Effects of consonantal context on the learnability of vowel categories from infant-directed speech

Frans Adriaans

Citation: The Journal of the Acoustical Society of America **144**, EL20 (2018); doi: 10.1121/1.5045192
View online: https://doi.org/10.1121/1.5045192
View Table of Contents: https://asa.scitation.org/toc/jas/144/1
Published by the Acoustical Society of America

Effects of consonantal context on the learnability of vowel categories from infant-directed speech

Frans Adriaans
Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, the Netherlands
f.w.adriaans@uu.nl

Abstract: Recent studies have shown that vowels in infant-directed speech (IDS) are characterized by highly variable formant distributions. The current study investigates whether vowel variability is partially due to consonantal context, and explores whether consonantal context could support the learning of vowel categories from IDS. A computational model is presented which selects contexts based on frequency in the input and generalizes across contextual categories. Improved categorization performance was found on a vowel contrast in American–English IDS. The findings support a view in which the infant's learning mechanism is anchored in context, in order to cope with acoustic variability in the input.

[DDO]
Date Received: April 3, 2018 **Date Accepted:** June 18, 2018

1. Introduction

During the first year of life infants begin to discover the vowel categories of their native language (e.g., Kuhl *et al.*, 1992; Polka and Werker, 1994). Infants initially have the ability to discriminate categories along both native and non-native phonetic contrasts (Eimas *et al.*, 1971; Trehub, 1976), and gradually become more attuned to the contrasts that they experience in their native language (Polka and Werker, 1994; Werker and Tees, 1984). By the age of 6 months infants are more sensitive to prototypical instances of vowel categories from their native language than to prototypes of non-native vowel categories, suggesting that experience with the native language is affecting infants' phonetic perception (Kuhl *et al.*, 1992). The process of phonetic adaptation continues to develop between 6 and 12 months, a period during which infants show further decline in non-native vowel discrimination (Polka and Werker, 1994).

The early age at which language-specific knowledge of vowels begins to manifest itself is puzzling in light of the phonetic properties of vowels in infant-directed speech (IDS). Swingley (2009) showed that vowels in IDS are characterized by highly variable formant distributions, and that vowel category structure is obscured by the large degree of overlap between categories. The lack of clear category boundaries indicates that there is no straightforward way for the infant to group together tokens based on their formants. Moreover, a growing number of studies have demonstrated that IDS provides the infant with a relatively complex acoustic environment, as compared to adult-directed speech (ADS). In various languages evidence has been found that IDS is characterized by decreased between-category distance and increased variability (e.g., Cristià and Seidl, 2014; Miyazawa *et al.*, 2017).

Given the nature of their acoustic input, how are infants able to acquire vowel categories? The leading view has been that infants use a distributional learning mechanism that groups together tokens based on similarity along acoustic dimensions (e.g., Maye *et al.*, 2002). While some analyses of IDS support this possibility (e.g., Werker *et al.*, 2007), the large amount of variability reported in recent IDS studies has led to a growing consensus that category learning cannot be solved by relying exclusively on acoustic clustering (Adriaans and Swingley, 2017; Dillon *et al.*, 2013; Feldman *et al.*, 2013; Swingley, 2009). It is therefore important to identify and assess alternative means through which infants might be able to detect category structure in IDS.

Computational models have been used to explore sources of contextual information which might guide the distributional learning of phonetic categories. One such source might be the infant's emerging lexicon. Infants around the age of 6–9 months have already acquired knowledge of some common words (Bergelson and Swingley, 2012; Tincoff and Jusczyk, 1999), and it is possible that the ability to recognize words in the input might help the infant to impose structure on the acoustic space

(Feldman *et al.*, 2013; Swingley, 2009). In addition, distributional learning could be facilitated by prosody. Adriaans and Swingley (2017) investigated whether the exaggerated prosody that is typical of IDS (e.g., high pitch, large intonation contours) might guide the infant to a subset of tokens that are relatively hyperarticulated, and thus easier to cluster as categories. It was found that models trained on tokens that showed prosodic exaggeration outperformed models that did not take prosodic status into account. Prosody thus has the potential to support distributional learning from highly variable acoustic input.

