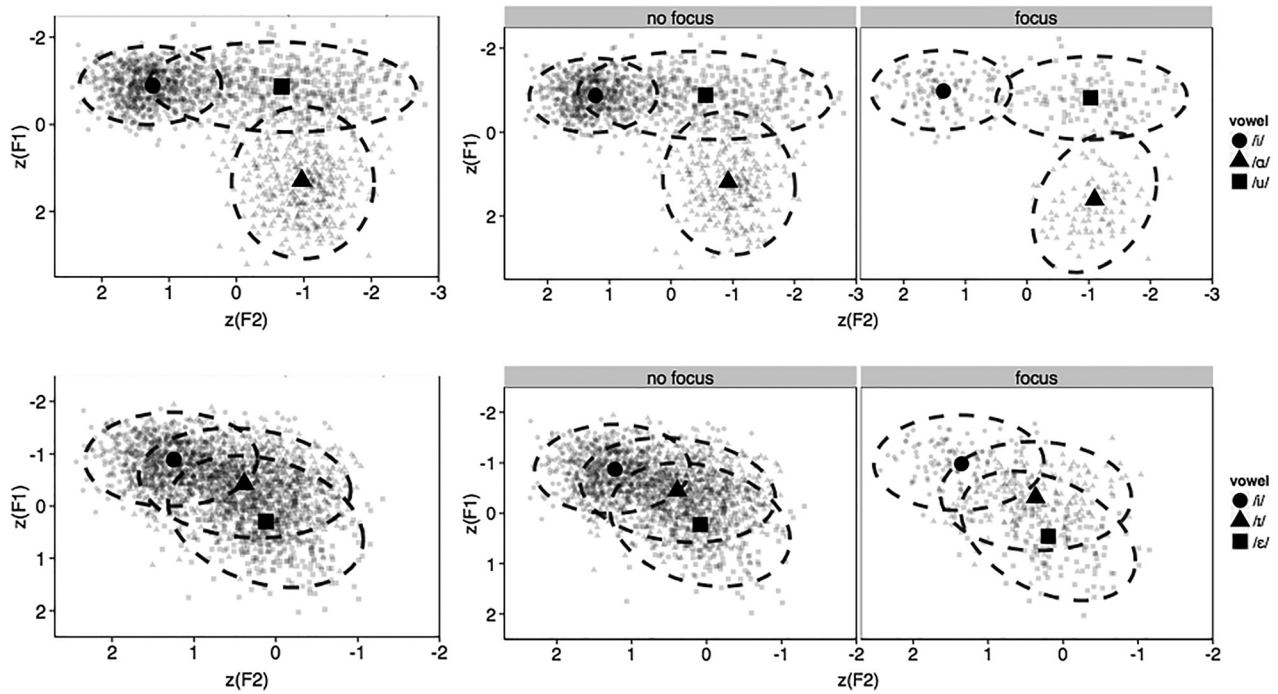

FIG. 5. Distributions of ALL TOKENS (left), NO FOCUS tokens (middle), and FOCUS tokens (right) for the /i/-/ɑ/-/u/ data set (top) and the /i/-/ɪ/-/ɛ/ data set (bottom).

Earlier studies have pointed out the importance of examining whether or not specific contrasts (such as /i/-/ɪ/) are enhanced (Cristià and Seidl, 2014). We therefore also examine the effects of prosodic exaggeration on a series of four specific vowel contrasts that lie between point vowels /i/ and /ɑ/: /i/-/ɪ/, /ɪ/-/ɛ/, /ɛ/-/æ/, and /æ/-/ɑ/. If there is a positive effect on the distributional learning of these category pairs, then this would provide additional support for our hypothesis that prosody in IDS supports distributional category learning.

Table V shows that all four contrasts were indeed easier to cluster in the FOCUS set, with the biggest increase in classification accuracy for the vowel pairs /i/-/ɪ/ and /æ/-/ɑ/. While the increase in classification accuracy was relatively small for /ɪ/-/ɛ/ and /ɛ/-/æ/, the learnability advantage was consistent across the four vowel pairs, providing additional evidence for a beneficial role for prosody in distributional category learning.

