COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2021) e13075

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13075

Quantifying Talker Variability in North-American Infants' Daily Input

Federica Bulgarelli,[a] Jeff Mielke,[b] Elika Bergelson[a]

[a]*Department of Psychology and Neuroscience, Duke University*
[b]*Department of English (Linguistics Program), North Carolina State University*

Received 15 December 2020; received in revised form 19 November 2021; accepted 22 November 2021

Abstract

Words sound slightly different each time they are said, both by the same talker and across talkers. Rather than hurting learning, lab studies suggest that talker variability helps infants learn similar sounding words. However, very little is known about how much variability infants hear within a single talker or across talkers in naturalistic input. Here, we quantified these types of talker variability for highly frequent words spoken to 44 infants, from naturalistic recordings sampled longitudinally over a year of life (from 6 to 17 months). We used non-contrastive acoustic measurements (e.g., mean pitch, duration, harmonics-to-noise ratio) and holistic measures of sound similarity (normalized acoustic distance) to quantify acoustic variability. We find three key results. First, pitch-based variability was generally lower for infants' top talkers than across their other talkers, but overall acoustic distance is higher for tokens from the top talker versus the others. Second, the amount of acoustic variability infants heard could not be predicted from, and thus was not redundant with, other properties of the input such as the number of talkers or tokens, or proportion of speech from particular sources (e.g., women, children, electronics). Finally, we find that patterns of pitch-based acoustic variability heard in naturalistic input were similar to those found with in-lab stimuli that facilitated word learning. This large-scale quantification of talker variability in infants' everyday input sets the stage for linking naturally occurring variability "in the wild" to early word learning.

Keywords: Acoustic variability; Corpus methods; Home language environment; Infancy; Language development; Naturalistic input; Talker variability

1. Introduction

Infants must sort through immense amounts of variability in order to learn their native language(s). For example, each instance of a given word is going to sound slightly different

Correspondence should be sent to Federica Bulgarelli, Department of Psychology and Neuroscience, Duke University, Durham, 417 Chapel Drive, Durham, NC 27708, USA. E-mail: fedebul@gmail.com

every time it is said, both within the same talker and across talkers, as a result of variables such as age, gender, dialect, and register (see Boland, Kaan, Valdés Kroff, & Wulff, 2016; Peterson & Barney, 1952). While previous research has investigated how learners contend with many of these sources of variability in the lab (e.g., Houston & Jusczyk, 2000; Schmale & Seidl, 2009; Singh, Morgan, & White, 2004), here we focus on quantifying one type of variability, namely talker variability, for highly common words in infants' input. That is, while previous research has tested the effects of talker variability on learning in controlled lab environments (Bulgarelli & Bergelson, accepted; Galle, Apfelbaum, & McMurray, 2015; Richtsmeier, Gerken, Goffman, & Hogan, 2009; e.g., Rost & McMurray, 2009), very little is known about how much acoustic variability infants hear in the real world.

In constrained laboratory experiments, previous research has shown that input that varies acoustically (as a function of a talker, accent, or affect) can be both helpful and challenging for learners. For example, after hearing familiar words from a single talker or in a single affect or accent, young infants do not always recognize those words when the talker, accent, or affect changes (Houston & Jusczyk, 2000; Schmale & Seidl, 2009; Singh, Morgan, & White, 2004; though see van Heugten & Johnson, 2012). Even for adults, who are highly proficient speakers of their native language, variability stemming from how different talkers produce words can slow down recognition of familiar words (Mullennix, Pisoni, & Martin, 1989). However, increasing variability during a learning phase can also help infant learners recognize trained words and reject mispronunciations (e.g., training with variable affect and testing words with an affect or pronunciation change; Singh, 2008). Similarly, with infant learners, increasing talker variability can be helpful for integrating input across talkers in a speech segmentation task (Estes & Lew-Williams, 2015), learning novel phonotactic patterns (Seidl, Onishi, & Cristia, 2014), and recognizing previously heard words when the talker changes (Houston, 1999).

Talker variability also shapes various components of early word learning. For instance, Bulgarelli and Bergelson (acepted under review) tested the effects of talker variability on early word representations in 8-month-olds. They found that learning a new word–object mapping with talker variability highlighted how the surface features of the word could change (e.g., that a change in talker does not break the word–object link) but made it difficult for infants to notice when the sounds of the word had changed (e.g., when it was mispronounced or a totally new word was said). In contrast, for slightly older infants, the same kind of talker variability manipulations have been found to help infants learn highly similar words. Specifically, when English-learning 14-month-olds are presented with novel minimal pairs (e.g., /buk/and/puk/) produced by a single talker repeating a small subset of tokens, they have difficulty associating these words with two novel objects (in a traditional Switch task; Stager & Werker, 1997). In contrast, when infants hear these words produced by a highly variable single talker or by multiple talkers, they succeed in learning the word–object pairings (Galle et al., 2015; Rost & McMurray, 2009); this effect holds in German as well (Hohle, Fritzsche, Meb, Philipp, & Gafos, 2020). Other manipulations also help 14-month-olds with this more challenging minimal pair switch task, such as highlighting the referential nature of the task, embedding the novel words in full sentences, and labeling familiar objects before introducing novel ones (Fennell & Waxman, 2010). Some types of variability, however, are not helpful for learning

novel minimal pairs. That is, when variability is added to the word tokens in a phonetically relevant dimension like VOT (voice onset time), 14-month-olds again fail to learn the two novel words (Rost & McMurray, 2010). Systematicity too appears to play a role in this task: if within-word variability is completely systematic (e.g., nine women produce word 1 /buk/, nine men produce word 2 /puk/), infants do not seem to leverage talker variability for learning minimal pairs (Quam, Knight, & Gerken, 2017). Thus, while talker variability is not *necessary* for learning highly similar words, adding variability in an *irrelevant* dimension seems to allow infants to attend to the differences between the words that remain consistent across the variable presentations, for example, their initial phonemes (Galle et al., 2015; Rost & McMurray, 2009; see also Tripp, Feldman, & Idsardi, 2021 for a proposal that the benefit of talker variability may be due in part to the group membership of the informant).

