PNAS

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Naturalistic speech supports distributional learning across contexts

Kasia Hitczenko[a,1] and Naomi H. Feldman[b,c]

Edited by Janet Werker, The University of British Columbia, Vancouver, BC, Canada; received December 23, 2021; accepted July 18, 2022

At birth, infants discriminate most of the sounds of the world's languages, but by age 1, infants become language-specific listeners. This has generally been taken as evidence that infants have learned which acoustic dimensions are contrastive, or useful for distinguishing among the sounds of their language(s), and have begun focusing primarily on those dimensions when perceiving speech. However, speech is highly variable, with different sounds overlapping substantially in their acoustics, and after decades of research, we still do not know what aspects of the speech signal allow infants to differentiate contrastive from noncontrastive dimensions. Here we show that infants could learn which acoustic dimensions of their language are contrastive, despite the high acoustic variability. Our account is based on the cross-linguistic fact that even sounds that overlap in their acoustics differ in the contexts they occur in. We predict that this should leave a signal that infants can pick up on and show that acoustic distributions indeed vary more by context along contrastive dimensions compared with noncontrastive dimensions. By establishing this difference, we provide a potential answer to how infants learn about sound contrasts, a question whose answer in natural learning environments has remained elusive.

phonetic learning | language acquisition | distributional learning

Languages differ in the speech sound inventories they use to reliably convey meaning. For example, Hindi has a distinction between unaspirated retroflex [t] vs. dental [t̪] that is used to distinguish meanings (e.g., [tal] means "postpone," while [t̪al] means "beat"), but English does not. Adult speakers are generally tuned to the properties of the language(s) they speak. For example, while most adult Hindi speakers can hear the difference between [t] and [t̪], most English-speaking adults cannot (1). Because speech sound inventories differ across languages, listeners must learn about the speech sounds of their language(s) from the input they hear.

The first signs of this phonetic learning appear within the first year of life. During their first couple of months, infants can discriminate most sounds of the world's languages, showing similar perceptual abilities regardless of their language experience. For example, both newborn English-learning and Hindi-learning infants can hear the distinction between retroflex [ta] vs. dental [t̪a], a speech contrast that exists in Hindi but not English. However, over the course of the first year of life, this changes. Infants become language-specific listeners, starting to more closely resemble adults in their discrimination abilities. Their ability to discriminate nonnative contrasts (e.g., between retroflex [ta] and dental [t̪a]) for English-learning infants) declines (2–4), whereas their ability to discriminate contrasts in their own language(s) improves (5).

These perceptual changes have generally been taken as evidence that infants are learning which acoustic dimensions are contrastive in their language(s): that is, which acoustic dimensions have multiple categories along them (6). Speech sounds differ in how they are acoustically produced, and one or more acoustic dimensions will be used to signal differences between sound contrasts. The idea, then, is that infants become aware of which acoustic dimensions are used to contrast the meaningful sounds in their language(s) and begin primarily focusing on those dimensions when perceiving speech.

Decades of research into how infants learn about contrastiveness in their first year of life has built a wealth of knowledge in this area; however, we still do not know what aspects of the speech signal allow infants to make these inferences from the acoustically variable speech they hear in their daily lives. One of the most well-studied current proposals for how infants learn which dimensions of their language are contrastive is known as distributional learning (6). It proposes that infants learn the contrastive dimensions of their language(s) by tracking the frequency distribution of sounds along acoustic cue dimensions. If an infant observes a bimodal (two-peaked) distribution along a dimension, then they learn that the dimension is contrastive, whereas if an infant observes a unimodal (one-peaked) distribution, then they learn that the dimension is not contrastive.

Significance

Languages differ in the speech sounds they use, and humans need to learn which sounds their language uses. This learning starts early. By age 1, infants have already tuned into their language(s): their ability to hear sound distinctions from their language(s) improves, while they often lose the ability to hear other sound distinctions. Understanding how this early learning proceeds is important as it serves as a foundation for later development; however, it has proven difficult to identify a learning mechanism that works on the true input infants hear. We present an account for how infants learn the speech contrasts of their language and show that the necessary signal is present in naturalistic speech, advancing our understanding of early language learning.

Author affiliations: [a]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Études Cognitives, Ecole Normale Supérieure, Ecole des Hautes Etudes en Sciences Sociales, CNRS, Paris Sciences et Lettres Research University, 75005 Paris, France; [b]Department of Linguistics, University of Maryland, College Park, MD 20742; and [c]University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742

Author contributions: K.H. and N.H.F. designed research; K.H. performed research; K.H. analyzed data; and K.H. and N.H.F. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: kasia.hitczenko@ens.psl.eu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2123230119/-/DCSupplemental.

Published September 12, 2022.

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

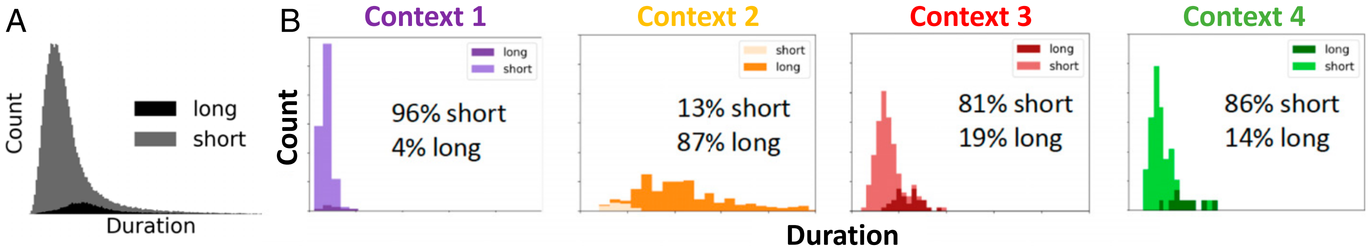

Fig. 1. (*A*) The frequency distribution of sounds along the duration dimension in Japanese is unimodal, despite vowel length being contrastive. (*B*) Vowel frequency distributions along duration, for four Japanese contexts (defined by prosodic position, neighboring sounds, and vowel quality). The relative proportion of phonemically short and long vowels changes substantially across contexts, which results in differently shaped distributions. The short vs. long categories are color-coded for the reader's benefit. Infants (and our analyses) do not have access to this color information when learning, only the overall distributions.

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

This account has experimental support: distributions of sounds affect infants' discrimination in the laboratory (6–10). In addition, when bimodality is present in the input, computational models successfully learn correct contrasts (11, 12). However, a key assumption underlying this proposal is that contrastive dimensions do indeed exhibit bimodality, and while this is the case for some contrasts, recent work looking at naturalistic speech corpora has shown that this is not a universal property of child-directed speech (13, 14).

For example, in Japanese, vowel length is contrastive (15), meaning that two different words like /toko/ ("bed") and /toko:/ ("travel") can be distinguished solely by how long a vowel is. However, analyses of a spontaneous corpus of Japanese child-directed speech reveal that the distribution along the duration dimension is unimodal despite being contrastive (13) (Fig. 1*A*; note that infants do not have access to the individual color-coded short and long vowel distributions shown in this figure, only the combined overall distribution). A similar finding has been reported for Dutch vowel length (14), as well as many other contrasts (5, 16, 17). That is, although infants are able to use distributional information for learning when available, it is not available for all of the contrasts they learn about, so distributional learning is not sufficient.

Many follow-up theories have been proposed to explain how infants learn in cases where bimodality is not present. This has included theories arguing that bimodality might be present when considering only the most prominent sounds (e.g., stressed vowels) (18, 19), when normalizing for effects of neighboring sounds or other factors (20), or when using word-level, visual, or referential information (14, 21–25). While many of them have experimental support and work on controlled laboratory speech, over the past 40 y, it has proven difficult to identify a learning mechanism that works on the true speech infants hear. Swingley (14) takes an important step in that direction by showing that, in Dutch, average vowel durations by word type are often longer in word types with long vowels than word types with short vowels (and, thus, that short and long vowels may be separable). However, we still do not have a measure that consistently separates vowels with a contrast from vowels without a contrast across corpora, languages, and vowel qualities. This problem is so extreme that recent work has suggested that infants might not actually be learning how many phonetic categories there are along a dimension at all, because this signal is not present in their input in a way that they have access to (26).

In this paper, we show that the necessary signal to learn which acoustic dimensions are contrastive may be present in naturalistic input and accessible to infants. Our proposal takes advantage of the contextual information of a sound, which infants are sensitive to (22–24, 27–31). In this work, we take the context of a sound to include factors like its neighboring sounds, its prosodic position in a word/utterance (i.e., if it immediately borders a word or utterance boundary), and its word frame; however, we think of context more broadly as any information that listeners track about where a sound occurs or who spoke it. When an acoustic dimension is contrastive, there are multiple categories along it, and the relative proportion of those categories may differ across contexts (e.g., if two categories are present, one context may be 50% category 1 and 50% category 2, whereas another context may be 90% category 1 and only 10% category 2). We show that such differences in category frequency—which are extremely common across languages (32–34)—can help infants distinguish contrastive from noncontrastive dimensions.