V. ACOUSTIC CORRELATES OF PROSODIC FOCUS

Tokens that were labeled as having prosodic focus were tokens that occurred in syllables that sounded (to our annotators) like they were being emphasized by the speaker. To what extent did these emphasized tokens stand out acoustically? In order to evaluate how consistently these vowels were marked with specific acoustic features, we compared focused and not-focused vowels along the acoustic dimensions of pitch ($F0$), pitch movement ($\Delta F0$), and duration.

Figure 6 shows the average mean $F0$, $F0$ movement, and duration for each speaker in our vowel database for vowels in focused ("foc") and unfocused ("nofoc") position. Across all nine vowel categories, all three speakers produced focused vowels with a higher $F0$, greater $F0$ movement, and longer duration. On average, vowels with focus were about 40 Hz higher than vowels without focus. Also, focused vowels showed 30 Hz more change in $F0$ throughout the vowel's duration. Finally, vowels with focus were on average 70 ms longer than vowels without focus. Linear regression analyses using "Speaker" (3 levels),[2] "Vowel" (9 levels), and "Focus" (2 levels) as predictors confirmed that the prosodic exaggeration was significant along each of the three dimensions. (See the appendix for the full analyses.) These results confirm that the focused tokens were, on the whole, exaggerated along dimensions that infants might attend to preferentially (e.g., Fernald and Kuhl, 1987).

TABLE V. Classification accuracies for adjacent vowel pairs. The displayed scores are the means obtained through 100 repeated runs, along with the 95% CI.

	i-ɪ			ɪ-ɛ			ɛ-æ			æ-ɑ		
		95% CI			95% CI			95% CI			95% CI	
Model	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper
ALL TOKENS	0.7885	0.7831	0.7953	0.7018	0.6933	0.7123	0.6801	0.6742	0.6867	0.7955	0.7901	0.8025
NO FOCUS	0.7766	0.7708	0.7838	0.6971	0.6894	0.7063	0.6780	0.6708	0.6863	0.7878	0.7829	0.7939
FOCUS	0.8669	0.8651	0.8690	0.7214	0.7144	0.7300	0.6955	0.6914	0.6999	0.8636	0.8618	0.8656