While both within- and between-talker variability during training help infants learn minimal pairs in the lab with similar effect sizes (Galle et al., 2015; Rost & McMurray, 2009; Tsui, Byers-Heinlein, & Fennell, 2019), these types of talker variability have different acoustic manifestations. Galle et al. (2015) quantified these differences by measuring a set of 10 acoustic properties that are irrelevant to word identity, and not contrastive in English (e.g., mean pitch, harmonics-to-noise ratio, duration). They then computed standard deviations of these measurements for the between- and within- talker stimuli used in Galle et al. (2015) and Rost and McMurray (2009). The authors found that for some of the acoustic measurements, within-talker stimuli were significantly more variable, whereas for others between-talker stimuli were significantly more variable. This suggests that even though these types of (putatively "irrelevant") variability influence word learning similarly using the switch task in the lab, the acoustic information they provide diverges.

Critically, while talker variability has been found to impact word learning in the lab (see also Richtsmeier et al., 2009), very little is known about how much between- or within-talker variability infants actually experience in their daily environments. Previous research using naturalistic recordings has shown that North American infants learning a spoken language receive input from an average of six different people every day (Bergelson & Aslin, 2017). Convergently, questionnaire studies have found that French infants in the first year of life hear between 2 and 21 talkers a week (Bergmann & Cristia, 2018). However, knowing how many talkers are in a child's input does not necessarily reveal how many different talkers a child is likely to hear a given word from, and it does not provide information about the acoustic variability provided by these talkers. Measuring the acoustic variability in infants' daily input is thus an important first step in investigating whether variability influences word learning beyond what is found in tightly controlled lab settings.

In the current research, we seek to quantify acoustic variability in infants' naturalistic input for highly frequent, early-learned nouns. Specifically, we measure acoustic variability for concrete nouns that are highly frequent in infants' input, based on naturalistic all-day longitudinal recordings of a sample of 44 infants, recorded monthly from 6 to 17 months. We chose nouns because they are the focus of the lab experiments with talker variability we are building on, and they are highly prevalent in English-learning children's early vocabularies (Fenson et al., 1994; Gentner, 1982). In contrast to Rost and McMurray (2009) and Galle et al. (2015), we do not limit our analysis to minimal pairs, as these are quite rare in a speech to (and by)

infants (Caselli et al., 1995). In what follows, we describe variability at three different levels of aggregation by talker type. Our *overall variability* measure includes all noun tokens in the input to infants, our *top-talker* measure uses noun tokens produced by the most frequent talker each infant heard, and our *other-talkers* measure uses tokens from all other talkers for each child (omitting the top-talker). Our analysis takes both a frequency-based approach and an acoustic approach. The frequency-based approach seeks to describe properties of the input that may vary across infants (e.g., the number of talkers, number of tokens). The acoustic approach instead seeks to establish *how much* acoustic variability infants hear from their top-talker and from other talkers, using both non-contrastive acoustic measures used in previous research (Galle et al., 2015) and holistic sound-based similarity metrics (see Mielke, 2012). We also compare acoustic variability for top- and other-talker input, asking whether and how each talker category provides differently varying acoustic information. We further ask whether the amount of acoustic variability infants hear is related to more straightforward input metrics (i.e., number of tokens, number of talkers), or the proportion of the input they hear from different talker categories (i.e., female adults, children, electronics). Understanding what properties may underlie acoustic variability has implications for ways to systematically increase this variability in cases where it may be helpful for learning.

2. Methods

2.1. Participants

Participants were 44 infants recruited for a year-long study of word learning, who participated monthly starting at 6 months of age until 17 months of age (23 males, 21 females). All infants were born full term (40 ± 3 weeks), had no known vision or hearing problems, and were monolingually raised (parents did not report $>25\%$ exposure to a language other than English). Forty-two of the infants were White and two were multiracial. Maternal education ranged from high school degree to advanced degree (high school degree: $n = 1$; some college: $n = 3$; associate or bachelor's degree: $n = 18$, advanced degree: $n = 22$). The sample includes one pair of dizygotic twins; both are included.[1]

2.2. Procedure

2.2.1. Home recordings, initial data processing, and manual language annotation

Over the course of the year-long study that began when infants were 6 months old, families were audio-recorded once a month for an entire day (up to 16 h), and video recorded once a month for an hour (on different days' see Bergelson, Amatuni, Dailey, Koorathota, & Tor, 2018; see also Bergelson, 2017, for data and fuller data description). For the monthly audio recordings, infants wore vests with built-in pockets that housed a small LENA audio recorder (LENA foundation). Families were given these materials in advance and instructed to have children wear the vest and the LENA recorder from the time they woke up until they went to sleep for the night, except for naps and baths. Families could pause the recorder but were asked to minimize these pauses. The monthly hour-long video recordings were obtained

during a typical hour of the infants' day. Infants wore a hat or headband affixed with two small Looxcie cameras (22 g each). One camera was oriented slightly down and the other slightly up, to best capture infants' visual field (verified by Bluetooth with an iPad/iPhone during setup). A standard camcorder (Panasonic HC-V100 or Sony HDR-CX240) on a tripod was positioned in the corner, which parents were asked to take with them if they changed rooms. Other than equipment drop-off and pickup, and video-recording setup, the research team was not present during recording sessions. Details on the entire data processing pipeline can be found on OSF, see also Bergelson et al. (2018).

For months 6 and 7, the full daylong recordings were used. For months 8–17, a script (see OSF) was used to demarcate the 4 (month 8–13) or 3 (month 14-17) hours in the file with the most talk, dubbed *subregions* (see Supplementals for additional information). Trained researchers annotated all videos and full daylong recordings (months 6 and 7) or subregions (months 8–17) for all audio recordings. Thus, the current paper uses input heard by these 44 infants from approximately 54 audio-recorded hours and 12 video-recorded hours per child, sampled sparsely (i.e., monthly) over the course of a year.

Based on the broader goals of this project (which focused on testing noun knowledge), each imageable concrete noun said directly to or near the target child was manually tagged by annotators. This included body parts (e.g., hand, foot) and foods (e.g., cracker, milk), but not occupations (e.g., teacher), or proper nouns. Concrete nouns produced in the distance (such as faint background television) were not included. Each noun token was labeled with an utterance-type (declarative, imperative, question, reading, singing, short phrase (i.e., less than three words with no verb), see Bergelson et al., 2018), a tag for whether the referent of the noun was present and attended to by the infant (yes, no, unclear), and individual talker labels (see Bergelson et al., 2018; Bulgarelli & Bergelson, 2029, for information on the reliability of speaker tags; reliability was high); the likely addressee was not coded.