We test our proposal on two test cases, Japanese and Dutch, which have been most problematic for both distributional learning and additional previous theories, and show that our proposal explains how infants could nonetheless learn the contrast from information available to them within their first year of life. Complemented by previous findings that 1) infants are sensitive to distribution shapes and contextual information and 2) changes in the relative proportion of sounds across contexts are a cross-linguistically widespread property of sound categories, these results are promising and suggest that infants may be able to learn about contrastiveness from naturalistic speech input, thus pointing toward a possible answer to a long-standing question in the field.

Distributional Learning across Contexts

The inspiration for our proposal comes from a finding showing that the context a sound occurs in (neighboring sounds, prosodic position, speaker, etc.) is predictive of its identity: just knowing what context a Japanese vowel appears in can predict its length with around 95% accuracy (35). This means that short and long vowels appear in different proportions in different contexts. Most contexts have almost all short vowels (e.g., context 1 in Fig. 1*B*), whereas some contexts have almost all long vowels (e.g., context 2 in Fig. 1*B*), and some are in between (e.g., contexts 3 and 4 in Fig. 1*B*). Fig. 1*B* reveals that these changes in the relative proportion of short and long vowels can change the overall shape of the frequency distribution in the context. All of the distributions in Fig. 1*B* are unimodal, despite the fact that there are two categories. Thus, they would not be conducive to the distributional learning theory proposed by ref. 6. However, this is only one aspect of a distribution's shape, and across contexts, the distributions differ in how wide or peaky they are, where they peak, and so forth. This arises because of two facts: 1) when a dimension is contrastive, the overall frequency distribution in each context is the sum of the short vowel distribution and the long vowel distribution, and 2) short and long vowels have different distributions, as can be seen in Fig. 1. Taken together, this means that in a language like Japanese, where there is a contrast, we would expect different relative proportions of short vowels and long vowels across different contexts, and since short vowels and long vowels have different acoustic distributions, we would expect the overall distribution to change across different contexts. On the other hand, in a language like French (where there is no length contrast), shape changes cannot arise from different relative proportions of short and long vowels because there is no short vs. long vowel distinction.

In light of this, we propose that infants might learn that a dimension is contrastive by tracking the acoustic distribution along that dimension across different contexts. They could compare the shapes of the distributions across those contexts and infer that a dimension is contrastive if the shape varies substantially across contexts but infer that it is not contrastive if the shape is largely the same across contexts. We operationalize a sound's context as 1) its (immediately) neighboring sounds, its prosodic position (whether it falls at a word or utterance boundary), and its quality (learned before length) or 2) its word frame, due to evidence that infants are sensitive to this information in their input (22–24, 27–31, 36). However, we are not tied to these particular factors. Any contextual factors that infants track, and that change the relative proportion of sound category membership, could work.

It is important to note that the learning outcome of this proposal is the same as in ref. 6 but differs from the learning outcomes of some phonetic learning theories that have arisen since then (21). In particular, the learning outcome here is whether or not an acoustic dimension is contrastive—i.e., whether it is used to distinguish multiple categories. The learning outcome of some other theories included this knowledge implicitly but often also included knowledge about what those categories were. Certainly listeners eventually learn about the categories, and a full learning account would need to eventually explain how that happens. However, the discrimination behavior infants exhibit in ref. 2 does not require them to have learned categories (2, 26, 37), so we follow the original literature and focus on how infants learn which dimensions are contrastive in their language.

In what follows, we ask whether the necessary signal for this learning account is present in naturalistic speech, that is, whether there are larger distribution shape changes across contrastive dimensions than noncontrastive dimensions. We focus on three test cases, which each involve some data in which vowel length is contrastive, and some data in which vowel length is not contrastive. We look at vowel length for two reasons. First, it has a largely agreed upon primary cue (duration) that can be easily extracted from any annotated corpus. Second, it is possibly the best known case of extreme overlapping categories that cannot be explained by previous theories like distributional learning (13). We focus on the test cases that have been problematic for past phonetic learning theories but argue in *Discussion* that this same approach to phonetic learning is likely to be effective across a wide range of languages and contrasts.

Results

Our results confirm that in spontaneous speech, there are more extreme distribution shape changes across contexts when a dimension is contrastive than when it is not.

For all of the corpora we study, we extract the acoustic distributions across a number of contexts and compare them pairwise, using Earthmover's distance (38), a commonly used metric of distribution shape difference (see *SI Appendix* for discussion of results using an alternative metric, KL divergence, instead). We operationalize "context" in two different ways, both of which rely on information that infants at the relevant age are sensitive to 1) a combination of prosodic position, neighboring sounds, and vowel quality (P + NS + VQ) and 2) word frames.

We first compare a spontaneous speech corpus of Japanese (which has a vowel length contrast) against a spontaneous speech corpus of French (which does not). We then test two spontaneous Dutch corpora. Dutch has the property that a subset of its vowels has a length contrast, whereas a different subset does not. Comparing the subset that has a contrast against the subset that does not allows us to control for any effects that may arise due to differences in how the French and Japanese corpora were collected and annotated. Two of our tests examine adult-directed speech (ADS) corpora because they allow us to test this proposal on large-scale, spontaneous speech corpora which do not exist for infant-directed speech (IDS), but we include results from a small corpus of infant-directed Dutch as well.

Japanese vs. French ADS. We first compared Japanese and French, defining context as a combination of prosodic position, neighboring sounds, and vowel quality (Fig. 2). Each data point

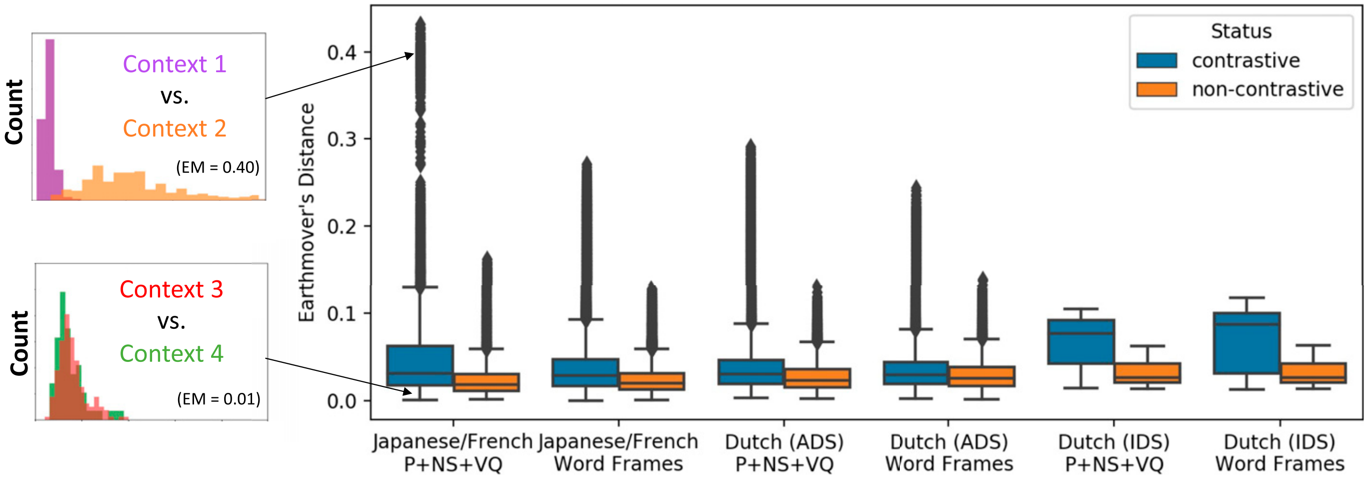

Fig. 2. Distribution of Earthmover's distances by test case. Each data point represents the pairwise Earthmover's distance (EM) between distributions from two different contexts (e.g., we show the comparison for Fig. 1*B*'s context 1 vs. context 2, which has a high Earthmover's distance, and Fig. 1*B*'s context 3 vs. context 4, which has a small Earthmover's distance). Across all test cases, the tail of the contrastive boxplot (*Left*) is longer than that of the noncontrastive boxplot (*Right*), suggesting that there are more extreme distribution shape changes across contexts when the acoustic dimension is contrastive. P + NS + VQ = prosodic position + neighboring sounds + vowel quality.

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

Table 1. Information about the Japanese contexts that drive the tail in the case of the P + NS + VQ analysis, including what percentage of vowels in that context are long, how many times that context occurred (count), and its frequency rank out of all contexts that occurred

Context	Percent long	Count	Frequency rank
Phrase-initial, word-final /e/	64.7	1,357	18
Phrase-initial, phrase-final /a/	56.7	255	95
Phrase-initial, phrase-final /e/	87.9	244	100

contributing to the boxplot represents the Earthmover's distance between a pair of contextual acoustic distributions. For example, the comparison between context 1 and context 2 in Fig. 1*B* has a high Earthmover's distance, whereas the distance for contexts 3 and 4 is much smaller because they are very similar.

