Distributional Learning of Vowel Categories is Supported by Prosody in Infant-Directed Speech

Frans Adriaans (adriaans@psych.upenn.edu) and **Daniel Swingley (swingley@psych.upenn.edu)**
Department of Psychology and Institute for Research in Cognitive Science, University of Pennsylvania
3401 Walnut Street, Suite 400A, Philadelphia, PA 19104, USA

Abstract

Infants' acquisition of phonetic categories involves a distributional learning mechanism that operates on acoustic dimensions of the input. However, natural infant-directed speech shows large degrees of phonetic variability, and the resulting overlap between categories suggests that category learning based on distributional clustering may not be feasible without constraints on the learning process, or exploitation of other sources of information. The present study examines whether mothers' prosodic modifications within infant-directed speech help the distributional learning of vowel categories. Specifically, we hypothesize that 'motherese' provides the infant with a subset of high-quality learning tokens that improve category learning. In an analysis of vowel tokens taken from natural mother-infant interactions, we found that prosody can be used to distinguish high-quality tokens (with expanded formant frequencies) from low-quality tokens in the input. Moreover, in simulations of distributional learning we found that models trained on this small set of high-quality tokens provide better classification than models trained on the complete set of tokens. Taken together, these findings show that distributional learning of vowel categories can be improved by attributing importance to tokens that are prosodically prominent in the input. The prosodic properties of motherese might thus be a helpful cue for infants in supporting phonetic category learning.

Keywords: Infant-directed speech; phonetic category learning; prosody; computational modeling.

Introduction

Infants in the first year of life develop knowledge of the phonetic categories that make up the consonants and vowels of their native language (e.g., Werker & Tees, 1984). The early age at which this takes place rules out learning accounts in which semantic contrast in phonologically similar words drives most category learning. As a result, it is assumed that infants learn phonetic categories using an implicit statistical clustering process that relies on separation of the categories in perceptual space. Indeed, 6- and 8-month-old infants have been found to form representations of two distinct categories (e.g., /d/ and /t/) when exposed to an artificially generated bimodal distribution on a distinguishing acoustic dimension, but not when exposed to a unimodal distribution (Maye, Werker, & Gerken, 2002; Maye, Weiss, & Aslin, 2008; see also Cristià, McGuire, Seidl, & Francis, 2011). Further evidence for the plausibility of distributional learning of phonetic category structure comes from analyses of infant-directed speech. Mothers appear to provide their infants with acoustic cues that support distributional learning of phonetic categories (Werker et al., 2007). In particular, infant-directed speech is characterized by expansion of the F1-F2 vowel formant space, which could enhance the separability of vowel categories (Kuhl et al., 1997). Several studies have used approximations of infant-directed speech tokens as input to computational procedures (such as multivariate Gaussian mixture models) that succeed in learning vowel categories, suggesting that distributional learning could be feasible for infants (de Boer & Kuhl, 2003; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007).

Some caution is appropriate in interpreting these findings. Studies that show the usefulness of distributional cues for category learning have, in large part, been based on analyses (and simulations) of vowel tokens that were elicited in a laboratory setting, and that occurred in a small number of words or nonwords. It is possible that maternal speech under these conditions is different from maternal speech in quotidian home contexts. Analyses of natural, unscripted infant-directed speech recordings show that vowel distributions are highly variable, and that overlap between categories poses a substantial problem for distributional category learning (Swingley, 2009). One possibility suggested by this result is that infants' learning of phonetic categories is guided by additional sources of information, such as the emerging lexicon (Feldman, Griffiths, & Morgan, 2009; Swingley, 2009).

Another possibility, explored here, is that infants are able to succeed in category learning because they have a bias to attend to some tokens more than to others, and that these salient tokens are clearer instances of their categories. If so, the difficulty of distributional category learning is overestimated by considering the whole mass of experienced speech sounds. This notion is indirectly supported by studies showing that infants prefer "motherese" speech over adult-directed speech. Across different languages motherese is characterized by acoustic exaggeration, including higher overall pitch, greater intonation contours, and longer durations (Fernald et al., 1989; Grieser & Kuhl, 1988; Kuhl et al., 1997). These properties have been found to modulate infants' attention, and possibly facilitate language learning by enhancing infants' speech discrimination skills (Fernald & Kuhl, 1987; Karzon, 1985; Liu, Kuhl, & Tsao, 2003; Trainor & Desjardins, 2002).