2.2.2. Extracting individual word tokens

For the present analysis, we identified the five most frequent nouns across the entire corpus (tokens = 21,499, types = 5), which were "baby," "book," "ball," "water." and "dog(gy)." For each of these, we extracted an audio-clip for each noun instance in all audio- and video-recordings based on its timestamp and a 0.5-s buffer on each side. Then, trained research assistants transcribed these segments (mean length = 1.83 s) using Praat. Using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), we next aligned the transcribed textgrids to the audio wav files. The first author reviewed all force-aligned files ($N = 21{,}499$) and made adjustments as necessary to the alignment of the target words, and then extracted the wav files containing the bare target words. This process resulted in individual wav files for each token of each target word for each participant.

2.3. Acoustic measurements

We took two approaches to quantifying acoustic variability in these word tokens. Our first set of analyses measured acoustic properties that are not lexically contrastive in English, as done by Galle et al. (2015) on their laboratory stimuli; hereafter, we refer to these as

non-contrastive acoustic measures. The second acoustic analysis employed a phonetic metric of sound similarity, based on Mielke (2012); hereafter we refer to this analysis as the **holistic measure of sound similarity**.

For our **non-contrastive acoustic measures**, we measured *mean pitch, median pitch, max pitch, mean pitch slope, duration*, and *harmonics-to-noise ratio* (all operationalized below). All of these (except median pitch) were measured by Galle et al. (2015), and we conduct the same measurements to facilitate comparisons across naturalistic input and in-lab stimuli. We further add median pitch as a pitch measure that is less sensitive to pitch extrema than the pitch's max or mean. Each of these measurements was conducted on the whole word using an automated approach in PraatR (Albin, 2014; see the script at https://osf.io/6gsez/ for details about how each measurement was calculated).

We divided our tokens into four categories: male adult, female adult, child, or electronic. We used a category-specific pitch range for the pitch-related measurements in Praat. For tokens in the male adult category, we used a pitch range of 50–400 Hz, for tokens in the female adult category we used a pitch range of 100–600 Hz, for children we used a pitch range of 200–800 Hz, and for electronics we used a pitch range of 50–800 Hz. Using these category-specific pitch ranges, we calculated several measures. *mean*, *median*, and *max* pitch reflect the average, median, and maximum pitch, in hertz, for the entire word (Galle et al., 2015, call max pitch "pitch excursion."). We further calculated *mean pitch slope* as the average in the differences between consecutive pitch measurements over the course of the word, in Hz/ms (Galle et al., 2015, call this *pitch slope*). Finally, we computed *duration* (the length of the word) in seconds (to the nearest 1/100th of a second), and *harmonics-to-noise ratio* (a ratio of the energy contained in the periodic part of the speech signal to the energy contained in the aperiodic part) using the harmonicity (cc) standards in Praat, with a time step of 0.01 s, the category-specific minimum pitch, silence threshold of 0.1, and number of periods per window of 1.

While Galle et al. (2015) measured an additional five spectral measures, here we opt not to include these as they are not intended to measure properties of an entire word, but are more traditionally measured on specific sounds, for example, affricates and fricatives (e.g., Jongman, Wayland, & Wong, 2000). Furthermore, spectral measurements are affected by the vowel in the word itself; while Galle et al. (2015)'s stimuli all had the same vowel, ours do not, and thus spectral measurements are challenging to compare.[2]

For a more **holistic measure of sound similarity**, we calculated acoustic distances between any two tokens of a word (within subject) using a dynamic time warping algorithm (see Mielke, 2012, for an earlier use of dynamic time warping to measure phonetic similarity). First, each sound file was converted into matrices of MFCCs with 12 coefficients using PraatR (see OSF for script). Using the DTW package in R (Giorgino, 2009), we used these matrices to calculate the normalized acoustic distance between every two tokens of a word for each subject, using the Euclidian distance method. The DTW algorithm calculates the acoustic distance between two tokens by first finding the optimal path between the matrices of two sounds; and then calculating the average of the distances for all points that the path passes through. We used the "symmetric2" slope constraint in the DTW package, which is normalizable and symmetric but does not impose any slope constraints. Due to differences in

lengths across files, imposing slope constraints such that files could not be more than three times as long as each other would have resulted in excluding 40,548 comparisons, involving 10,359 different files. Since we are interested in capturing variability across instances of words, imposing these constraints would not capture the full range of variability.[3] If the two sounds were identical, the optimal path would be perfectly diagonal, and the normalized acoustic distance would be 0. More similar sounds or words will therefore have a smaller normalized acoustic distance, while less similar sounds or words will have a larger one. A normalized acoustic distance was calculated for each unique pair of tokens of, for example, the word ball for subject 1 (repeated for each word, for each subject).

2.4. Excluding unusable tokens

There were a few types of problematic word tokens that we excluded before analysis: those with likely pitch tracking errors, overlapping environmental sounds, and high static.

Since we were specifically interested in quantifying the amount of variability infants hear, we did not exclude any extreme values (such as those more than 3 standard deviations away from the mean). However, extreme values in pitch measurements can be due to erroneous pitch tracking as opposed to actually extreme values. While manually inspecting the pitch track for >20,000 word tokens was not feasible, we sought to minimize errors (particularly for our extrema measurements such as max pitch and mean pitch slope) by excluding any tokens that had consecutive pitch measurements that differed by more than an octave (double or half the previous pitch). Such jumps in pitch are a classic signature of pitch-tracking errors and are unlikely to occur in natural speech. We identified $n = 4{,}176$ that met this criteria and excluded them from our analyses.

Word tokens from naturalistic recordings can include background noise that human listeners readily filter or source-separate, but which potentially skew the acoustic measures we sought to characterize (e.g., other speakers in the background, music, toy noises, or animals). To protect against this possibility, the first author listened to each individual token and determined whether it needed to be excluded either due to overlap with other speakers, or other environmental noise (e.g., ball bouncing, hands clapping, toy noises, etc). Using this method, 446 tokens (out of 17,570) were excluded due to environmental noise (music, toy noises, other loud noises), while 570 were excluded due to overlap with other speakers. To ensure consistency, a trained research assistant listened to $\sim 10\%$ ($N = 542$), of the tokens of "baby," the most frequent of our words, to independently assess whether they should be excluded for background noise. Agreement was 92.79%, kappa = 0.65.