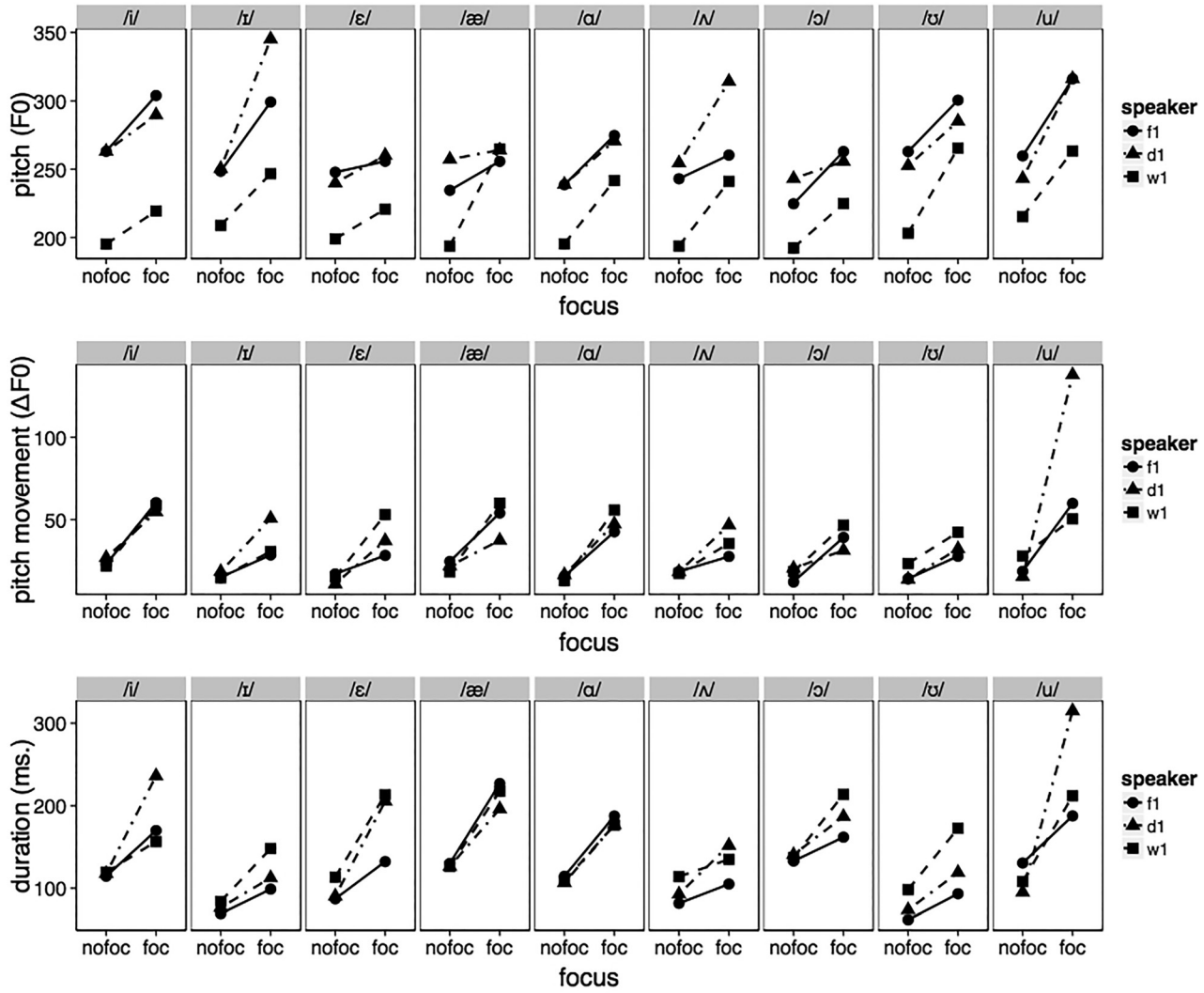

FIG. 6. Acoustic correlates of prosodic focus (from top to bottom: pitch, pitch movement, duration).

VI. GENERAL DISCUSSION

Recently there has been debate about whether or not IDS has properties that would support language acquisition (e.g., Eaves *et al.*, 2016). On one hand, IDS appears to be hyperarticulated and to have certain attention-grabbing prosodic properties. On the other hand, IDS appears to be highly variable, and in fact hyperarticulation of point vowels might obscure contrasts with non-point vowels (Cristià and Seidl, 2014). In the current study we aimed to investigate whether one beneficial property of IDS (prosodic exaggeration) might help overcome a problematic property (overlap between categories). In analyses of the vowel space, as well as in a series of distributional learning simulations, we found that prosodic exaggeration in IDS has the potential to support distributional vowel learning by presenting the learner with a subset of "high-quality" tokens that the infant is likely to pay attention to. Models trained on tokens that showed prosodic exaggeration outperformed models that were trained on tokens that did not show prosodic exaggeration.

The models we used to simulate phonetic category learning in the current study assessed the separability of categories based on formant distributions. While this gives insight into properties of the input that could support learning, these models are fairly simplistic mixture models, and are not meant to represent a realistic model of phonetic category learning. The models were trained on particular subsets of categories in the vowel space, and were given the number of categories that needed to be found. The models did not attempt to learn the full set of categories, which would introduce the problem of estimating not just the shapes, orientations, and sizes of each category, but also estimating the number of vowel categories in the language. A more realistic model of phonetic category learning would also need to take into account other factors beyond formant distributions and prosody. Specifically, category learning might need crucial guidance from the infant's emerging lexicon (Swingley, 2009; Feldman *et al.*, 2013). Prosody could help such a model by providing the learner with a better bottom-up cue to category structure.

Several studies have argued against the view that hyperarticulation in IDS would be the product of didactic intent from the parent (Benders, 2013; McMurray *et al.*, 2013; but see Eaves *et al.*, 2016). The current study does not address underlying motivations or intentions from the part of the speaker. Rather, what we find is that prosodic exaggeration within IDS appears to be accompanied by relatively careful,

hyperarticulated speech. This suggests that mothers might simultaneously get the infant's attention (using prosody) and provide the infant with a clearer signal (supporting category learning). Whether this behavior is intentional, or is even true for languages other than American English, remains to be seen.