In addition to lexical and prosodic factors, phonetic category learning might be supported by phonological context. It has long been known that surrounding consonants affect vowel formants in adult speech (e.g., Hillenbrand *et al.*, 2001). A study by Dillon *et al.* (2013) found that the learnability of vowel categories in Inuktitut is affected by consonantal context. The three vowels of Inuktitut (/i/, /a/, /u/) are systematically lowered before uvular consonants. In a series of simulations Dillon *et al.* (2013) found that a model that links phonetic subcategories via a contextual phonological rule outperformed a Gaussian Mixture Model (GMM) without context. The model was tested on ADS in a language with a relatively simple vowel system, and the model employed one single phonological context ("uvular" or "other"). Therefore, fundamental questions regarding the role of context in category learning remain. For example, how could an infant discover relevant phonological contexts from input data? To what extent does consonantal context help the infant navigate through the large amounts of variability found in IDS?

The current study investigates whether consonantal context could contribute to the learning of vowel categories from IDS. An analysis of natural American–English IDS explores whether surrounding consonants affect the formant structure of vowels. A computational model is then presented, which is used to assess the effects of consonantal context on the learnability of vowel categories in IDS. The study contributes to the identification of sources of variability found in IDS, and explores ways in which an infant could discover phonetic categories in a highly variable linguistic environment.

2. Contextual vowel distributions

Vowel measurements were taken from the Adriaans and Swingley (2017) vowel database.[1] The database contains a total of 4435 vowel tokens which were extracted from recordings of natural mother–infant interactions in the Brent corpus (Brent and Siskind, 2001). Infants were 10 months old at the time of recording. The database represents a relatively large sample of IDS vowels, reflecting the high degree of variability that is found in IDS.

Formants (*F*1 and *F*2, measured at midpoint) were analyzed for two vowels: /i/ and /ɪ/. This pair of vowels was chosen for two reasons. First, these two vowels are by far the most frequently occurring vowels in the data set. There is a total of 1800 /i/ and /ɪ/ tokens (/i/: 801, /ɪ/: 999), which is 41% of the entire set of nine monophthongs in the database. Second, /i/ and /ɪ/ are close neighbors in the vowel space, displaying a large degree of overlap (illustrated in Fig. 1). The /i/-/ɪ/ contrast is thus an important one because the infant encounters it frequently, and it is a difficult one because of the acoustic overlap between the two vowels.

To explore whether consonantal context affects the distribution of vowels in IDS, five different preceding consonants were selected: /d/, /w/, /t/, /m/, and /s/. These

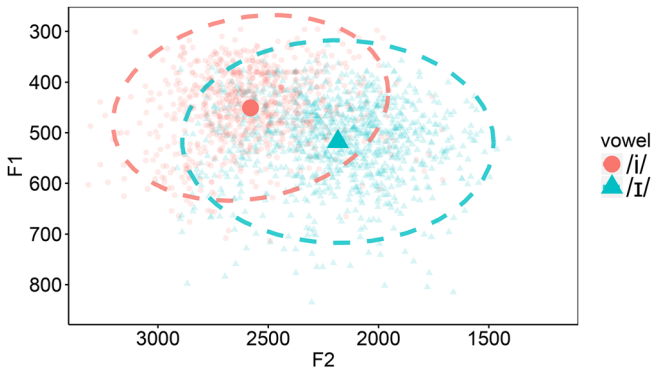

Fig. 1. (Color online) The distribution of /i/ and /ɪ/ tokens in IDS, along with their means and 95% confidence ellipses.

Table 1. Frequency of occurrence of vowels in different consonantal contexts.

	d_	w_	t_	m_	s_	Total
/i/	52	67	100	90	76	385
/ɪ/	112	65	30	32	30	269
Total:	164	132	130	122	106	654

consonants were selected because they are the most frequent consonants preceding /i/ and /ɪ/ (see Table 1). This means that infants often hear these vowels in these contexts, and these contexts thus possibly influence the vowel category learning process. All /i/ and /ɪ/ tokens that occurred in these contexts (regardless of syllable structure) were used in the current study. This resulted in a dataset of 654 vowel tokens.

The distribution of /i/ and /ɪ/ for the five different consonantal contexts is shown in Fig. 2. The graph illustrates that there are differences in the distribution of /i/ and /ɪ/ across different contexts. For example, /i/ is lower after /d/ ($F1 = 488$ Hz) than after /t/ ($F1 = 400$ Hz). Also, formant variability is larger when the preceding consonant is /m/. These differences in category means and variances result in a different overlap between /i/ and /ɪ/ for different contexts.