We also excluded any token with a harmonics-to-noise ratio <1 ($N = 865$). While a harmonics-to-noise ratio of 1 is not intrinsically meaningful, setting this cutoff excludes the small tail of tokens with a relatively high ratio of aperiodic noise relative to periodic speech, thus excluding tokens where the word of interest was muffled by static. A histogram of the distribution of harmonics-to-noise ratio values for all tokens (included and excluded for all the reasons above) can be found in Fig. 1. Lastly, we excluded any token for which any of the acoustic measurements could not be measured ($N = 334$); this was usually the case when pitch information was missing. While a relatively small number of tokens were excluded

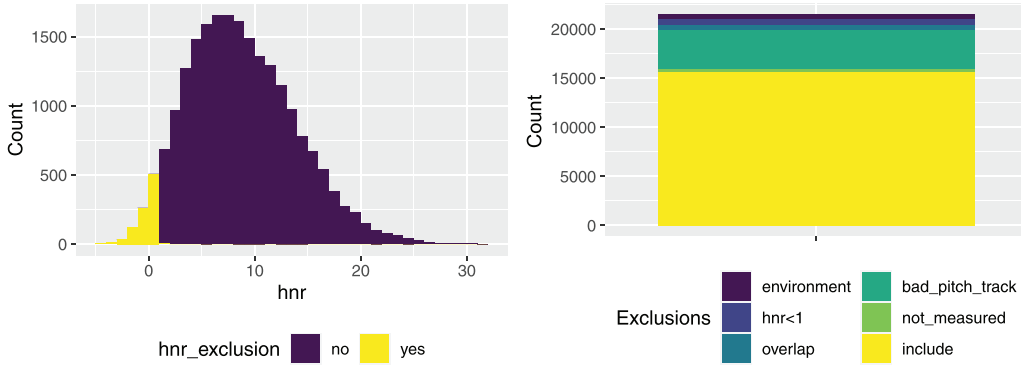

Fig. 1. Left: Histogram of the distribution of the harmonics-to-noise ratio values. Those in yellow are below 1, and were excluded. Right: Stacked bar plot of inclusions and exclusions.

Table 1
Per-word details about tokens and talkers across the sample of 44 subjects. The "Tokens" column shows the total number of tokens included, with the percentage of tokens included after our exclusion process in parentheses. The "Talkers" columns shows mean(range) number of talkers. The "Others", "Top," and "Overall' columns show the mean (SD)" for tokens in each of these talker-types. The "All" row represents the entire dataset

Word	Tokens	Talkers	Between	Within	Overall
Baby	3825 (75.55%)	5 (2–10)	30 (26)	57 (52)	87 (60)
Ball	2662 (71.97%)	4 (2–8)	19 (20)	42 (34)	60 (43)
Book	3242 (68.63%)	4 (1–8)	20 (17)	54 (53)	74 (58)
Dog	3188 (72.44%)	5 (1–9)	28 (28)	45 (40)	72 (58)
Water	2695 (74.61%)	5 (2–13)	18 (14)	43 (30)	61 (36)
All	15,612 (72.62%)	5 (1–13)	23 (22)	48 (43)	71 (53)

due to missing values, excluding them avoided unnecessary researcher degrees of freedom in data imputation approaches in our analyses. While we find these exclusions methodologically appropriate for our goals, it is in principle possible that they could impact our results. We make all tabular data available alongside this paper for interested researchers who would like to consider different criteria.

Thus, after all exclusions, the current dataset (see Table 1) includes 3824 tokens of "baby"; 2662 tokens of "ball"; 3242 tokens of "book"; 3187 tokens of "dog/doggy"; and 2695 tokens of "water."

2.5. *Data aggregation*

Given our central goal of characterizing talker *variability* across our measures, we aggregated the data in three ways: *overall*, *top-talker*, and *other-talkers*. Specifically, the *overall variability* analyses combined all tokens heard by infants. The *top-talker* analyses were done using all tokens produced by each subjects' top talker (i.e., the talker who produced the most nouns across each infants' entire corpus). In most cases, the top talker was the infant's mother,

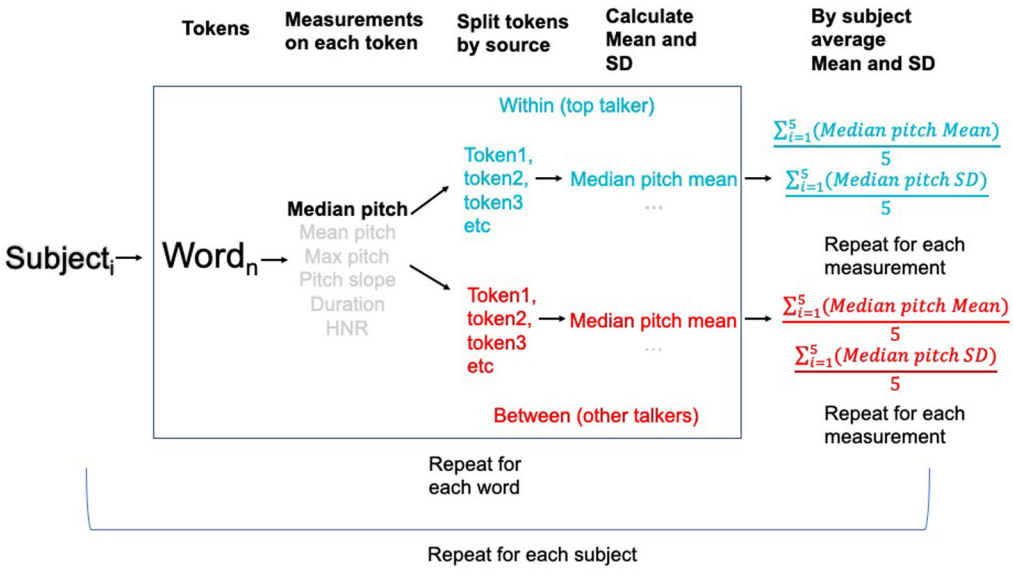

Fig. 2. Data aggregation flow. This process results in a other-talker and top-talker mean and SD for each of the six non-contrastive acoustic measurements, for each of the five words, for each of the 44 infants.

but for some it was their father or grandparent. The *other-talkers* analyses were then conducted by grouping together all tokens produced by other talkers from each child's input, excluding their top-talker. In the remainder of the paper, we use *top-talker*, *other-talkers*, and *overall variability* as described here, and we call this variable *talker-type*. As it remains an open question whether learners track talker-specific distributions (see, e.g., Choi & Shukla, 2021) and to maintain independence of observation, we only conduct statistical analyses comparing the *top-talker* and *other-talkers* levels, as comparing *overall* to either *top-talker* or *other-talkers* would violate independence assumptions.