It remains to be demonstrated that motherese effectively guides the infant's attention to those vowel tokens that are most useful for category learning. Computational models that aim to explain category learning are typically fit to isolated, equally weighted vowel tokens (de Boer & Kuhl, 2003; Vallabha et al., 2007). Such models overlook prosodic context which might make certain vowel tokens more attractive than others, and which thus potentially affects the learnability of vowel categories.

The current study examines the relation between prosodic exaggeration and vowel learning from infant-directed speech. Specifically, we hypothesize that motherese provides the infant with a subset of high-quality learning tokens that improves distributional category learning. First, we analyze prosodic determinants of vowel expansion within infant-directed speech, thereby attempting to predict which vowel tokens in the infant's speech input could be particularly beneficial for phonetic category learning. Second, we simulate the distributional learning of phonetic categories in order to examine whether prosodic focus helps in discovering category structure in cases of large overlap between categories. Importantly, analyses and simulations are done on realistic data, using vowel tokens taken from recordings of natural mother-infant interactions. We thus provide a test of distributional learning in a setting that acknowledges the variability and complexities that are found in real everyday speech.

Vowel Expansion in Infant-Directed Speech

Earlier studies on vowel expansion compared speech directed to adult listeners and speech directed to infant listeners (Kuhl et al., 1997). While infant-directed speech is often hyperarticulated compared to adult-directed speech, the mechanisms underlying vowel expansion in infant-directed speech are not yet fully understood. It seems likely that the prosodic exaggeration notable in infant-directed speech has an effect on vowel expansion. Here we explore this possibility by asking whether prosodically prominent vowels in infant-directed speech are hyperarticulated relative to parts that are not prosodically highlighted. In analyses of recordings of natural mother-infant interactions, we examine whether prosodic focus predicts vowel expansion (see also Mo, Cole, & Hasegawa-Johnson, 2009). We examine expansion in tokens that were labeled to have focus by human assessors (what we define as "annotated focus"), and also in tokens that were defined as exaggerated on acoustic grounds (higher pitch, greater pitch change, and longer duration; what we define as "acoustic focus"), to determine whether such vowels are more differentiable. Evidence of vowel expansion at prosodically predictable locations in infant-directed speech would indicate that attention to prosody could aid in vowel category learning.

Methods

Vowel expansion was examined by analyzing vowel productions by one mother ('f1') in the Brent corpus (Brent & Siskind, 2001), available through CHILDES (MacWhinney, 2000). These recordings consist of natural, unscripted infant-directed speech and therefore have no restrictions on the words or vowel types that may occur. Formant (F1, F2) measurements were obtained and hand-checked for 1,166 vowel tokens. Tokens covered the monophthongs of American English (/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /ɔ/, /ʊ/, /u/). Measurements taken at 33% and 50% of the vowel's duration were averaged and transformed into z scores to neutralize scale differences. Vowel expansion was measured by calculating the Euclidean distance of each token to the center of the mother's vowel space (Bradlow, Torretta, & Pisoni, 1996). In order to measure prosodic prominence in infant-directed speech each vowel token was judged by a human assessor who indicated whether the vowel occurred in a syllable that the mother was trying to emphasize (*focus* vs. *no focus*). Potential acoustic correlates of focus that were considered were: duration (logarithm of the absolute duration in ms.), pitch (F0 averaged over 33% and 50% measurements), and pitch change (the absolute value of the difference in F0 between measurements at 33% and 50%). The label of "acoustic focus" was assigned to vowels that exceeded the z-score of 0.5 for at least one of the three dimensions.