The boxplot corresponding to Japanese (where vowel length is contrastive) has a much larger tail, extending upward toward large Earthmover's distances, than the boxplot corresponding to French (where vowel length is not contrastive). This means that as predicted, there are many more pairs of contexts that have substantially different shapes (like context 1 vs. context 2 in Fig. 1*B*) when there is a contrast than when there is not. The maximum distance, the mean distance, and the distance variance are all larger for Japanese than French (maximum = 0.43 vs. 0.16; mean = 0.05 vs. 0.02; variance = 0.003 vs. 0.0004). Analyzing the contents of the tail in Japanese reveals that the tail is driven by contexts that have a much higher percentage of long vowels than observed overall and that occur frequently in the input (see Table 1 for frequency counts and ranks of the contexts that drive the signal).

These same patterns hold when we continue looking at French vs. Japanese, but instead use word frames as contexts. As before, there are more contexts with more extreme distribution shape changes in Japanese than French (i.e., along contrastive than noncontrastive dimensions), as seen by the longer tail in the second pair of boxplots in Fig. 2. As before, the maximum distance, the mean distance, and the distance variance are all larger for Japanese than French (maximum = 0.27 vs. 0.12; mean = 0.04 vs. 0.02; variance = 0.001 vs. 0.0002).

Dutch ADS and IDS. To test our proposal using a within-language comparison, we compare the subset of Dutch vowels that do contrast in length and the subset of Dutch vowels that do not. We find that the predicted pattern still holds, and it holds for both ways of defining context and both the ADS and IDS corpora (Fig. 2). This confirms that the results are not merely an artifact of using different corpora, as in the French vs. Japanese case, but seem to reflect something inherent to the existence or nonexistence of categories along an acoustic dimension. It is worth noting that Dutch-learning infants would not be able to perform this exact analysis to learn whether there is a length contrast, because they would not yet know enough to separate the vowels into contrastive and noncontrastive subsets. We return to the issue of what a learning account might look like in *Discussion*. Meanwhile, we conclude from this analysis that the signal our account predicts exists in Dutch: contrastive dimensions differ from noncontrastive dimensions.

Despite the qualitative similarity in results across all test cases, the scale of the difference in tail length varies. For example, in the French vs. Japanese P + NS + VQ case, the maximum Earthmover's distance in the contrastive Japanese case is 0.43, whereas for the other ADS cases, the maximum is less than 0.3. In the Dutch IDS corpus, which only has 284 vowel tokens, the maximum is only around 0.1. One possibility is that these differences arise because of the large differences in corpus size. The Japanese vs. French corpora considered 132,037 tokens, and the Dutch ADS corpus considered 21,187 tokens, but the Dutch IDS corpus only considered 284 tokens.

Corpus Size Analyses. To test how corpus size impacts results, we used bootstrap samples to run each analysis 50 times for 10 different corpus sizes ranging from the size of smallest corpus (284) to the size of the largest corpus (132,037). This also allowed us to test how much the size of the tail varied and whether differences observed between contrastive vs. noncontrastive cases were meaningful. Fig. 3 shows these results when calculating the maximum Earthmover's distance across all 50 runs; analogous plots for mean are provided in *SI Appendix*. First, this analysis reveals that the differences observed are meaningful: across many runs, at large enough corpus sizes, the contrastive line is higher than the noncontrastive line. That being said, in the Japanese vs. French case, the difference does not emerge until around 2,000 vowel tokens have been observed, so input size does matter. Second, this analysis reveals that differences in scale may be partially, but not entirely, due to corpus size. When subsetted to the size of the Dutch ADS corpus, the Japanese vs. French word frame maximum matches the remaining ADS results. However, the results are less clear for Dutch IDS: subsetting the Dutch ADS corpus to the size of the Dutch IDS corpus yields results more in line with each other for the P + NS + VQ analysis but less so for the word frames analysis.

From a learning perspective, this means that an ideal learner would need to observe around 2,000 vowel tokens and track the acoustic distribution within the 20 most frequent contexts in order to observe the difference (although we discuss potential ways to reduce the memory demands of the proposal next as well as in *Discussion*).

Relaxing Our Assumptions About Infants' Knowledge. Our analyses so far have been conducted assuming that 1) infants can track acoustic distributions across combinations of three contexts (prosodic position, neighboring sounds, and vowel quality); 2) infants can perfectly identify neighboring consonants; and 3) infants can perfectly segment words from speech. Although some of these assumptions have yet to be tested (e.g., we do not know whether infants can track distributions along multiple contextual dimensions), we know that others are likely overestimating infants' prior knowledge when learning about contrastive dimensions. Here we show that the same qualitative results still emerge even when we weaken these three assumptions.

The first assumption we revisit is whether infants can track distributions across combinations of contexts (i.e., prosodic position, neighboring sound, and vowel quality). We test what happens when we study each of these three factors individually. Fig. 4 shows that especially for prosodic position, but also for the other factors, the same patterns, for the most part, emerge. While it is still quite conceivable that infants track distributions across combinations of contexts, these results suggest that this need not be the case in order for our account to be successful. It suffices to track distributions across individual contexts (e.g., word frames or prosodic position). From a learning perspective, this means that infants would only need to be tracking the acoustic distribution across ∼4 to 10 contexts (rather than ∼20 as observed before), and these contexts include extremely prominent contexts (e.g., utterance-final and utterance-initial tokens).

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

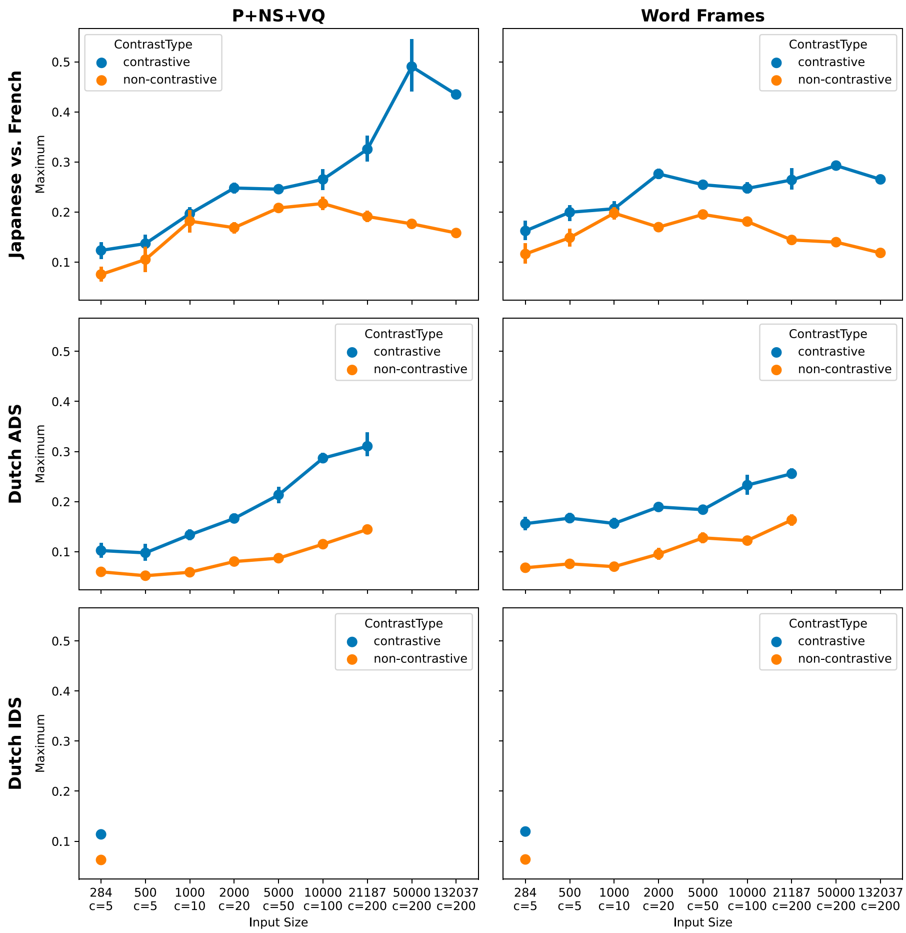

Fig. 3. Bootstrapped analyses reveal that observed differences between contrastive (top blue line) vs. noncontrastive (bottom orange line) dimensions are meaningful but that input size does matter. "C" refers to the number of contexts included in the analysis. The maximum input size for which data are shown depends on the corpus size: 284 for Dutch IDS, 21,187 for Dutch ADS, and 132,037 for Japanese vs. French.

The next two assumptions we revisit are whether infants can perfectly encode the identity of neighboring sounds and whether they have a solid enough grasp on word segmentation to have access to the prosodic position information we use. On one hand, prosodic boundaries are one of the first signals that infants are sensitive to (28, 29); however, on the other hand, there is concurrent evidence that infants make missegmentation errors (39–41). To address these two assumptions, we test what happens when we rerun the P + NS + VQ simulations, with neighboring sounds defined by their broad class (i.e., stop, fricative, vowel, approximant, etc.) rather than their particular identity (e.g., /k/, /g/, and /b/) and with noise added to the prosodic information (we simulate a 20% error rate here, such that 20% of the time, the infant misrepresents the prosodic position of the vowel, but the results generalize across error rates). Fig. 5 shows that these differences do not qualitatively change the results, suggesting that even with a more rudimentary grasp on contextual factors, infants could still use this method to learn the sound contrasts.