One way to assess the between-category distance in each context is by calculating the Bhattacharyya distance (Hennig, 2010), which is a measure of the distance between two Gaussian distributions based on their means and covariance matrices. A distance of 0 would mean that the two distributions occupy the exact same acoustic space, whereas larger values indicate greater distances between the distributions. Using this measure, the distance between /i/ and /ɪ/ in specific consonantal contexts can be compared to the distance between /i/ and /ɪ/ in the overall distribution. The Bhattacharyya distance between /i/ and /ɪ/ in the overall distribution (Fig. 1) was 0.3888. For each consonantal context the between-category distance was larger than in the overall distribution: /d/: 0.5108 (+31.4%), /m/: 0.5586 (+43.7%), /s/: 0.4679 (+20.3%), /t/: 0.5366 (+38.0%), and /w/: 0.4978 (+28.0%).

Vowel category distance is thus affected by consonantal context, which suggests that consonantal context is a factor that contributes to the overall variability in IDS vowel distributions. The question is whether a distributional learner that has access to a small selection of highly frequent consonantal contexts could learn vowel categories more effectively than a distributional learner without context. This possibility is explored in a series of simulations assessing the effects of consonantal context on the learnability of vowel categories from IDS.

3. Anchored distributional learning

A new model was developed to assess the effects of consonantal context on distributional category learning. This model, which will be referred to as Anchored

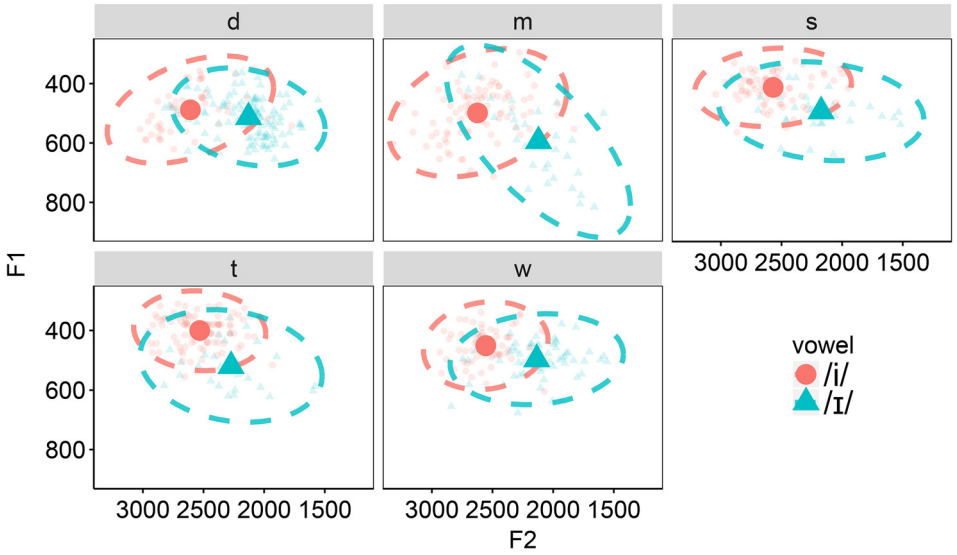

Fig. 2. (Color online) Distribution of /i/ and /ɪ/ tokens in the five most frequent preceding consonantal contexts (/d_/, /m_/, /s_/, /t_/, and /w_/), along with their means and 95% confidence ellipses.

Distributional Learning (ADL), selects multiple consonantal contexts based on their frequency in the input and uses these contexts to learn contextual categories. During training, vowel tokens are anchored in their context (in this case, the particular consonant preceding the vowel). This leads to a set of anchored distributions, from which contextual category parameters are estimated. The learner then estimates the vowel categories by generalizing across different contexts.

In line with earlier studies, formant distributions are represented as multivariate GMMs (e.g., de Boer and Kuhl, 2003), where each category is defined by its mean, covariance matrix, and mixing proportion. For the current study on /i/ and /ɪ/, the model builds separate GMMs for each of the five most frequent consonantal contexts (/d_/, /m_/, /s_/, /t_/, and /w_/). This means that the model learns a mean, covariance matrix, and mixing proportion for each vowel in each context. As a final step the model makes a generalization across different contexts. The generalized parameters are obtained by averaging over the contextual category parameters. The prediction is that this model would obtain higher vowel classification accuracies than a distributional learner that estimates vowel categories directly from formant distributions.