For our **non-contrastive acoustic measurements**, we then calculated **means**, **ranges,** and **standard deviations** (SD) for each infant, for each word, at each *talker-type* level (see Fig. 2 for data aggregation flow). The SDs for mean pitch, max pitch, mean pitch slope, duration, and harmonics-to-noise ratio for the *top-talker* and *other-talkers* levels can be directly compared to the measurements reported for *within-talker* and *between-talker* stimuli (respectively) by Galle et al. (2015). The means and ranges are reported for transparency in the Supplemental materials.

For our **holistic measurement of sound similarity**, we calculated an average normalized acoustic distance for each subject for each word,[4] at each *talker-type* level. For this measurement, the *top-talker* average reflects the average acoustic distance for all comparisons of tokens produced by the top talker, while the *other talkers* average reflects the average acoustic distance for all comparisons of tokens produced by any other talker. That is, the *other talkers* average includes dynamic time warping comparisons of tokens produced by all talkers excluding the top talker (e.g., subject 1's dad tokens to each other and to subject 1's

grandmother tokens). The *overall variability* average reflects comparisons for all tokens (subject 1's top talker tokens to each other, and all other tokens (e.g., dad, grandmother)).

3. Results

In what follows, we quantified talker variability by taking a frequency-based and an acoustic-based approach. In the frequency-based section, we asked how many different talkers infants heard producing highly frequent nouns, and how many tokens of those words infants heard overall from their top-talker and other talkers. For the acoustic-based section, we ask how much acoustic variability infants heard for these highly common words overall, and from their top-talker and all other talkers, using both the **non-contrastive acoustic properties** and the **holistic similarity metric** described above. Then, we investigated whether the amount of acoustic variability infants heard is related to other properties of the input, such as the number of tokens or the number of talkers infants heard, and the proportion of input from different sources (e.g., female adults, children, electronics). Lastly, we asked how the patterns of acoustic variability for naturalistic input differed from those reported for in-lab stimuli that are designed to feature high talker variability, and which support novel word learning in prior work.

3.1. Frequency-based analyses

3.1.1. Top concrete nouns

As described above, our analyses focus on the top five concrete nouns in the corpus overall: baby, ball, book, water, and dog(gy). On average, 4.60 of these five words occurred in the top 25 concrete nouns for all families (see also Bergelson et al., 2018, for the top 10 words for just months 6 and 7 of the corpus). Further, an analysis of CHILDES using childes-db (Braginsky, Sanchez, & Yurovsky, 2019) confirms that these five words are in the top 25 concrete nouns for a large sample of corpora (see Moore & Bergelson, under review, for full details on the corpus analysis). Thus, while family-by-family frequency of individual concrete nouns varied, the nouns analyzed here provided a representative sample of highly frequent concrete nouns for this corpus, and for many other corpora in CHILDES.

3.1.2. Number of talkers and number of tokens

Over the course of the $\sim$ 66 h annotated from each infants across 24 days over the sparsely sampled year, infants heard an average of 71 tokens of each of these words, produced by five talkers on average (including both the child's top talker and all other talkers; see Table 1 for a breakdown by word).[5] These values were modestly correlated: hearing more talkers produce a specific word resulted in hearing more tokens of that word overall, $\tau = 0.21$, $z = 4.36$, $p < .001$.

We then divided the tokens into top-talker and other-talkers tokens (as detailed above) and counted the instances of each word for each subject. For example, one subject heard the word "baby" 36 times from their top talker and 19 times from all other talkers ($n = 4$ talkers). Each

subject's average number of tokens was calculated as the mean token count across words for each talker-type (see Fig. 2). On average, infants heard 48 tokens of each word from their top talker and 23 tokens of each word from all other talkers (see Table 1 for breakdown by word). There was a modest correlation between the number of tokens infants heard from their top-talker and from other-talkers $\tau = 0.17$, $z = 3.73$, $p < .001$, suggesting that infants who hear more input tend to hear more input from both talker-types.

As evidenced by the ranges and SD reported in Table 1, there was considerable variability across subjects in how many tokens of these words infants heard and how many talkers they heard them from. What these values cannot tell us, however, is (1) whether there was relative consistency in which talkers produced these words and (2) how many instances of a given word each talker produced. To address the first question, while infants heard five unique talkers on average for any given word, infants heard an average of 9 ($SD = 3$) distinct talkers across these five words. Additionally, any given talker produced, on average, three of these words; with some producing only one word and others producing all five ($SD = 2$). Together, this suggests that there is also variability in *who* infants heard produce even highly common words.

Answering the second question, looking across all talkers (including the child's top talker), infants heard an average of 16 tokens of each word per talker, though this varied substantially across subjects (range = 2–116, $SD = 22$). However, this value is inflated by the child's top talker. Focusing only on other-talker tokens, infants heard an average of six tokens of each word per talker (range = 1–41, $SD = 7$).

We further sought to describe the distribution of other-talker tokens from different talker categories. The proportion of other-talker tokens produced by female adult talkers (33%) did not differ from the proportion of other talker tokens produced by male adult talkers (36%), $t(431.66) = -0.84$, $p = .399$. In addition to other-talker tokens produced by adults, 22% of other-talkers tokens were produced by child talkers, and 9% of other-talkers tokens were produced by electronics.

Taken together, our descriptive, frequency-based analyses found wide variability in the number of talkers and tokens infants heard producing highly common nouns. Moreover, this variability was seen both in the number of tokens infants heard across all talkers, as well as within individual "top" talkers.