Overall, we show that the necessary linguistic knowledge and capabilities can be considerably reduced and yet the correct finding still emerges. This suggests that this finding is a robust one that immature learners could learn from even in noisy learning environments.

Analyses with Long Vowels Removed. Finally, to test whether these results arose because of the contrast, we removed all vowels labeled as long from the corpora and reran the same analyses. We predicted that removing the long vowels would cause the tail for the contrastive dimension to disappear, such that the results for the contrastive dimension with long vowels removed (i.e., with the contrast artificially removed) would resemble those of the noncontrastive dimension. As can be seen in Fig. 6, in Japanese and in Dutch IDS (the two hand-annotated corpora we use), the tail disappears or is reduced once the long vowels are removed, suggesting that it is at least partially the presence of the long vowels that causes these large changes in distribution shape. However, this is not case in Dutch ADS: the contrastive dimension still has a longer tail even when long vowels are removed. While this is not predicted by our account, there are a number of reasons why we may observe this result here. One possibility is that the annotations are imperfect. The remaining datasets studied here include hand-corrected segmental annotations, while the Dutch ADS data included force-aligned annotations which were not validated for their duration. Another possibility is that it has to do with the fact that these data were phonetically annotated, rather than phonemically annotated. That is, some phonemically long vowels were marked as being short. In fact, ref. 42 reports that as many as 20% of word tokens that had long vowels underwent a shortening process. Especially combined with the fact that these data were automatically annotated, this could mean that we are unable to actually remove all long vowels and that some long vowels remain which are driving the differences observed. Finally, another possibility is that the presence of long vowels in a context changes the distribution of short vowels. For example, if a context is 50% short vowels and 50% long vowels, then the short vowels may be pronounced with shorter durations than in a context

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

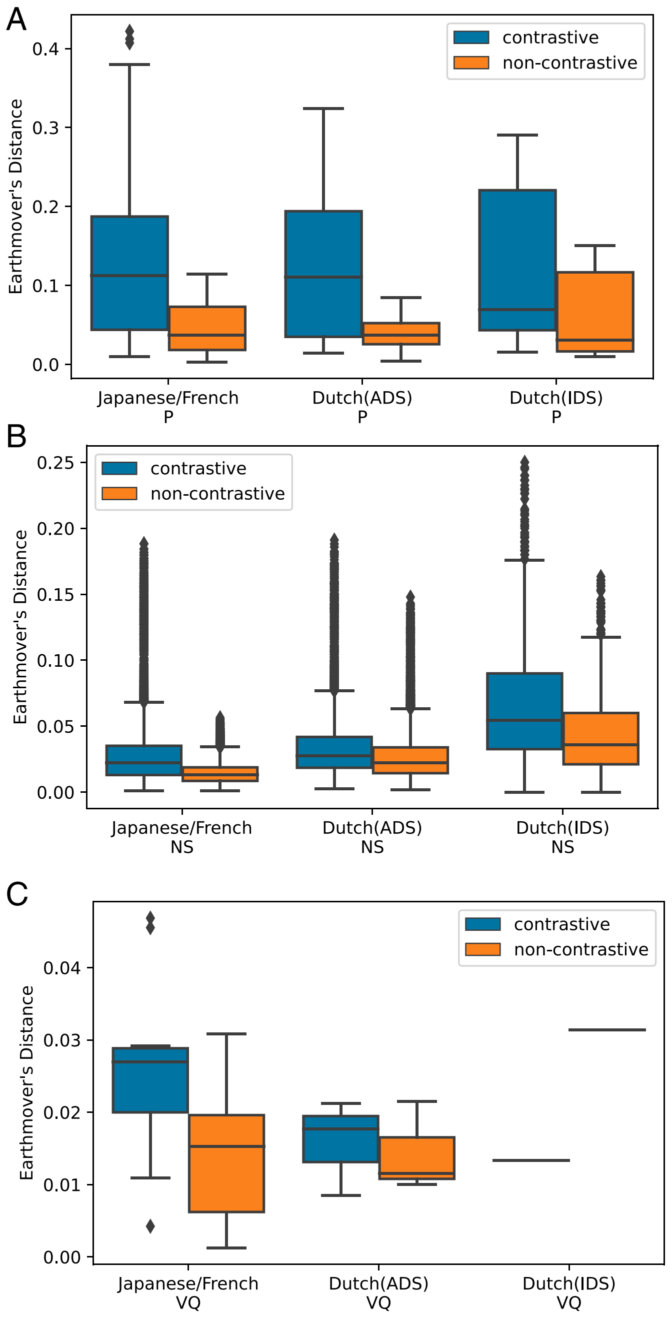

Fig. 4. Results are similar when we relax our assumptions about infants' knowledge. Here, instead of studying a combination of factors, we study (*A*) prosodic position (P), (*B*) neighboring sounds (NS), and (*C*) vowel quality (VQ) individually (left boxplot in each panel shows contrastive dimension; right boxplot in each panel shows noncontrastive dimension). While less clear for vowel quality, we see the same pattern of results in the prosodic position and neighboring sounds only cases.

with 90% short vowels and only 10% long vowels (i.e., to better differentiate the vowel types). If this is the case, then even if we remove the long vowels, we should expect to see differences in the short vowel distributions across contexts. However, this finding is nonetheless different from what our account predicts, and more work should be done with hand-annotated corpora to see whether this finding remains and, if so, what is driving it.

Nonetheless, across the board, we do observe that the contrastive dimensions have a longer tail than the noncontrastive case, suggesting that contrastive dimensions exhibit more extreme distribution shape changes across contexts (even though all distributions are unimodal) than noncontrastive dimensions. These results show that contrastive dimensions look different from noncontrastive dimensions and that infants would learn the correct generalization about their language by using this signal. As such, this is a piece of signal that has been shown to successfully differentiate contrastive and noncontrastive dimensions using spontaneous speech.

Discussion

In this paper, we proposed an account for how infants could learn which acoustic dimensions of their language are contrastive. The idea is that infants track the frequency distributions of sounds in different contexts and that they learn that an acoustic dimension is contrastive if the distribution shape along that dimension varies substantially across different contexts. For this account to work, it needs to be the case that the distribution shape varies more across contexts when a dimension is contrastive than when it is not. We tested this prediction in three test cases, with two different definitions of context. Across the board, our results show that the distribution shape along an acoustic dimension changes more across different contexts when that dimension is contrastive than when it is noncontrastive. This is a signal that differentiates contrastive and noncontrastive dimensions, and it is the kind of signal that listeners are likely to be sensitive to. As such, this is a phonetic learning account that has been shown to work on spontaneous data and suggests that infants could be learning which acoustic dimensions are contrastive after all. In the remainder of the paper, we discuss the promise and open questions of this proposal, including its generalizability, evidence on whether infants have the necessary sensitivities, and how infants could use this signal to learn.

Generalizability. We considered the test cases of Japanese and Dutch vowel length because they are famous problem cases for many of the phonetic learning theories that already exist (13, 43). However, they are unique contrasts in a number of ways. They have low functional load in that they are not frequently used to distinguish different meanings and they have particularly overlapping acoustic distributions. In addition, the Japanese contrast is primarily signaled by one acoustic cue (i.e., duration), while most contrasts are signaled by more (44, 45), and as around 90% of Japanese vowels are short, it is less balanced than many other contrasts. Future work will need to test the generalizability of this

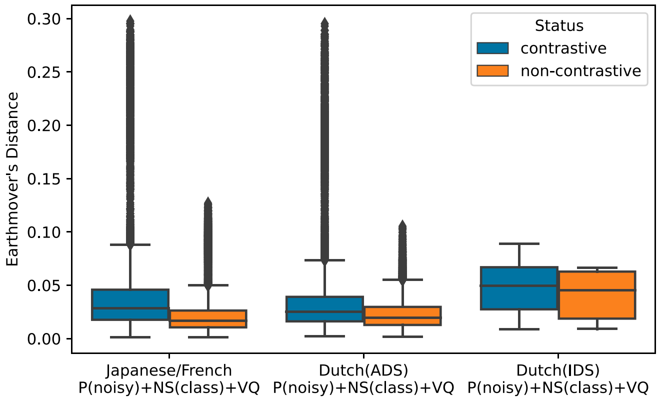

Fig. 5. Results are similar when we relax our assumptions about infants' knowledge. Here we define neighboring sounds by their broad class (stop, fricative, etc.) and add noise to the prosodic position with a noise rate of 20% (left boxplot shows contrastive dimension; right boxplot shows noncontrastive dimension).