Results

Table 1 shows the number of focused and unfocused tokens for each vowel. The annotated-focus set contained 336 vowel tokens (28.8% of the total set). The acoustic-focus set had 543 tokens (46.6% of the total set). Figure 1 shows the mean formant frequencies of vowels in focused and unfocused position. Vowels in focused position were further away from the center of the vowel space than vowels in unfocused position.[1]

Stepwise linear regression analyses revealed that annotated focus is a significant predictor of the vowel's distance from the center of the vowel space, independent of vowel type (adjusted $R^2 = 0.4221$; vowel***, focus**, vowel:focus *ns*). Vowels in syllables with annotated focus were thus hyperarticulated relative to vowels in unfocused syllables. This confirms the intuition that in natural infant-directed speech mothers exaggerate certain vowels by marking them with sentence focus. Interestingly, vowel expansion did not manifest itself through stretching of the triangle defined by the "point vowels" (/i/, /ɑ/, /u/), but rather followed a consistent pattern of expansion throughout the entire set of monophthongs.

The tokens that had acoustic focus showed very similar results. Stepwise regression revealed that acoustic focus is a significant predictor of vowel expansion (adjusted $R^2 = 0.4300$; vowel***, focus***, vowel:focus*). These results indicate that whether infants are able to judge focus (as our annotators did), or whether they simply pay attention to tokens that have extreme values on prosodic dimensions (i.e., "acoustic focus"), the tokens that have focus show expansion, and are thus possibly particularly helpful for the learning of phonetic categories.

In sum, vowels that are prosodically exaggerated might be particularly useful for phonetic learning because they have distributional properties that enhance the separability of vowel categories. The overlap between categories, however, is still substantial. It thus remains to be demonstrated that prosodic highlighting makes a meaningfully large difference in the learnability of vowel categories.

[1]The exception was /ɔ/ and /ʊ/ in the acoustic-focus set. The means of these vowels are unreliable due to their low frequency of occurrence in the data set. (See Table 1.)

Table 1: Frequency of occurrence of vowels in focused and unfocused position.

	/i/ (182)	/ɪ/ (320)	/ɛ/ (163)	/æ/ (139)	/ɑ/ (112)	/ʌ/ (130)	/ɔ/ (21)	/ʊ/ (22)	/u/ (77)	Total: (1,166)
Focus (annotated)	41	72	51	67	32	36	12	5	20	336
No focus (annotated)	141	248	112	72	80	94	9	17	57	830
Focus (acoustic)	105	112	65	95	55	45	16	10	40	543
No focus (acoustic)	77	208	98	44	57	85	5	12	37	623