To assess the model's performance, a series of simulations was performed which compared the model to a baseline model. The baseline model was a Distributional Learning (DL) model, implemented as a single GMM which does not make use of context to estimate category parameters. In addition to this baseline model, a supervised learning model was implemented which estimates category parameters based on labeled training data. This was done to obtain an upper bound on the classification accuracy that can be obtained given the structure of the data and the maximum likelihood classification criterion that is imposed on the overlapping Gaussian distributions. The ADL model is expected to outperform the baseline model while approaching the performance of the supervised learner.

Model parameters were estimated using the EM algorithm in *MCLUST for R* (Fraley and Raftery, 2006). The algorithm searches for two ellipse categories, with no constraints on the ellipses' size, shape, and orientation. To compensate for scale differences, formant values ($F1$, $F2$) were transformed to z scores. For each model 2000 training points were sampled from each vowel category's multivariate normal distribution in the appropriate context (e.g., Vallabha *et al.*, 2007). The test set that was used to assess each model's classification accuracy consisted of 2000 newly sampled data points for each category. To obtain a reliable estimate of each model's performance, a total of 100 repeated runs were done for each model.

Table 2 shows the classification accuracies for the three tested models. The ADL model outperforms the baseline DL model, which means that the category parameters that are obtained by generalizing over different anchored distributions (i.e., formant distributions in different phonological contexts) are more accurate than parameters that are estimated directly from the entire formant distribution. While the increase in classification accuracy from the DL to the ADL model might appear modest (from 0.7797 to 0.8169), the ADL model's performance approaches that of the supervised learning model (0.8211). This means that there is not much more room for improvement beyond the ADL model's performance in this experimental setup. The classification performance of each model is illustrated in Fig. 3. Here it can be seen that the DL model has difficulty estimating parameters for /ɪ/, while the ADL and supervised models look nearly identical, and have a closer resemblance to the underlying distribution of the test data.

4. Discussion

A growing number of studies have reported on the large degree of variability found in IDS. In order to advance our knowledge of how phonetic categories are formed during early language acquisition, there is a need to identify factors that may contribute to this variability, and to explore ways in which infants could discover category structure

Table 2. Classification accuracies on the /i/-/ɪ/ dataset. For each model the mean accuracy over 100 repeated runs is displayed, along with the 95% confidence interval (CI).

		95% CI	
Model	Mean	Lower	Upper
Distributional learning	0.7797	0.7742	0.7868
Anchored distributional learning	0.8169	0.8158	0.8180
Supervised learning	0.8211	0.8204	0.8218

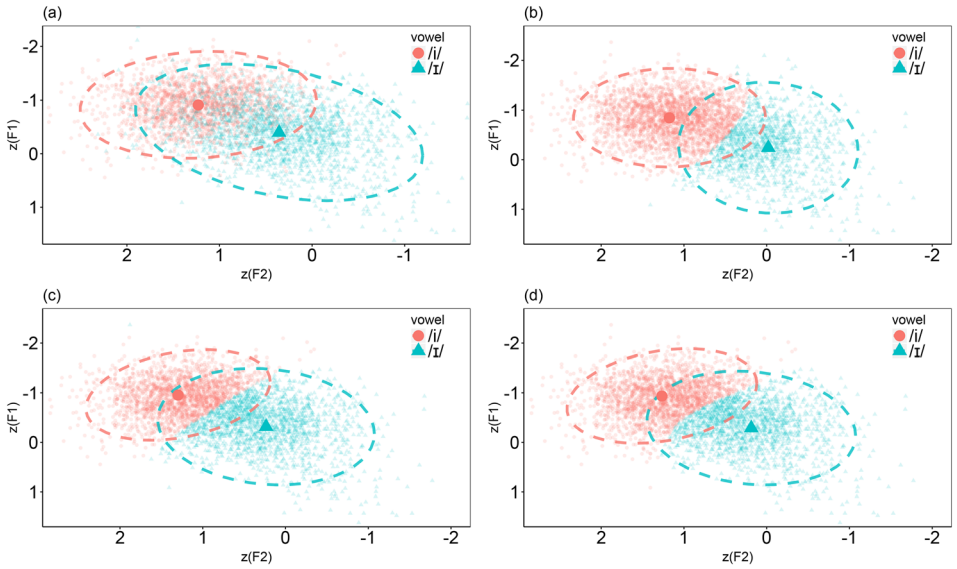

Fig. 3. (Color online) Distribution of the /i/-/ɪ/ test data and categorizations by three different models: (a) test data, (b) Distributional learning, (c) Anchored distributional learning, and (d) supervised learning.

in light of this variability. The results of the current study suggest that consonantal context could be used as a possible source of information to facilitate distributional vowel learning from IDS. The ADL model generalizes across different contexts, and this strategy improved categorization performance on a difficult contrast in natural American–English IDS.