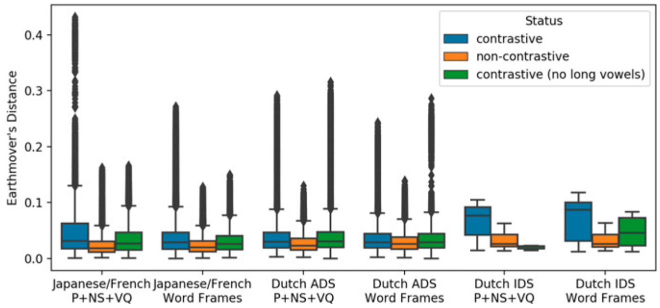

Fig. 6. Results from artificially removing long vowels. Within each case, the left boxplot corresponds to the contrastive dimension, the middle boxplot corresponds to the noncontrastive dimension, and the right boxplot corresponds to the contrastive dimension with long vowels removed. We observe that the tail length is reduced for the Japanese/French and Dutch IDS analyses, however, against predictions, not for the Dutch ADS analysis.

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

proposal. Nonetheless, we think it is likely that this signal will generalize to other contrasts because the signal we illustrate in Fig. 1*B* does not stem from these idiosyncrasies. Rather it arises because of a handful of contexts that have particularly different frequency distributions, and we think these exist because of properties of language that phonologists have argued are universal (32).

All languages are thought to have phonotactic or cooccurrence constraints. For example, for the English contrast [n]–[ŋ] (e.g., "sin" vs. "sing"), [n] can occur at the beginning or end of a syllable, while [ŋ] can only occur at the end of syllables. This means that the relative proportion of [n]s and [ŋ]s will change across those contexts and could show patterns like in Fig. 1. Similarly, sometimes sounds will be pronounced differently based on their context. For example, if we consider the contrast between [n] and [m], the [n] sound never occurs before the sounds [p,b] but is instead pronounced as [m] in those contexts (as in "impossible"). These types of phonological alternations also create differences in which sounds occur in which contexts and are cross-linguistically widespread (32–34). Finally, there are systematic regularities based on the words in the language. For example, for the [p]–[b] contrast in English, [b] is more likely in the word frames _aby and _ar ("baby" is a word, but "paby" is not, and "bar" is more frequent than "par"), whereas [p] is more likely in the word frames _lay and _in ("play" is a word, but "blay" is not, and "pin" is more frequent than "bin"). Taken together, phonotactics, phonological alternations, and word regularities all create systematic regularities in the contexts that sounds occur in, such that different sounds occur in different contexts and different contexts are made up of different ratios of sounds. This is currently thought to be true of all contrasts—even those that are equally balanced, have higher functional load, or are multidimensional—and will be the case in ADS as well as IDS. Indeed, ref. 35 showed the context is similarly predictive of which category a sound belongs to in both ADS and IDS.

Nonetheless, while our results are promising, they ultimately come from one acoustic dimension, one contrast type, and only three languages. Furthermore, there are additional complexities that we have not considered here, like how extralinguistic factors (e.g., speaking style/register, emotional content of speech, and speech rate) could affect the signal and how this proposal interfaces with other learning strategies that infants have been shown to use (e.g., using visual and referential information).

To test generalizability, future work should replicate these findings on other datasets as they become available (we have made our code publicly available to facilitate this effort) (46). Particular focus should be placed on replicating these results on a large ecological corpus of IDS, replicating these results on other contrasts whose acquisition is difficult to explain [e.g., Filipino nasals (47)], and studying Dutch to understand why removing long vowels does not always change the result, as well as replicating these results in languages where the identity of a particular sound cannot be predicted as well from its context, due to having fewer phonotactic and other systematic restrictions (48). It will also be important to test this proposal on contrasts that are signaled by multiple acoustic dimensions (as the contrasts we study here are unique in primarily being signaled by one, duration). In order for this account to work in those cases, we would need to observe the signal we report here along at least one of the acoustic dimensions that signal the contrast (e.g., along VOT or F0 for stop contrasts). Alternatively, rather than operating over individual acoustic dimensions, infants could search for this signal along composite acoustic dimensions that are discovered from the input by combining dimensions that are highly correlated in the input (e.g., as discussed in refs. 37, 49). Finally, we note that Earthmover's distance can be calculated over multidimensional distributions, so it should be possible to scale this approach up to multidimensional contexts if so desired; however, for that to work, the infant would have to first identify the relevant combinations of acoustic dimensions they should focus on. Overall, it will be important to replicate these findings across highly variable corpora that adequately represent the full range of speech types that infants could encounter.

Finally, it will also be important to test that these results do not overgeneralize and wrongly label noncontrastive dimensions as contrastive. Here it will be particularly important to test behavior on allophonic variation, where a particular sound is realized differently depending on the context it occurs in. While this pattern is similar to that of different phonemes occurring in different contexts, our analyses provide preliminary evidence that allophonic variation is not labeled as contrastive. In particular, in French, vowel length varies allophonically (vowels are lengthened depending on the following consonant) (50), yet our analyses reliably treated French (allophonic variation in vowel length) differently from Japanese (phonemic variation in vowel length). This suggests that this method may correctly differentiate contrastive and allophonic variation, although it will be important to study this further and understand how it does so (if it does). We offer two speculative reasons why allophonic variation may not be detected to the same degree as phonemic variation, although they will need to be tested. First, this proposal relies on contrasting sounds having sufficiently different acoustic distributions (so that changing the relative proportion of the sounds changes the shape of the distribution). It is possible that allophonic variation changes how a sound is produced to a lesser degree than phonemic variation, although this may be difficult to assess. A second possibility is that allophony often, although not always, affects all of the sounds produced in a particular context (e.g., the duration of all French vowels is affected by neighboring consonants). This may lead to a shift in distribution between contexts, without a change in shape, which may lead to smaller distribution shape changes as measured by Earthmover's distance. Certainly, more research studying how this proposal handles allophony will be crucial.

Could Infants Do This? This proposal places higher computational and memory demands on infant learners than many past theories have (e.g., distributional learning). Infants would need to be able to track distribution shapes across many contexts and then compare their shapes pairwise.

Could infants track distributions across different contexts? Although this has not been tested, we know that listeners track the shape of frequency distributions overall and use this for phonetic learning and processing. For example, infants make different phonetic inferences depending on whether the distribution they hear is bimodally or unimodally shaped (6). Another study showed that adult listeners are sensitive to the variance of the sound distributions they are exposed to, another property of a distribution's shape (51). In that experiment, adults categorized sounds differently depending on whether they heard a distribution with high or low variance, although this has not been tested in infants.

In addition, we know that listeners are sensitive to the context of a sound and use it for phonetic learning and processing. Both toddlers 12 mo and older and adults have been argued to track acoustic distributions across speakers (which can be thought of as a context), can adapt to speakers who have different accents (i.e., different distributions of sounds) (52–58), and mirror the speech of their interlocutors. In addition, infants are sensitive to phonotactics (59), as well as phonological alternations—the fact that sounds tend to be pronounced differently in different contexts (60, 61). Additionally, multiple studies have shown that infants use the word frame of a sound in phonetic learning (22, 24). That is, infants seem to assign acoustically similar sounds to different categories if they occur in different word frames, suggesting that infants can track the context that a sound occurs in and use it for phonetic learning. Adult speech perception is affected by contextual factors, like neighboring sounds. For example, ref. 62 showed that Japanese perception of whether the final vowel in CoC'V was phonemically short or long depended on the identity of both C and C'. Furthermore, ref. 63 showed that adults continually track how informative particular acoustic cues are and will selectively reweight these cues in some contexts but not others (e.g., when categorizing /b–p/ in beer–pier but not for /d–t/ in deer–tier and vice versa). Indeed, listeners may even track information across contexts defined by multiple factors. For example, in English, to learn whether a stop consonant will be aspirated or not, the speaker must track whether the sound is voiceless or voiced (akin to vowel quality), whether it occurs at the start of a stressed syllable or not (akin to prosodic position), and what the neighboring sounds are and then notice the change in pronunciation/distribution across those different contexts.

Taken together, while these results do not provide direct evidence that infants track distribution shapes across contexts, they suggest that listeners can track complex statistical regularities across complex contexts and use them in real-time phonetic learning and processing.

Could infants compare distribution shapes? Finally, the last skill necessary for the proposal is for infants to be able to compare distributions. While this has again not been directly tested, one possible clue is that listeners seem to reweight acoustic cues depending on how variable/informative they are, with cues that have narrower distributions being more informative than cues with wider distributions (51, 64, 65). Another possible clue is that toddlers and adults are able to identify when they need to adapt their representations to speech they hear. Being able to identify an accent implies that listeners can identify when the speech they are hearing differs from the speech they usually hear, a computation that is likely to involve tracking at least some properties of the distributions (66).

Overall, given infants' demonstrated sensitivity to distribution shape and to changes across context, there is good reason to believe that infants could be sensitive to the type of distributional information that our account assumes, but future work should test whether infants/listeners can track distributions across different contexts (defined by one or more contextual factors) and compare distributions' shapes. In addition to testing whether infants can, in theory, perform the computations this account requires, future work should also test whether infants actually use them to learn about contrastiveness in the way we propose here. One approach would be to test this experimentally, by exposing infants to acoustic distributions that differ or remain the same across contexts and seeing if this affects their learning/behavior. Another approach would be to use cross-linguistic corpora to identify contrasts that should be easier/harder to learn according to our proposal and compare that against age of acquisition and speech perception/production data. For example, controlling for degree of acoustic overlap, this theory would predict that contrasts that have stronger phonotactic restrictions or that are more predictable (i.e., it is easier to predict which member of the contrast occurred based on the context it occurred in) should be easier to learn through this method. These approaches will allow us to overcome the next big hurdle for this account, which is determining whether infants use this signal to learn.