3.2. Acoustic variability

3.2.1. Acoustic measurements of naturalistic tokens

Having established how many talkers and how many tokens of words infants heard, we now turn to describing the acoustic measurements of these tokens. For our **non-contrastive acoustic measurements**, the central focus is the *standard deviations* of these acoustic measurements, which allowed us to quantify talker *variability*. As we did for the frequency measurements, we examined the acoustic measurements overall, for the top-talker alone, and for the other talkers.[6] Specifically, we calculated the standard deviation for each acoustic

Table 2
Standard deviations of acoustic measurements for other-talkers and top-talker tokens in the current dataset and in Galle et al (2015) (for non-contrastive acoustic measurements) and mean normalized acoustic distance in the current dataset. Asterisks indicate which talker-type (others vs. top) was found to be significantly more variable. *p*-Values are uncorrected. Max pitch, mean pitch, and median pitch comparisons withstand the Bonferroni correction for multiple comparisons ($p = .007$)

	Current dataset				Galle et al. (2015)	
Measurement	Overall	Other-talkers	Top-talker	*p* value	Between	Within
SD Mean_pitch(Hz)	85.66	90.55*	69.18	<.001	77.54*	37.73
SD Max_pitch(Hz)	119.26	123.97*	100.73	<.001	137.22*	64.48
SD Median_pitch(Hz)	86.91	91.72*	70.6	<.001	NA	NA
SD duration(ms)	170.87	160.17	161.57	0.602	112.03	203.65*
SD Harmonic_to_Noise(dB)	4.37	4.01	4.18*	0.048	3.07	3.95*
SD Mean_Pitch_slope(Hz/ms)	4.57	4.63*	3.94	0.014	9.57*	6.67
Mean acoustic distance	294.61	224.46	297.87*	<.001	NA	NA

measurement for each word, for each subject (e.g., the SD of duration across tokens of "ball" heard by subject 1 from all talkers (overall variability), their top talker, and their other talkers).

For our **holistic sound similarity measure**, the central focus is the average acoustic distance calculated using the dynamic time warping approach described above. This measurement allowed us to capture overall similarity across tokens of words in the *overall* dataset, as well as for the *top-talker* or *other-talkers*. As described above, we calculated the mean acoustic distance for each word, for each subject at each talker-type level (overall, top, other). Values reported in Table 2 aggregate across words and subjects. That is, each *overall variability* value and each *top-talker* standard deviation was made up of 220 data points (5 words * 44 subjects). Due to a lack of other-talkers tokens for some words/subjects, the *other-talkers* values were made up of 214–217 data points.

Using these sets of acoustic measurements, we tested whether infants heard more acoustic variability from their *top-talker* or *other-talkers*. This let us assess whether the child's top and other talkers provide different evidence for the limits of how words can vary. As noted above, we only conducted statistical comparisons across these two *talker-type* levels, as these maintain independence of observations. We compared values (SDs for non-contrastive acoustic measurements and means for the holistic sound similarity measure) for the *top-talker* and the *other-talkers* using a Wilcoxon test.

We found that variability from *other-talkers* was significantly greater than that from the top-talker for mean pitch, max pitch, median pitch, and mean pitch slope, while variability from the *top-talker* was significantly greater for the harmonics-to-noise ratio, though the differences for mean pitch slope and harmonics-to-noise ratio did not withstand correction for multiple comparisons (see Table 2). Variability from the top and other talkers did not differ for the duration. We conducted a similar analysis for holistic measurement of sound similarity to assess whether the tokens infants heard are overall more similar for their top talker relative to all other talkers. We found that the mean normalized acoustic distance was significantly larger for *top-talker* tokens relative to *other-talkers* tokens.

The preceding analysis of SDs and holistic acoustic distance collapsed across all words in our dataset. To test whether word-level effects contributed to the variance of these measurements (i.e., our measure of talker variability), we next conducted a series of 5 (word) $\times$ 2 (talker-type: top/other-talker) ANOVAs, one for each **non-contrastive acoustic measurement** and the **holistic similarity measure**, testing the effects of word, talker-type, and critically, their interaction.

As expected, the main effect of word was significant for all models (all $p < 0.02$). Additionally, the main effect of talker-type was significant for mean pitch $F(1, 424) = 58.96$, $MSE = 829.96$, $p < .001$, max pitch $F(1, 424) = 39.17$, $MSE = 1,488.88$, $p < .001$, median pitch $F(1, 424) = 51.59$, $MSE = 913.16$, $p < .001$ and mean pitch slope $F(1, 424) = 12.33$, $MSE = 52,070.44$, $p < .001$, but not for harmonics-to-noise ratio $F(1, 424) = 3.30$, $MSE = 1.08$, $p = .070$, or duration $F(1, 424) = 0.06$, $MSE = 0.00$, $p = .808$ consistent with our prior analysis; see Table 2. The main effect of talker-type was also significant for normalized acoustic distance, $F(1, 424) = 70.71$, $MSE = 7,301.35$, $p < .001$.

We were particularly interested in the word-by-talker-type interaction, as a significant interaction would suggest that top- versus other-talker variability varied as a function of the word itself. The word by source interaction was significant for two of our acoustic measures: mean pitch ($F(4, 424) = 3.62$, $MSE = 829.96$, $p = .006$) and median pitch ($F(4, 424) = 3.48$, $MSE = 913.16$, $p = .008$). The pattern of larger standard deviations for other-talker tokens relative to top-talker tokens held for all words, but was noticeably weaker for "ball"; see Fig. 3. All other interactions were not significant (max pitch: $F(4, 424) = 2.19$, $MSE = 1,488.88$, $p = .069$; duration: $F(4, 424) = 1.52$, $MSE = 0.00$, $p = .194$; mean pitch slope: $F(4, 424) = 0.90$, $MSE = 52,070.44$, $p = .465$; harmonics-to-noise ratio: $F(4, 424) = 2.06$, $MSE = 1.08$, $p = .085$; or normalized acoustic distance $F(4, 424) = 1.79$, $MSE = 7,301.35$, $p = .131$), see Supplementals for similar graphs for the remaining acoustic measures. These are exploratory analyses, and, therefore, both significant and non-significant differences should be considered cautiously.