The improved performance of the model could be due to the fact that it compensates for consonant-vowel (CV) coarticulation. Vowel variability is partially predictable from the consonantal context, and knowing the context can thus reduce unexplained variability, leading to improved classification performance (see also Dillon *et al.*, 2013). It is important to note that the vowel tokens used in this study had been measured at midpoint. This means that only those coarticulation effects that were still present halfway through the vowel's duration had an impact on the results. It is possible that the learnability advantage might have been larger if formant measurements had been taken closer to the actual CV transition point.

From a language development perspective, it is important to note that the model presented here only relies on a small set of highly frequent consonant environments, and does not need the full consonant inventory. This is important because consonants are generally thought to be acquired later than vowels (e.g., Werker and Tees, 1984). However, 1-month-old infants can already discriminate consonant contrasts (Eimas *et al.*, 1971), and 6- to 8-month-old infants show evidence of consonant distributional learning (Maye *et al.*, 2002). While the exact developmental path toward acquiring phonetic categories remains unclear, these findings give some support for the use of consonants as anchors.

There is an ongoing debate regarding the role of variability in early language acquisition. For instance, it has been argued that talker variability might help the learner to identify and focus on contrastive acoustic dimensions (e.g., Rost and McMurray, 2010). The current study adds to earlier computational studies exploring how variability in IDS might be navigated successfully by using contextual information such as lexical context (Feldman *et al.*, 2013) and prosodic context (Adriaans and Swingley, 2017). It is possible that multiple contextual factors affect early language development, and the general model presented here could be used to integrate different factors. The findings support a view in which the infant's distributional learning mechanism is anchored in various types of context, in order to cope with puzzling amounts of variability in the input.

References and links

[1]The data are available from www.fransadriaans.nl/resources/ids-vowels/.

Adriaans, F., and Swingley, D. (**2017**). "Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability," J. Acoust. Soc. Am. **141**(5), 3070–3078.

Bergelson, E., and Swingley, D. (**2012**). "At 6–9 months, human infants know the meanings of many common nouns," Proc. Natl. Acad. Sci. U.S.A. **109**, 3253–3258.

Brent, M. R., and Siskind, J. M. (**2001**). "The role of exposure to isolated words in early vocabulary development," Cognition **81**, B33–B44.
Cristià, A., and Seidl, A. (**2014**). "The hyperarticulation hypothesis of infant-directed speech," J. Child Lang. **41**, 913–934.
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.
Dillon, B., Dunbar, E., and Idsardi, W. (**2013**). "A single-stage approach to learning phonological categories: Insights from Inuktitut," Cognitive Sci. **37**, 344–377.
Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (**1971**). "Speech perception in infants," Science **171**, 303–306.
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.
Fraley, C., and Raftery, A. E. (**2006**). "MCLUST Version 3 for R: Normal mixture modeling and model-based clustering," Technical Report.
Hennig, C. (**2010**). "Methods for merging Gaussian mixture components," Adv. Data Anal. Class. **4**, 3–34.
Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (**2001**). "Effects of consonant environment on vowel formant patterns," J. Acoust. Soc. Am. **109**(2), 748–763.
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.
Maye, J., Werker, J. F., and Gerken, L. (**2002**). "Infant sensitivity to distributional information can affect phonetic discrimination," Cognition **82**, B101–B111.
Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., and Mazuka, R. (**2017**). "Vowels in infant-directed speech: More breathy and more variable, but not clearer," Cognition **166**, 84–93.
Polka, L., and Werker, J. F. (**1994**). "Developmental changes in perception of nonnative vowel contrasts," J. Exp. Psychol. **20**, 421–435.
Rost, G. C., and McMurray, B. (**2010**). "Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning," Infancy **15**, 608–635.
Swingley, D. (**2009**). "Contributions of infant word learning to language development," Phil. Trans. R. Soc. B **364**, 3617–3622.
Tincoff, R., and Jusczyk, P. (**1999**). "Some beginnings of word comprehension in 6-month-olds," Psychol. Sci. **10**, 172–175.
Trehub, S. E. (**1976**). "The discrimination of foreign speech contrasts by infants and adults," Child Develop. **47**, 466–472.
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.