Reducing the Computational Complexity of the Proposal. It is also possible that the memory and computational restrictions of the proposal could be reduced. On one hand, we saw that considering individual contexts (e.g., just the most frequent word frames or just prosodic position) was still effective, as was introducing some parsing errors and considering broader segment classes rather than individual neighboring sounds. On the other hand, online approximations or metrics that do not require the whole distribution to be tracked in order to get a measure of distribution shape distance could also reduce the computational and memory complexity of the proposal. For example, rather than exactly representing the distribution, this proposal could operate over a compressed representation of the distribution that keeps track of how many points fall within larger bins/bands (similar to reducing the number of bins in a histogram). It is likely that we would still observe the critical pattern even with this less detailed representation of distribution shape, and as this only requires keeping track of one number per bin (its count), it could reduce the size of the representation of a distribution to just 5 to 10 numbers. Even considering all 200 of the most frequent contexts, this could involve storing as few as 1,000 numbers. In addition, there may be a way to zero in on the contexts that yield the necessary signal without doing all of the pairwise comparisons represented in the boxplots. It is possible that the key contexts that drive the signal are overrepresented in the outliers of the overall distribution (i.e., particularly short or particularly long vowels could be more likely to occur in a context that drives the tail). If this were true, infants could arrive at the same signal we observe here, by focusing in on the contexts of outlier sounds, rather than tracking the distribution across all contexts. Finally, another possibility is that infants could compare contextual distributions using higher-order measures of distribution shape (e.g., variance) rather than tracking the entire distribution. The fact that the distribution shape changes across contexts could also mean that the variance of the distributions changes across contexts. If so, infants could pick up on this difference without encoding the entire distribution across contexts.

Moving from Signal to Learning Account. The data we report plot Japanese and French side by side, but most Japanese infants do not get French input to compare against. Assuming that this pattern generalizes to other contrasts and that infants have the necessary sensitivities to detect this signal, how could infants actually use it to learn?

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

One possibility is that infants use a built-in threshold to determine whether a dimension is contrastive: if the metric (this could be something like the average, range, variance, or maximum Earthmover's distance) exceeds the threshold, they learn the dimension is contrastive; otherwise, they learn that it is not. Another possibility is that infants compare against other acoustic dimensions of their own language (instead of against other languages, as we did). If these metrics turn out to be larger for all contrastive dimensions than all noncontrastive dimensions, infants could easily separate contrastive vs. noncontrastive dimensions.

One complication for this possibility, however, is that the metrics we report are sensitive to the scale of a dimension, making it difficult to compare across dimensions with different scales (e.g., formants vs. duration). To overcome this problem, we tried z-scoring the dimensions but found that the key effect partially disappeared: the effect was retained for French vs. Japanese and when considering only the most frequent contexts, but when more contexts were considered, the pattern reversed in Dutch ADS. This happened because z-scoring is sensitive to variance, and as there was more variability along the contrastive than noncontrastive dimensions, z-scoring led to artificially lowered Earthmovers' distances along the contrastive (more variable) dimension. Nonetheless, a method that standardizes the scales with less sensitivity to the overall variance could allow for comparison across dimensions.

Finally, given the overall distribution along an acoustic dimension, infants could have a probabilistic model of how different they should expect distributions to be across contexts, depending on whether the overall distribution is made up of one vs. two categories. With this, they could compare how likely a one-category vs. two-category solution is to have generated the observed Earthmover's distances (controlling for phonological processes, it would be unlikely for one category to produce extremely different distribution shapes as in Fig. 1*B*).

One issue that should be considered in future work is whether learners consider aggregate distributions across all of the vowel qualities when comparing distribution shapes across contexts. In our analyses of Japanese and French, we have assumed that they do, and in fact, vowel quality is one of the contextual factors we analyze. However, this creates a potential problem for Dutch because only a subset of Dutch vowels contrast in length. If Dutch infants were to initially rely on the aggregated distribution shape comparison approach we put forth for Japanese, this might lead them to conclude that vowel duration is a contrastive dimension that they should tune into (as the context pairs that showed high Earthmover's distance would still be in this analysis), but they would not realize that only some of the vowels contrast. It is possible that they could later learn which specific vowel qualities contrast in length. It is also possible that Dutch infants use a different strategy entirely for discovering the vowel length contrast: those Dutch vowels that contrast in length also contrast in vowel quality (e.g., [ɑ] vs. [aː] contrast in Dutch, but [a] vs. [aː] contrast in Japanese). Having already separated [ɑ] from [aː] using their vowel qualities, Dutch infants could simply notice that these vowels differ systematically in their durations, without doing any distribution shape comparisons of the type we propose here. Ultimately, we remain agnostic as to how exactly infants learn. Given the complexity of the task infants are faced with as well as past experimental findings showing infants use many types of information in phonetic learning, it likely involves a combination of strategies (e.g., using word-level, visual, referential, and other distributional information in addition to the types of analyses we report here). Having established that a signal exists in naturalistic data, we hope future research will investigate how this signal might best be used for learning and how it integrates with other promising accounts.

Conclusion

Infants need to learn which acoustic dimensions of their language are contrastive in order to learn the sound system of their language. However, we still do not know what aspects of naturalistic input provide the necessary signal for them to do so. In this paper, we propose a potential account for how infants learn this and show that there is a signal about whether a dimension is contrastive in noisy, spontaneously produced input. This account is particularly promising for two reasons. First, the signal that we pick up on is a direct consequence of multiple categories exhibiting properties that hold true across most languages, so we think it is likely that this result will generalize to other contrasts. In addition, the signal is something that even infants may be sensitive to. Past work has shown that infants track the shapes of overall frequency distributions and know about how sounds are likely to sound in different contexts (6, 60, 61). Adults have been shown to track distributions across situations (i.e., across different talkers) (52). In conclusion, we show that even when two sounds overlap acoustically, the fact that they occur in different contexts leaves signal to their contrastiveness. These results provide initial support for a phonetic learning account that works on highly acoustically variable spontaneous speech.

Materials and Methods

Methods. For each test case, one of the datasets (contrastive or noncontrastive) was larger than the other. To correct for this, we only considered the first N tokens of the larger dataset, where N was the size of the smaller dataset. We extracted the duration, the primary acoustic cue to length, of each vowel token in seconds, rounding to the same degree of precision. In addition, we extracted all contextual information that was available across all of the corpora we study and that infants of the relevant age are sensitive to the following.

- Vowel quality: For Japanese, this was /a/, /e/, /i/, /o/, or /u/. For French, this was /a/, /e/, /i/, /o/, /u/, /y/, /ø/, /ã/, /õ/, or /ø̃/. For Dutch, this was /ɑ-a/, /ɔ-o/, /œ-ø/, /ɛ-e/, /ɪ/, /u/, /y/, or /i/. The first four listed pairs are differentiated by quality and length, but we do not incorporate these vowel quality differences into this paper. Vowel quality is thought to be learned before vowel length (27).
- Prosodic position: We represented prosodic position (a vowel's position relative to prosodic boundaries) with four indicator values: 1) whether the vowel was word-initial or not, 2) whether the vowel was word-final or not, 3) whether the vowel was phrase-initial or not, and 4) whether the vowel was phrase-final or not. Infants have been shown to be sensitive to prosodic boundaries quite early (28, 29).
- Neighboring sounds: We extracted the identity of the immediately previous sound and the immediately following sound, as labeled by the phonetic transcription, ignoring length information. Again, vowel length contrast is thought to be learned later than other types of contrasts (27).
- Word frame: We extracted the word frame that the vowel occurred in, excluding all length information. For example, one word frame could have been [b_i_ru], which would include both [biːru] and [biru]. We chose to include word frames as infants know and can segment words early (30, 31, 36, 67) and use word frames in phonetic learning (22–24).

We looked at two main ways of defining context, although we do not have any commitments about which contexts infants would compute over. In the first way of defining context, we used a combination of vowel quality, prosodic position, and neighboring sounds (e.g., /o/ vowels that follow a /t/ and precede a /k/ that are word- and phrase-internal), although we also consider each of these three contextual factors individually. This combined set of factors corresponds to the subset of factors considered in ref. 35 that were available for the corpora

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.

we study and that infants are most sensitive to. In the second way of defining context, we used word frames, as has been done in refs. 14 and 21, among others.

Because most contexts occur very infrequently, we looked at a subset of all possible contexts. We subsetted the contexts in two qualitatively different ways: either by taking the top X most frequent contexts or by taking all contexts that had at least N tokens, varying X and N. Results were qualitatively similar in all cases, so we present results from including the 200 most frequent contexts for the French vs. Japanese and Dutch ADS analyses and only the 5 most frequent contexts for the Dutch IDS analysis due to its much smaller size. Once we had the contexts, we extracted the vowel duration frequency distributions in each context (examples shown in Fig. 1).