3.2.2. Predictors of acoustic variability

Our next analysis investigated whether the acoustic measures described above provided unique information about infants' naturalistic input, above and beyond other common descriptive measures. Specifically, it is possible that hearing more tokens or hearing more talkers would correlate with the amount of top- or other-talker variability infants heard, as hearing more tokens or more talkers may lead to more variability in how those tokens are produced. Similarly, speech from certain categories of talkers may be more or less likely to contribute to overall variability. For example, toys or media are likely the only sources that provided highly consistent instances of words, as they are repeated identically every time. Do children who heard a larger proportion of these highly frequent words from electronics receive relatively less variable input overall? On the other hand, talker age and gender may also be related to the amount of variability infants received. That is, the phonetic properties of child speech vary from those of female speech and from those of male speech. Further, female talkers may vary their speech more when talking to infants than male talkers (e.g., Fernald et al., 1989; Gleason, 1975; though see Benders, StGeorge, & Fletcher, 2021, for analyses suggesting the

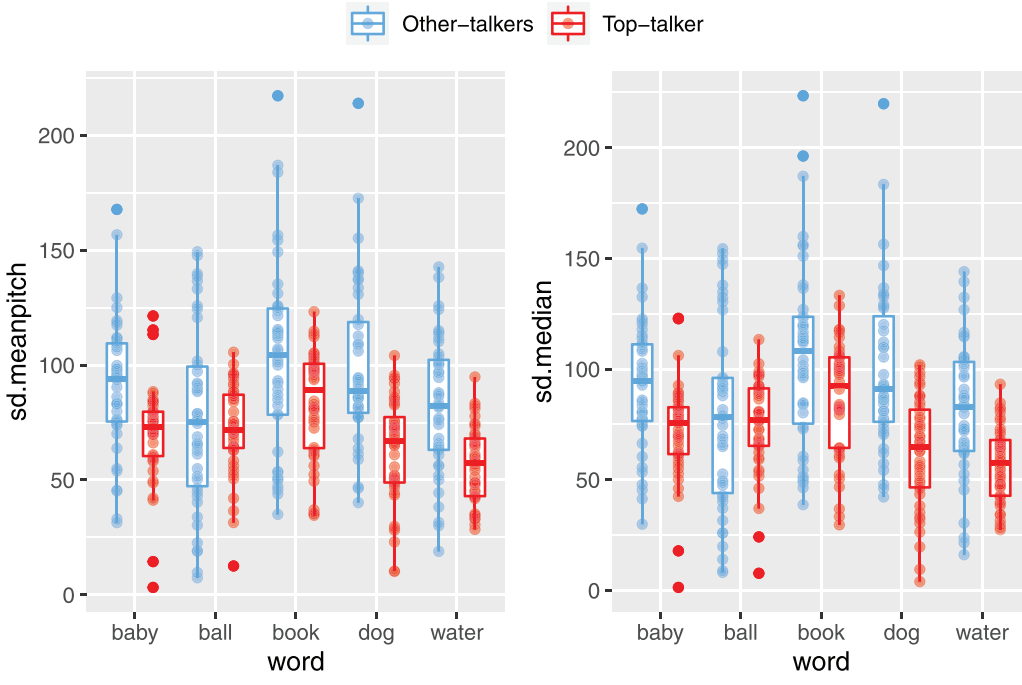

Fig. 3. By word mean pitch SD (left) and median pitch SD (right) for top-talker (red) and other-talkers (blue) tokens. Each dot is the value for one participant, line corresponds with median value, ends of the boxplot correspond to the first and third quartiles.

magnitude of the difference may be similar). That is, hearing a higher proportion of female speech, or speech from other children (which is likely higher overall in pitch-related properties), may be correlated with hearing more variability.

To test this possibility, we report two sets of correlations. In the first, we tested the correlation between acoustic variability (both the non-contrastive acoustic measurements and the holistic acoustic distance) and the number of tokens infants heard within our set of top-talker tokens. This let us assess whether hearing more tokens lead to more variability in what those words sounded like from the top-talker. In the second set, we tested the correlations between the other-talkers acoustic variability and the number of talkers, as well as the number of tokens (controlling for the number of talkers), and the proportions of electronic, female adult, and child input (prepubescent, collapsing gender). This let us assess whether any of these properties of the input influenced the relative variability infants heard from other-talkers.

All correlations can be found in Table 3. Since all the variables were not normally distributed as determined by the Shapiro–Wilks test, we used Kendall correlations. Together, these correlations suggest that hearing more tokens resulted in hearing more acoustic variability within infants' top-talker tokens. In contrast, while hearing more talkers resulted in more acoustic variability for other-talker tokens, the correlations between acoustic variability and other-talker tokens, after controlling for the number of talkers, did not withstand correction.

Table 3
Correlation matrix (Kendall's tau). Column two reports correlations between top-talker acoustic measurements and top-talker tokens. The remaining columns report correlations between other talker tokens and other-talker token count (column 3, controlling for number of talkers), talker count (column 4), proportion female tokens (column 5), proportion child tokens (column 6) and proportion of electronic tokens (column 7)

	Other-talkers					
Measurement	Top #Tokens	#Tokens	#Talkers	%Female	%Child	%Electronic
SD Max_pitch(Hz)	0.11*	0	0.22**	−0.02	0.33**	0.15**
SD Mean_pitch(Hz)	0.15**	0.02	0.24**	0.01	0.29**	0.15**
SD Median_pitch(Hz)	0.17**	0.03	0.23**	0.02	0.28**	0.14**
SD duration(ms)	0.15**	0.08	0.16**	0.05	0.13*	0.05
SD Harmonic_to_Noise(dB)	0.16**	0.1*	0.11*	0.11*	0.22**	−0.02
SD Mean_Pitch_slope(Hz/ms)	0.19**	0.1*	0.17**	0	0.36**	0.08
Mean acoustic distance	0.15**	0.02	0.07	0.02	0.06	0.16**

[a]**significant after correcting for multiple comparisons ($n = 7, p < .007$), *$p < .05$.

Further, the proportion of input from female or electronic sources seems to be only slightly related to acoustic variability. In contrast, the proportion of tokens produced by other children was strongly correlated with most of the acoustic measurements, suggesting that hearing more input from other children increases the acoustic range of possible speech, leading to more variability.

That said, the modest size of these correlations (no Kendall correlation larger than $\tau = 0.36$, mean $\tau = 0.12$) suggests that acoustic variability was not simply redundant with these other properties. That is, though tokens, talkers, and proportion of speech from different talker categories are far easier to quantify than acoustic measurements, the acoustic measurements characterized word variability above and beyond these other aspects of the input.