While the total number of vowels (both types and tokens) used in the current study is larger than is typically used in earlier studies, our database was based on speech from a relatively small number of mothers. The vowel space analyses were therefore completed separately for each individual speaker (Figs. 2 and 6). These three speakers showed similar patterns, suggesting that mothers emphasize vowels in similar ways. While it remains an open question whether this is true for most mothers (or in other languages), our study provides a detailed picture of within-speaker variability across the vowel space, in everyday mother-infant interactions, and thereby provides ecologically valid training materials for computational models of phonetic category learning.

Infants' successful perception of native phonetic categories is positively correlated with vocabulary size (Cristia *et al.*, 2014; Kuhl *et al.*, 2005; Tsao *et al.*, 2004; Yeung *et al.*, 2014), as is skill in spoken-word recognition (Kemp *et al.*, 2017; Marchman *et al.*, 2016), indicating the likely relevance of early perceptual learning for language development in early childhood. The substantial experimental literature on children's early linguistic trajectories has not been matched by a thorough characterization of children's language environments, and as a result tracing children's typical developmental pathways quantitatively has been difficult. The present dataset is part of a broader effort at quantitative modeling of early phonological and lexical development (e.g., Cristià, 2013). Our analyses support the possibility that attention to prosodically salient instances of vowels helps infants solve what might appear to be an insurmountable computational problem.

ACKNOWLEDGMENTS

This work was funded by the Netherlands Organisation for Scientific Research (NWO) Grant No. 446.010.027 to F.A. and NIH Grant No. R01-HD049681 to D.S. Part of the research reported here was presented at the International Child Phonology Conference and the 34th Annual Conference of the Cognitive Science Society in 2012. We would like to thank former members of the Penn Infant Language Center for their assistance, especially Ashley Baldwin, Allison Britt, Joe Fruehwald, and Becky Mead. We also thank Zachary Jaggers and James Whang for their assistance with annotating the data, and we would like to thank two anonymous reviewers for helpful comments and suggestions.

APPENDIX: REGRESSION ANALYSES

The following three linear regression analyses test whether there were acoustic differences (along the dimensions of pitch, pitch movement, and duration, respectively) between our sets of focused and unfocused vowels, taking into account other factors that may affect these dimensions. The analyses take into account the effects of Speaker, Vowel, and Focus, the latter being our variable of interest.

1. Acoustic correlates of focus: pitch ($F0$ in Hz). Intercept: Speaker = $d1$, Vowel = /i/, Focus = NO FOCUS.

Coefficient	*Estimate*	*Std. Error*	Pr(> |z|)
Intercept	259.95	3.34	<0.001***
Speaker			
$f1$	−4.72	2.83	0.0953
$w1$	−50.76	3.13	<0.001***
Vowel			
/ɪ/	−2.86	3.51	0.4141
/ɛ/	−18.07	4.24	<0.001***
/æ/	−18.44	4.42	<0.001***
/ɑ/	−17.84	4.51	<0.001***
/ʌ/	−12.78	4.28	0.0028**
/ɔ/	−23.17	6.20	<0.001***
/ʊ/	−0.18	5.93	0.9763
/u/	3.30	4.36	0.4496
Focus	39.18	2.67	<0.001***

2. Acoustic correlates of focus: pitch movement ($\Delta F0$ in Hz). Intercept: Speaker = $d1$, Vowel = /i/, Focus = NO FOCUS.

Coefficient	*Estimate*	*Std. Error*	Pr(> |z|)
Intercept	27.86	1.67	<0.001***
Speaker			
$f1$	−4.84	1.41	<0.001***
$w1$	−3.08	1.56	0.0485*
Vowel			
/ɪ/	−10.72	1.75	<0.001***
/ɛ/	−11.36	2.11	<0.001***
/æ/	−1.81	2.20	0.4101
/ɑ/	−9.06	2.25	<0.001***
/ʌ/	−10.21	2.13	<0.001***
/ɔ/	−9.46	3.09	0.0022**
/ʊ/	−11.17	2.96	<0.001***
/u/	4.75	2.17	0.0288*
Focus	29.53	1.33	<0.001***