We compared the shape of each pair of contextual frequency distributions, using a metric known as Earthmover's distance or Wasserstein distance (37, 68), which is commonly used to measure the difference in shape between two distributions (see *SI Appendix* for methods and results using KL divergence instead). Earthmover's distance is often talked about in terms of two piles of dirt, which represent the two distributions being compared. In this context, Earthmover's distance can be thought of as the minimum cost of turning one earth pile into the other, where cost corresponds to a combination of the amount of earth being moved as well as the distance it has to be moved. In other words, the distance is the minimum average distance a piece of dirt will have to be moved in order to turn one pile into the other. A higher distance means there was a greater shape mismatch. We plot the distribution of Earthmover's distances and report its mean, variance, and maximum.

Having found that the distribution of Earthmover's distances for contrastive dimensions had a longer tail than for noncontrastive dimensions, we qualitatively analyzed the contents of this long tail to determine which individual contexts led to the pattern observed. We identified the contexts that showed up most frequently in the tail and analyzed how frequent they were (both in terms of absolute count and their frequency ranking relative to all contexts), as well as what the relative frequency of short and long vowels was in each of these key contexts.

To assess if the reliability of these differences between contrastive and noncontrastive dimensions were reliable, we used bootstrap statistics. We sampled (with replacement) particular vowel tokens with their contexts, creating a new contrastive and a new noncontrastive dataset. We then recalculated Earthmover's distances across the contexts using these bootstrap samples, repeating the process 50 times. We plot the maximum Earthmover's distance with the SD, which allows us to observe the reliability of these differences. To study the effect of the input corpus size, we varied the number of vowel tokens sampled from 284 vowel tokens (the size of the Dutch IDS corpus) to 132,037 (the size of the French vs. Japanese corpora).

In our final simulation, we relaxed our assumptions about infants' prior knowledge. First, while previous analyses used neighboring segment identity directly (e.g., /k/, /t/, and /s/), this simulation only used the segment's broad class (e.g., stop, fricative, and vowel), which infants are more sensitive to. Second, to simulate imperfect segmentation, we added noise to the prosodic position factor. Prosodic position is represented with four indicator values (depending on whether the vowel in question is word/utterance-initial and word/utterance-final). To add noise, we changed 20% of these values (making sure that the resulting prosodic position was real; e.g., sounds considered to be utterance-final were necessarily also considered to be word-final). We then used the same procedure from above with these updated factors.

Corpora. The French vs. Japanese analysis compared the Corpus of Spontaneous Japanese (CSJ) against the Nijmegen Corpus of Casual French (NCCFr). The Dutch analyses looked at the Ernestus Corpus of Spontaneous Dutch (ECSD) (ADS) and the Levelt/Fikkert corpus (IDS).

CSJ. The CSJ is a large corpus of spontaneously produced ADS (69). Around 90% of the speech consists of spontaneously produced monologues about academic fields, their favorite memory, and so forth. The remaining 10% consists of spontaneous dialogues either in free conversation with the experimenter or engaged in a task. Our analysis focuses on the core portion of the corpus, which was force-aligned and hand-corrected with the segmental information required for our analyses (see ref. 69 for more details). The core portion consists of 811,731 total vowel tokens of which 89.1% are phonemically short and 10.9% are phonemically long, but only the first 132,307 tokens were used to match the size of the French corpus.

NCCFr. The NCCFr is a corpus of spontaneously produced ADS (70). Unlike the CSJ, however, the NCCFr consists exclusively of conversational speech between close friends. Topics included upcoming examinations, travel plans, an ongoing strike, and so forth. The corpus consists of speech by 46 French speakers and includes 132,307 vowel tokens. The corpus was orthographically transcribed by two professional transcribers. The corpus was transcribed at the segmental level by Martine Adda Decker (personal communication with M. Ernestus, 14 January 2019).

ECSD. The ECSD consists of adult-directed, conversational speech, with speakers talking with a friend, at first freely and then engaged in a task-oriented discussion (42). The corpus has speech by 20 different speakers and includes 60,955 tokens with a length contrast and 21,187 tokens without. Professional transcribers created an orthographic transcription of the interactions, which was manually aligned to the speech. The corpus was also phonetically transcribed using a forced alignment model (details can be found in ref. 71). Validations revealed a 14% discrepancy between manual annotations and forced-aligned annotations, which is in the range of human disagreement. However, these analyses did not directly validate durational information, so it is unclear how accurate annotations of the start and end points of the phones are. This could introduce some noise into our analyses as it could affect how accurate the vowel durations are and how accurately we can determine which word a vowel belonged to.

Fikkert/Levelt/Swingley IDS corpus. We also tested our account on a corpus of Dutch IDS collected by Fikkert (72) and Levelt (73). The annotated portion of this corpus is small: it contains a total of 300 utterances, with a total of only 1,296 vowel tokens, but each of the contrastive and noncontrastive datasets had to be subsetted to 284 to make equally sized subsets. The corpus consists of naturalistic longitudinal speech interactions with one child (Catootje) aged 1 year 10 months. The corpus was transcribed at the word level. Time-aligned phonetic annotations were created by Dan Swingley (DS) (14). Given the transcriptions, the speech toolkit HTK (Hidden Markov Model Toolkit) (74) was used to estimate the boundaries of the phones using the HVITE forced-alignment tool. The output of the forced-alignment tool was manually corrected by DS, a speaker of Dutch. K.H. time-aligned the word-level transcription to the time-aligned phonetic transcriptions based on the location of the phones in Praat (75).

Data, Materials, and Software Availability. Code for all analyses has been deposited in GitHub (https://github.com/khitczenko/contextual-dl) (46). Previously published data were used for this work (14, 42, 69, 70, 72, 73).

ACKNOWLEDGMENTS. This work was supported by NSF grants IIS-1421695 and BCS-1734245 and National Science Foundation Research Traineeship 1449815. We thank Thomas Schatz, Mirjam Ernestus, Dan Swingley, and Francisco Torreira for help acquiring corpora, as well as Micha Elsner, Bill Idsardi, Jeff Lidz, Reiko Mazuka, Rochelle Newman, Thomas Schatz, Dan Swingley, and the reviewers and editor for helpful feedback and discussion.

1. J. F. Werker, J. H. Gilbert, K. Humphrey, R. C. Tees, Developmental aspects of cross-language speech perception. *Child Dev.* **52**, 349–355 (1981).
2. J. F. Werker, R. C. Tees, Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* **7**, 49–63 (1984).
3. P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, B. Lindblom, Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* **255**, 606–608 (1992).
4. S. Tsuji, A. Cristia, Perceptual attunement in vowels: A meta-analysis. *Dev. Psychobiol.* **56**, 179–191 (2014).
5. C. R. Narayan, J. F. Werker, P. S. Beddor, The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Dev. Sci.* **13**, 407–420 (2010).
6. J. Maye, J. F. Werker, L. Gerken, Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* **82**, B101–B111 (2002).
7. J. Maye, D. J. Weiss, R. N. Aslin, Statistical phonetic learning in infants: Facilitation and feature generalization. *Dev. Sci.* **11**, 122–134 (2008).
8. K. A. Yoshida, F. Pons, J. Maye, J. F. Werker, Distributional phonetic learning at 10 months of age. *Infancy* **15**, 420–433 (2010).
9. A. Cristia, Can infants learn phonology in the lab? A meta-analytic answer. *Cognition* **170**, 312–327 (2018).
10. R. K. Reh, T. K. Hensch, J. F. Werker, Distributional learning of speech sound categories is gated by sensitive periods. *Cognition*, **213**, 104653 (2021).
11. G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, S. Amano, Unsupervised learning of vowel categories from infant-directed speech. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 13273–13278 (2007).