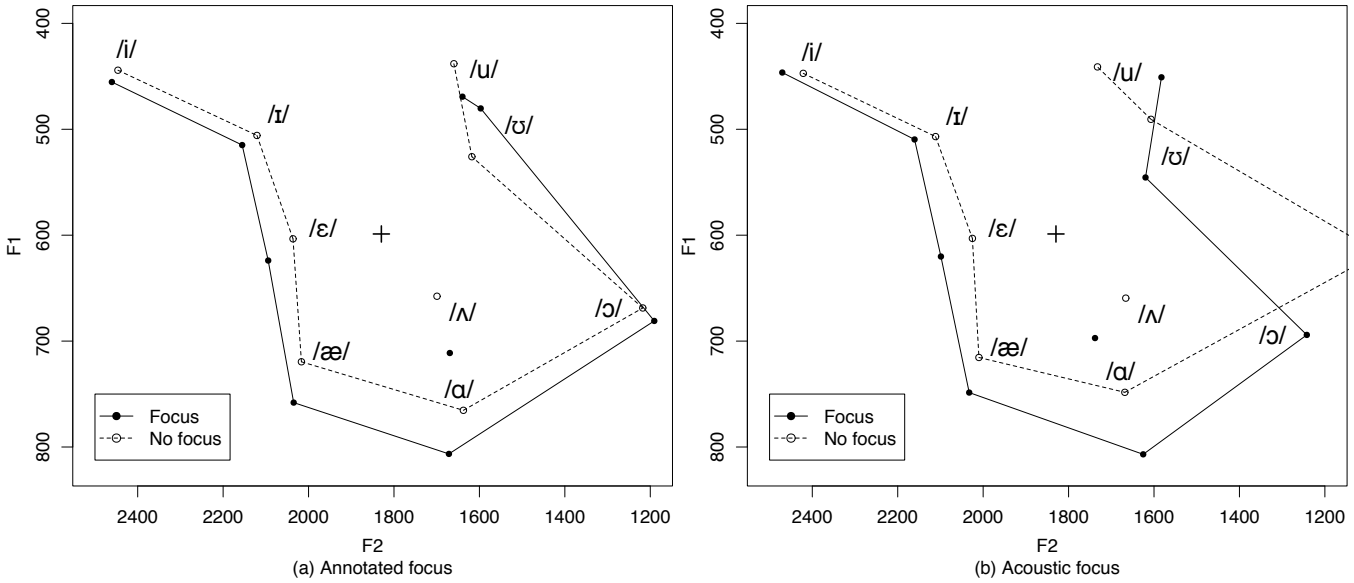

Figure 1: Vowel expansion within infant-directed speech. '+' indicates the center of the mother's vowel space.

The Learnability of Vowel Categories

In order to see if prosodically highlighted vowels would be beneficial to infant language learners, we simulate the distributional learning of vowel categories from infant-directed speech. In particular, we examine whether prosodic focus helps 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 prosodic properties of motherese support phonetic category learning.

Methods

The learnability of vowel categories is simulated for two different sets of vowels: /i/, /ɪ/, /ɛ/ and /ɛ/, /æ/, /ɑ/. These sets were chosen because they each contain three vowels that are close in the F1-F2 formant space. As a consequence, the overlap between categories is large, and the learning of these categories poses a substantial problem for distributional learning models. In line with earlier work on computational modeling of phonetic category learning (e.g., de Boer & Kuhl, 2003; McMurray et al., 2009; Vallabha et al., 2007), we treat categories as multivariate Gaussian distributions. The learning problem is characterized as estimating the parameters (means, covariances and mixing proportions) for these distributions. In our case, categories are defined as 2-dimensional distributions (the z scores of the first and second formants). Data points are assigned to the category that has the maximum likelihood for that point. Parameters of the Gaussian distributions are estimated using the EM algorithm (Dempster, Laird, & Rubin, 1977) as implemented in the *MCLUST for R* software package (Fraley & 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, Getty, Clark, & Wheeler, 1995), the models were given no information or constraints with respect to volume, shape, or orientation.

In order to assess whether focused tokens were helpful for category learning, models were trained on a subset of the data (either the annotated-focus set or the acoustic-focus set). The Gaussian distributions that were estimated from these subsets were subsequently used to classify all vowel tokens in the data set. We predicted that Gaussian mixture models trained

on a relatively small set of prosodically prominent vowel tokens would provide a better classification of the data than Gaussian mixture models that were trained on the complete set of vowel tokens. Performance of the unsupervised clustering models was assessed by comparing their classification accuracy to a supervised learner that learned three Gaussian categories based on actual vowel category labels. The supervised learner represented an upper bound on the classification accuracy that can be obtained given the maximum likelihood classification criterion that is imposed on the overlapping Gaussian distributions.

Results

Table 2 shows the classification accuracy for models trained using all tokens, annotated-focus tokens, acoustic-focus tokens, and all tokens' category labels (this last being the supervised "ideal"). The first thing to note is that the classification accuracy of the supervised learners was below 80%, confirming that overlap between categories was substantial. Considering the unsupervised "All tokens" model, the 12- to 15-percentage-point decline relative to the supervised model shows that the categories are not trivially detectable in the distributions.[2] Using vowel tokens annotated as focused aided accuracy to a small degree in the i-ɪ-ɛ data set, a result that nevertheless reveals some utility to focus marking given that this model was trained on only 164 data points rather than the entire dataset (which consisted of a total of 665 i-ɪ-ɛ tokens). However, for the ɛ-æ-ɑ data set the clustering algorithm was unable to fit a model to the annotated-focus tokens. We believe that this is due to the small size of the annotated-focus data set for ɛ-æ-ɑ ($n = 150$, with only 32 tokens for /ɑ/, see Table 1). Thus, focused vowels are, in at least some cases, variable enough that category solutions are difficult to determine when the quantity of data is very small.