3. Acoustic correlates of focus: duration (ms). Intercept: Speaker = $d1$, Vowel = /i/, Focus = NO FOCUS.

Coefficient	*Estimate*	*Std. Error*	Pr(> |z|)
Intercept	117.75	4.10	<0.001***
Speaker			
$f1$	−9.02	3.47	0.0093**
$w1$	5.85	3.84	0.1277
Vowel			
/ɪ/	−44.92	4.30	<0.001***
/ɛ/	−18.43	5.20	<0.001***
/æ/	22.81	5.42	<0.001***
/ɑ/	−1.22	5.54	0.8249
/ʌ/	−30.49	5.25	<0.001***
/ɔ/	13.82	7.60	0.0691
/ʊ/	−42.75	7.28	<0.001***
/u/	11.98	5.35	0.0251*
Focus	66.62	3.27	<0.001***

[1]The database will be made publicly available on the authors' websites.
[2]"Speaker" is included as a fixed factor, because in our dataset it has only three levels, which is insufficient to treat it as a random factor.

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

Boersma, P., Escudero, P., and Hayes, R. (**2003**). "Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, pp. 1013–1016.

Boersma, P., and Weenink, D. (**2012**). "Praat: Doing phonetics by computer" (Computer program).

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Brent, M. R., and Siskind, J. M. (**2001**). "The role of exposure to isolated words in early vocabulary development," Cognition **81**, B33–B44.

Cooper, R. P., Abraham, J., Berman, S., and Staska, M. (**1997**). "The development of infants' preference for motherese," Infant Behav. Develop. **20**, 477–488.

Cooper, R. P., and Aslin, R. N. (**1990**). "Preference for infant-directed speech in the first month after birth," Child Develop. **61**, 1584–1595s.

Cristià, A. (**2013**). "Input to language: The phonetics and perception of infant-directed speech," Lang. Linguist. Compass **7**, 157–170.

Cristià, A., McGuire, G. L., Seidl, A., and Francis, A. L. (**2011**). "Effects of the distribution of acoustic cues on infants' perception of sibilants," J. Phonetics **39**, 388–402.

Cristià, A., and Seidl, A. (**2014**). "The hyperarticulation hypothesis of infant-directed speech," J. Child Lang. **41**, 913–934.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., and Hagoort, P. (**2014**). "Predicting individual variation in language from infant speech perception measures," Child Develop. **85**, 1330–1345.

de Boer, B., and Kuhl, P. K. (**2003**). "Investigating the role of infant-directed speech with a computer model," Acoust. Res. Lett. Online **4**, 129–134.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Stat. Soc. B **39**, 1–38, see http://www.jstor.org/stable/2984875.

Dillon, B., Dunbar, E., and Idsardi, W. (**2013**). "A single-stage approach to learning phonological categories: Insights from Inuktitut," Cognitive Sci. **37**, 344–377.

Eaves, B. S., Jr., Feldman, N. H., Griffiths, T. L., and Shafto, P. (**2016**). "Infant-directed speech is consistent with teaching," Psychol. Rev. **123**, 758–771.

Feldman, N. H., Griffiths, T. L., Goldwater, S., and Morgan, J. L. (**2013**). "A role for the developing lexicon in phonetic category acquisition," Psychol. Rev. **120**, 751–778.

Fernald, A. (**1985**). "Four-month-old infants prefer to listen to motherese," Infant Behav. Develop. **8**, 181–195.

Fernald, A., and Kuhl, P. (**1987**). "Acoustic determinants of infant preference for motherese speech," Infant Behav. Develop. **10**, 279–293.

Fernald, A., and Simon, T. (**1984**). "Expanded intonation contours in mothers' speech to newborns," Develop. Psychol. **20**, 104–113.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., and Fukui, I. (**1989**). "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," J. Child Lang. **16**, 477–501.

Fraley, C., and Raftery, A. E. (**2006**). *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*, Technical Report (University of Washington, Seattle, WA).

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Karzon, R. G. (**1985**). "Discrimination of polysyllabic sequences by one- to four-month-old infants," J. Exp. Child Psychol. **39**, 326–342.