12. B. McMurray, R. N. Aslin, J. C. Toscano, Statistical learning of phonetic categories: Insights from a computational approach. *Dev. Sci.* **12**, 369–378 (2009).
13. R. A. Bion, K. Miyazawa, H. Kikuchi, R. Mazuka, Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS One* **8**, e51594 (2013).
14. D. Swingley, Learning phonology from surface distributions, considering Dutch and English vowel duration. *Lang. Learn. Dev.* **15**, 199–216 (2019).
15. T. J. Vance, *An Introduction to Japanese Phonology* (SUNY Press, 1987).
16. C. Narayan, The acoustic–perceptual salience of nasal place contrasts. *J. Phon.* **36**, 191–217 (2008).
17. D. Swingley, C. Alarcon, Lexical learning may contribute to phonetic learning in infants: A corpus analysis of maternal Spanish. *Cogn. Sci. (Hauppauge)*, 10.1111/cogs.12620 (2018).
18. F. Adriaans, D. Swingley, Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *J. Acoust. Soc.* **141**, 3070–3078 (2017).
19. F. Adriaans, D. Swingley, "Distributional learning of vowel categories is supported by prosody in infant-directed speech" in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Miyake, Peebles, Cooper, Eds. (Cognitive Science Society, Austin, TX, 2012), pp. 72–77.
20. B. Dillon, E. Dunbar, W. Idsardi, A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cogn. Sci.* **37**, 344–377 (2013).
21. N. H. Feldman, T. L. Griffiths, S. Goldwater, J. L. Morgan, A role for the developing lexicon in phonetic category acquisition. *Psychol. Rev.* **120**, 751–778 (2013).
22. N. H. Feldman, E. B. Myers, K. S. White, T. L. Griffiths, J. L. Morgan, Word-level information influences phonetic learning in adults and infants. *Cognition* **127**, 427–438 (2013).
23. D. Swingley, Contributions of infant word learning to language development. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 3617–3632 (2009).
24. E. D. Thiessen, The effect of distributional information on children's use of phonemic contrasts. *J. Mem. Lang.* **56**, 16–34 (2007).
25. H. H. Yeung, L. M. Chen, J. F. Werker, Referential labeling can facilitate phonetic learning in infancy. *Child Dev.* **85**, 1036–1049 (2014).
26. T. Schatz, N. H. Feldman, S. Goldwater, X. N. Cao, E. Dupoux, Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2001844118 (2021).
27. Y. Sato, Y. Sogabe, R. Mazuka, Discrimination of phonemic vowel length by Japanese infants. *Dev. Psychol.* **46**, 106–119 (2010).
28. A. Christophe, E. Dupoux, J. Bertoncini, J. Mehler, Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *J. Acoust. Soc. Am.* **95**, 1570–1580 (1994).
29. A. Christophe, J. Mehler, N. Sebastián-Gallés, Perception of prosodic boundary correlates by newborn infants. *Infancy* **2**, 385–394 (2001).
30. P. W. Jusczyk, R. N. Aslin, Infants' detection of the sound patterns of words in fluent speech. *Cognit. Psychol.* **29**, 1–23 (1995).
31. P. W. Jusczyk, D. M. Houston, M. Newsome, The beginnings of word segmentation in english-learning infants. *Cognit. Psychol.* **39**, 159–207 (1999).
32. J. H. Greenberg, C. A. Ferguson, E. A. Moravcsik, *Universals of Human Language: Phonology* (Stanford University Press, 1978), vol. 2.
33. E. Moreton, Structural constraints in the perception of English stop-sonorant clusters. *Cognition* **84**, 55–71 (2002).
34. D. M. Gómez *et al.*, Language universals at birth. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5837–5841 (2014).
35. K. Hitczenko, R. Mazuka, M. Elsner, N. H. Feldman, When context is and isn't helpful: A corpus study of naturalistic speech. *Psychon. Bull. Rev.* **27**, 640–676 (2020).
36. C. Bergmann, A. Cristia, Development of infants' segmentation of words from native speech: A meta-analytic approach. *Dev. Sci.* **19**, 901–917 (2016).
37. N. H. Feldman, S. Goldwater, E. Dupoux, T. Schatz, Do infants really learn phonetic categories? *Open Mind (Camb.)* **5**, 113–131 (2021).
38. Y. Rubner, C. Tomasi, L. J. Guibas, "A metric for distributions with applications to image databases" in *Sixth International Conference on Computer Vision* (IEEE, 1998), pp. 59–66.
39. M. Babineau, R. Shi, "Processing of French liaisons in toddlers" in *BUCLD 35 Proceedings* (Cascadilla Press, Somerville, MA, 2011), pp. 25–37.
40. S. L. Mattys, P. W. Jusczyk, Phonotactic cues for segmentation of fluent speech by infants. *Cognition* **78**, 91–121 (2001).
41. A. Seidl, E. K. Johnson, Infant word segmentation revisited: Edge alignment facilitates target extraction. *Dev. Sci.* **9**, 565–573 (2006).
42. M. T. C. Ernestus, *Voice Assimilation and Segment Reduction in Casual Dutch: A Corpus-Based Study of the Phonology-Phonetics Interface* (Holland Institute of Generative Linguistics, Utrecht, 2000).
43. S. Antetomaso *et al.*, "Modeling phonetic category learning from natural acoustic data" in *BUCLD 41: Proceedings of the 41st Annual Boston University Conference on Language Development*, M. LaMendola, J. Scott, Eds. (Cascadilla Press, 2017), pp. 32–45.
44. L. Lisker, "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang. Speech* **29**, 3–11 (1986).
45. C. Narayan, "Developmental perspectives on phonological typology and sound change" in *Origins of Sound Change: Approaches to Phonologization*, A. C. L. Yu, Ed. (Oxford University Press, 2013), pp. 128–146.
46. K. Hitczenko, contextual-dl. GitHub. https://github.com/khitczenko/contextual-dl. Deposited 2 August 2022.
47. C. Narayan, A. Peters, V. Woldenga-Racine, "Fragile phonetic contrasts in longitudinal infant-directed speech: Implications for infant speech perception" in *41st Annual Boston University Conference on Language Development* (2017). https://saplabyork. github.io/pdfs/BUCLD2017_poster.pdf (Accessed 24 August 2022).
48. T. Pimentel, B. Roark, R. Cotterell, Phonotactic complexity and its trade-offs. *Trans. Assoc. Comput. Linguist.* **8**, 1–18 (2020).
49. C. L. Roark, D. C. Plaut, L. L. Holt, A neural network model of the effect of prior experience with regularities on subsequent category learning. *Cognition* **222**, 104997 (2022).
50. B. Tranel, *The Sounds of French: An Introduction* (Cambridge University Press, 1987).
51. M. Clayards, M. K. Tanenhaus, R. N. Aslin, R. A. Jacobs, Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* **108**, 804–809 (2008).
52. J. Maye, R. N. Aslin, M. K. Tanenhaus, The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cogn. Sci.* **32**, 543–562 (2008).
53. A. Cristia *et al.*, Linguistic processing of accented speech across the lifespan. *Front. Psychol.* **3**, 479 (2012).
54. M. van Heugten, E. K. Johnson, Infants exposed to fluent natural speech succeed at cross-gender word recognition. *J. Speech Lang. Hear. Res.* **55**, 554–560 (2012).
55. K. S. White, R. N. Aslin, Adaptation to novel accents by toddlers. *Dev. Sci.* **14**, 372–384 (2011).
56. D. Weatherhead, K. S. White, He says potato, she says potahto: Young infants track talker-specific accents. *Lang. Learn. Dev.* **12**, 92–103 (2016).
57. P. K. Kuhl, Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *J. Acoust. Soc. Am.* **66**, 1668–1679 (1979).
58. P. K. Kuhl, Perception of auditory equivalence classes for speech in early infancy. *Infant Behav. Dev.* **6**, 263–285 (1983).
59. P. W. Jusczyk, P. A. Luce, J. Charles-Luce, Infants' sensitivity to phonotactic patterns in the native language. *J. Mem. Lang.* **33**, 630 (1994).
60. P. W. Jusczyk, E. A. Hohne, A. Bauman, Infants' sensitivity to allophonic cues for word segmentation. *Percept. Psychophys.* **61**, 1465–1476 (1999).
61. K. S. White, S. Peperkamp, C. Kirk, J. L. Morgan, Rapid acquisition of phonological alternations by infants. *Cognition* **107**, 238–265 (2008).
62. E. Moreton, S. Amano, "Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies" in *Proceedings of the 6th European Conference on Speech Communication and Technology* (Eurospeech, 1999), pp. 2679–2682.
63. K. Idemaru, L. L. Holt, Word recognition reflects dimension-based statistical learning. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 1939–1956 (2011).
64. L. L. Holt, A. J. Lotto, Cue weighting in auditory categorization: Implications for first and second language acquisition. *J. Acoust. Soc. Am.* **119**, 3059–3071 (2006).
65. J. C. Toscano, B. McMurray, Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* **34**, 434–464 (2010).
66. L. Liu, T. F. Jaeger, Inferring causes during speech perception. *Cognition* **174**, 55–70 (2018).
67. E. Bergelson, D. Swingley, At 6-9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3253–3258 (2012).
68. C. Villani, *Optimal Transport: Old and New* (Springer Science & Business Media, 2008), vol. 338.
69. K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation" in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (ISCA/IEEE, 2003).
70. F. Torreira, M. Adda-Decker, M. Ernestus, The Nijmegen corpus of casual French. *Speech Commun.* **52**, 201–212 (2010).
71. B. Schuppler, M. Ernestus, O. Scharenborg, L. Boves, Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *J. Phon.* **39**, 96–109 (2011).
72. P. Fikkert, "On the acquisition of prosodic structure," PhD thesis, Leiden University, Leiden, the Netherlands (1994).
73. C. Levelt, "On the acquisition of place," PhD thesis, Leiden University, Leiden, the Netherlands (1994).
74. S. Young *et al.*, *The HTK Book* (Cambridge University Engineering Department, 2002), vol. 3, p. 12.
75. P. Boersma, Praat: A system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2001).

Downloaded from https://www.pnas.org by "ACQUISITIONS/SERIALS DEPT, UNIV OF MARYLAND" on September 12, 2022 from IP address 128.8.120.3.