Training on the bigger set of acoustic-focus tokens helped learning substantially, bringing the model within 3 percentage points of the supervised model in the i-ɪ-ɛ data set, and within 6 percentage points in the ɛ-æ-ɑ data set. Models that were trained on tokens that were acoustically prominent (long duration, high pitch, greater pitch movement) thus showed substantial classification improvement as compared to models that were prosodically uninformed. To illustrate the performance of different learning models, we display the i-ɪ-ɛ data along with the classifications that are predicted by different models in Figure 2. Figure 2 shows that only the acoustic-focus training set is able to predict three clearly distinct categories.

As it turns out, tokens that have focus or show acoustic exaggeration have a positive effect on the unsupervised learning of vowel categories. Importantly, these high-quality tokens are easily identifiable based on their prosodic properties. It is thus likely that these tokens are identifiable for infant language learners, and contribute to language learning.

[2] Such a decline is not found in models of the point vowels (i-ɑ-u) alone, for which we found accuracy $> 90\%$ for both the supervised and unsupervised learner.

Table 2: Classification accuracy on two different sets of overlapping vowel categories.

Model	Accuracy	
	i-ɪ-ɛ	ɛ-æ-ɑ
All tokens	0.6060	0.6449
Annotated focus	0.6331	-
Acoustic focus	0.7008	0.7343
Supervised	0.7278	0.7947

Discussion

In learning the phonetic categories of their native language, infants face large amounts of variability in the acoustic realizations of different vowel tokens. This poses a substantial problem for the purely bottom-up distributional learning of vowels. Here we presented one possible source of information that may guide phonetic category learning. If infants are able to detect high-quality learning tokens in the input, then they could make considerable progress in category learning. Motherese may play an important role in this process, by bringing such "high-quality" tokens to the infant's attention through prosodic modifications of the speech stream.

In our clustering experiments, focus as annotated by human listeners was not as effective as "focus" estimated using simple, one-dimensional acoustic measures. It is possible that this difference derived from sample-specific gaps in the number or quality of human-annotated focus tokens for some vowel types; this cannot be ruled out without examining other samples. Furthermore, it is likely that annotators' judgments of focus were, in some cases, based on their interpretation of the speaker's intentions: an adult listener might judge a word as being the one the speaker wished to emphasize even if the phonetics were not particularly marked. Still, the superior performance of the model that learned from the tokens that were simply more extreme on at least one of the acoustic dimensions shows that the benefits of "motherese" prosodic highlighting do not depend on possession of a mature capacity for interpreting focus. Sensitivity to simple dimensions like duration or pitch goes a long way.

Infant-directed speech prosody, with its exaggerated prosodic variation, certainly captures infants' attention, and this may be important for learning. Earlier studies have shown that pitch contours enhance infants' discrimination skills, since contours increase the acoustic salience of formant frequencies (Trainor & Desjardins, 2002). Such perceptual salience is not taken into account by our model. Our results show that prosody has additional benefits. We find that acoustically exaggerated tokens show a different distribution in the F1-F2 space, with greater distances from the center and enhanced separability of categories. The picture that emerges from earlier studies, combined with the current findings, is