Kemp, N., Scott, J., Bernhardt, B. M., Johnson, C. E., Siegel, L. S., and Werker, J. F. (**2017**). "Minimal pair word learning and vocabulary size: Links with later language skills," Appl. Psycholing. **38**, 289.

Kuhl, P., Jean, K., Andruski, E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (**1997**). "Cross-language analysis of phonetic units in language addressed to infants," Science **277**, 684–686.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., and Pruitt, J. (**2005**). "Early speech perception and later language development: Implications for the 'critical period'," Lang. Learn. Develop. **1**, 237–264.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (**1992**). "Linguistic experience alters phonetic perception in infants by 6 months of age," Science **255**, 606–608.

Liu, H.-M., Kuhl, P. K., and Tsao, F.-M. (**2003**). "An association between mothers' speech clarity and infants' speech discrimination skills," Develop. Sci. **6**, F1–F10.

Ma, W., Golinkoff, R. M., Houston, D. M., and Hirsh-Pasek, K. (**2011**). "Word learning in infant- and adult-directed speech," Lang. Learn. Develop. **7**, 185–201.

MacWhinney, B. (**2000**). *The CHILDES Project: Tools for Analyzing Talk*, Vol. 2: The database, 3rd ed. (Lawrence Erlbaum Associates, Mahwah, NJ).

Marchman, V. A., Adams, K. A., Loi, E. C., Fernald, A., and Feldman, H. M. (**2016**). "Early language processing efficiency predicts later receptive vocabulary outcomes in children born preterm," Child Neuropsychol. **22**, 649–665.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., and Cristià, A. (**2015**). "Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis," Psychol. Sci. **26**, 341–347.

Maye, J., Weiss, D. J., and Aslin, R. N. (**2008**). "Statistical phonetic learning in infants: Facilitation and feature generalization," Develop. Sci. **11**, 122–134.

Maye, J., Werker, J. F., and Gerken, L. A. (**2002**). "Infant sensitivity to distributional information can affect phonetic discrimination," Cognition **82**, B101–B111.

McMurray, B., Aslin, R. N., and Toscano, J. C. (**2009**). "Statistical learning of phonetic categories: Insights from a computational approach," Develop. Sci. **12**, 369–378.

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

Mo, Y., Cole, J., and Hasegawa-Johnson, M. (**2009**). "Prosodic effects on vowel production: Evidence from formant structure," in *Proceedings of Interspeech 2009*, pp. 2535–2538.

Pegg, J. E., Werker, J. F., and McLeod, P. J. (**1992**). "Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants," Infant Behav. Develop. **15**, 325–345.

Polka, L., and Werker, J. F. (**1994**). "Developmental changes in perception of nonnative vowel contrasts," J. Exp. Psychol. **20**, 421–435.

Soderstrom, M. (**2007**). "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants," Develop. Rev. **27**, 501–532.

Swingley, D. (**2009**). "Contributions of infant word learning to language development," Phil. Trans. Royal. Soc. B **364**, 3617–3622.

Thiessen, E. D., Hill, E. A., and Saffran, J. R. (**2005**). "Infant-directed speech facilitates word segmentation," Infancy **7**, 53–71.

Trainor, L. J., and Desjardins, R. N. (**2002**). "Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels," Psychonomic Bull. Rev. **9**, 335–340.

Tsao, F.-M., Liu, H.-M., and Kuhl, P. K. (**2004**). "Speech perception in infancy predicts language development in the second year of life: A longitudinal study," Child Develop. **75**, 1067–1084.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. (**2007**). "Unsupervised learning of vowel categories from infant-directed speech," Proc. Natl. Acad. Sci. U.S.A. **104**, 13273–13278.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., and Amano, S. (**2007**). "Infant-directed speech supports phonetic category learning in English and Japanese," Cognition **103**, 147–162.

Werker, J. F., and Tees, R. C. (**1984**). "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," Infant Behav. Develop. **7**, 49–63.

Yeung, H. H., Chen, L. M., and Werker, J. F. (**2014**). "Referential labeling can facilitate phonetic learning in infancy," Child Develop. **85**, 1036–1049.