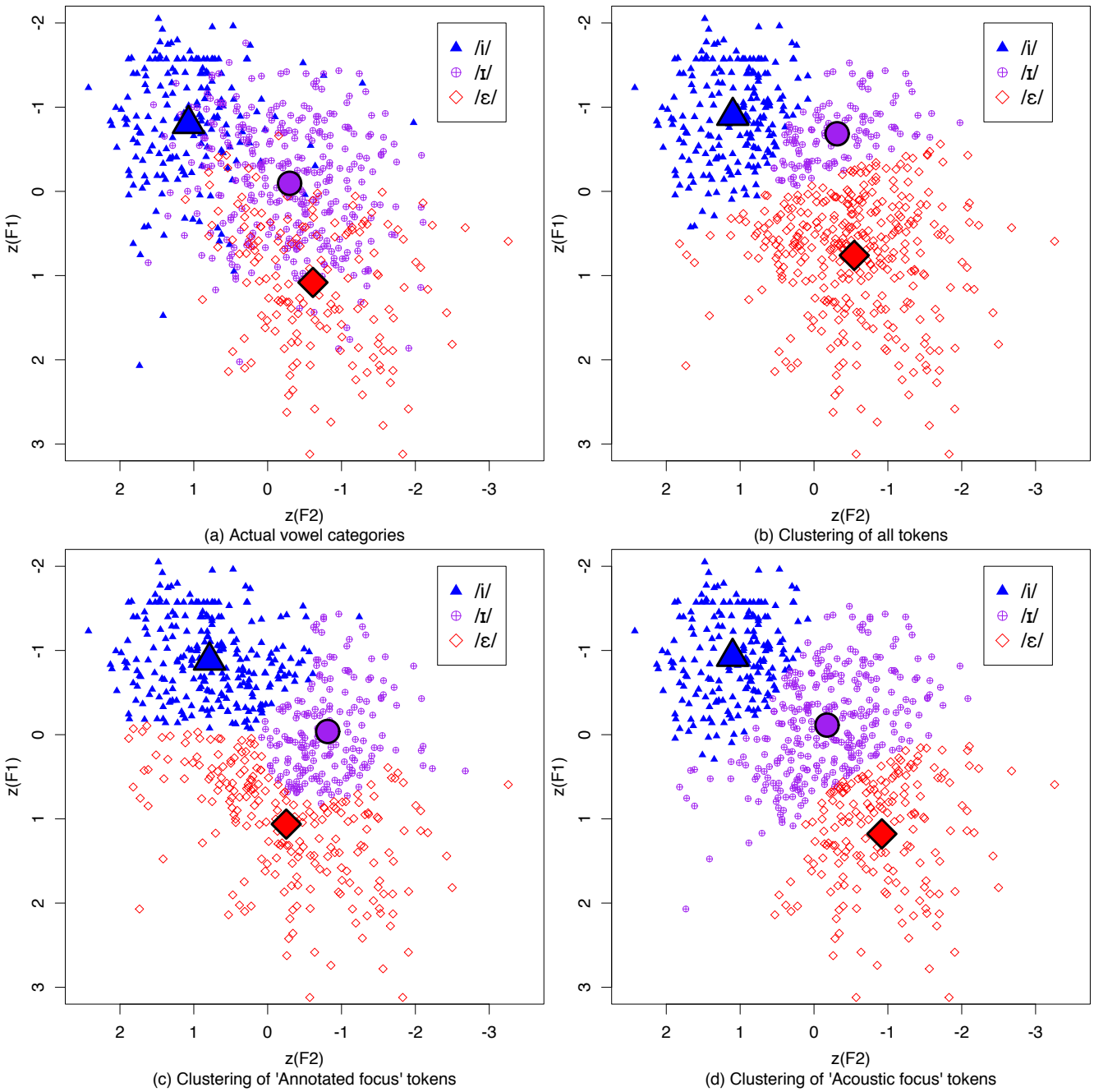

Figure 2: The i-ɪ-ɛ data set with (a) actual categories, (b) predicted categories based on all tokens, (c) predicted categories based on focused tokens, and (d) predicted categories based on acoustically exaggerated tokens. The means are plotted for each (predicted) category.

that the exaggerated prosody of infant-directed speech may capture infants' attention to speech in a general fashion, and at the same time provide an enhanced speech signal that supports language learning – if infants' category learning favors attention to the most salient instances.

Acknowledgments

This work was funded by the Netherlands Organisation for Scientific Research (NWO) grant 446.010.027 to F.A. and NIH grant R01-HD049681 to D.S.

References

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

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cogni-*

tion, *81*, B33-B44.

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

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.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1-38.

Feldman, N. H., Griffiths, T. H., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 2208-2213).

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*, 279-293.

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

Fraley, C., & Raftery, A. E. (2006). *MCLUST Version 3 for R: Normal mixture modeling and model-based clustering* (Tech. Rep.). Seattle, WA: University of Washington.

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*, 14-20.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099-3111.

Karzon, R. G. (1985). Discrimination of polysyllabic sequences by one- to four-month-old infants. *Journal of Experimental Child Psychology*, *39*, 326-342.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*, 684-686.

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

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, volume 2: The database* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*, 122-134.

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

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*, 369-378.

Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: evidence from formant structure. In *Proceedings of Interspeech 2009* (p. 2535-2538).

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*, 3617-3622.

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*, 335-340.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 13273-13278.

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

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