3.2.3. Comparison of real-world variability and in-lab variability

Lastly, we compared our measurements of acoustic variability for the top-talker and other-talkers to those reported by Galle et al. (2015) for within- and between-talkers stimuli used in in-lab studies (respectively). While there were many differences between our corpus-based analyses and those reported by Galle et al. (2015), we compared the direction of differences to begin to assess whether variability that had been shown to be useful for early word learning in the lab was similarly available to infants in the real world. As a reminder, Galle et al.'s "between-talker" stimuli reflected tokens of / buk/ and /puk/ spoken by 18 different talkers, half female, from Rost and McMurray (2009); their "within-talker" stimuli reflected tokens of /buk/ and /puk/ spoken by a single talker trained to maximize variability. Galle et al. (2015) found significant differences for each of their acoustic measurements; see Table 2. The mean pitch, max pitch, pitch slope, duration, and harmonics-to-noise ratio measurements on our corpus tokens could be directly compared to those reported by Galle et al. (2015). Their within-talker measurements map to our top talker, and their between talker maps to our other-talkers, but for ease of exposition we refer to them as top and other-talkers.

Overall, we found similar patterns of acoustic variability in our set of 15,610 naturalistic tokens as those reported for the ∼ 100 tokens used for previous lab stimuli. Namely, we see parallels in the *direction* of differences (i.e., whether other-talker or top-talker tokens were more variably produced). For 4/5 acoustic measurements that were measured for both sets, the directional differences in our naturalistic tokens matched those from the single pair of words used in prior in-lab experiments. Namely, mean pitch, max pitch, and mean pitch slope varied significantly more for other-talkers than for top-talker tokens, and harmonics-to-noise ratio varied significantly more for top-talker relative to other-talkers tokens (though as noted above, only mean and max pitch withstand the Bonferroni-corrected $p = .007$; see Table 2). Duration did not vary across types of talker variability for our naturalistic stimuli, with the magnitude of top- and other-talker variability falling between those found for within- and between-talker stimuli in lab studies. This may be due to the challenges inherent in comparing SDs of the duration of five words in naturalistic speech versus those of a minimal pair of items recorded in the lab. At the same time, this pattern may indicate that highly variable single talker recordings introduce more variability in duration than is likely to occur in the real world, while single tokens from multiple talkers provide less. (See Supplemental Tables 4 and 5 for a similar pattern of results on a direct (but limited) comparison of a subset of the corpus items and the same words recorded in the lab.)

Taken together, we found that comparing variability for top-talker and other-talker tokens from naturalistic input resulted in similar patterns of variability (particularly for the pitch-based measurements) relative to those reported for in-lab stimuli intended to systematically maximize talker variability.

4. Discussion

In the current study, we sought to quantify the sound-based variability in highly common nouns in infants' everyday learning environments, using both a frequency- and an acoustic-based approach. Our frequency-based analysis revealed several important descriptive properties of naturalistic input to infants. First, we found that across our 44 infants, five highly frequent words (baby, ball, book, dog(gy), and water) were heard from an average of five distinct talkers over the course of ∼ 66 hours over a sparsely sampled year (from 6 to 17 months). Furthermore, we found that the number of talkers infants heard producing these common words was quite variable, ranging from 1 to 13. At the same time, across the group, the number of talkers was highly similar for each word (see Table 1). Speculatively, the number of talkers producing these words is likely similar to other highly frequent aspects of the input, like closed-class words, though perhaps dissimilar from low-frequency words that might only be said in a limited range of circumstances or households (e.g., "edamame" or "forthwith"), though this remains an open question. We also found that while on average infants heard 71 tokens of each of these words, they heard considerably more input from their top talker relative to all other talkers in their input, with 64% of tokens being produced by the child's top-talker.

We also took two approaches to quantifying acoustic variability in the tokens of words infants heard. Both the **non-contrastive acoustic measures** and the **holistic measure of sound similarity** revealed significant differences in the variability of top-talker versus other-talker tokens. The directionality of these differences was often consistent with previous research using lab-based stimuli (Galle et al., 2015; Rost & McMurray, 2009; see Table 2). That is, infants in this longitudinal corpus heard more variability in pitch properties (mean, max, and median pitch) from other talkers, which is intuitive since individual talkers are more likely to vary in their fundamental frequency. By contrast, infants heard more variability from their top-talker for variables such as the harmonics-to-noise ratio and the mean normalized acoustic distance. These differences likely suggest that the child's top-talker modulates their voice quality more widely than the other talkers do, as both of these measures capture spectral properties of the speech sounds (e.g., breathiness).

We also found that while our measurements of acoustic variability were related to other, more easily measured properties of the input like the number of tokens and talkers infants heard, and the proportion of tokens from different talker sources, these correlations were small in magnitude. This suggests that acoustic measurements of naturalistic input provide novel information about the variability infants receive and must learn to contend with, beyond simple word or talker counts. That said, these correlational analyses suggest that some aspects of the input might be manipulable in order to increase acoustic variability (e.g., by increasing speech from other children), though the extent to which more variability is useful remains to be determined by future research.

What other kinds of factors might explain the range of acoustic variability in naturalistic input? We measured tokens of words produced in naturalistic speech to and around infants, which likely contained both infant-directed speech and adult-directed speech. Infant-directed speech is characterized by shorter utterances and more positive affect, but also more variability in pitch and pitch contour, compared to adult-directed speech (e.g., Cooper & Aslin, 1994; Fernald & Simon, 1984). It is possible that the proportion of child and adult-directed speech in the input may (at least partially) explain the amount of acoustic variability infants hear. Relatedly, talker gender also affects the prevalence of infant-directed speech, based on results showing that phonetic measurements of infant-directed speech vary between mothers and fathers (e.g., Fernald et al., 1989; Gleason, 1975) and that male and female talkers provide different proportions of infant-directed speech relative to adult-directed speech, both in North American samples (Bergelson et al., 2018) and cross-culturally Bunce et al. (in prep). While the proportion of speech produced by female talkers was not correlated with most acoustic measurements in the current analyses, future research could more pointedly examine the proportion of infant-directed speech produced by different talker categories, which is beyond the scope of the present work.

One important contribution of this work is our finding that the types of acoustic variability profiles that have been shown to shape word learning in limited and controlled lab stimuli (Galle et al., 2015; Rost & McMurray, 2009)—particularly those regarding pitch—are similarly present in infants' daily lives. As the stimuli used in these prior lab studies were recorded in a sound-proof booth by trained young adults, it could well have been the case that the

acoustic properties of those stimuli would not vary in similar ways to parents and other caregivers naturally interacting with their infants. At the same time, some of the acoustic measurements revealed less stable patterns across real-world and lab-recorded stimuli, particularly for the acoustic measurements that likely capture voice quality modulations. Specifically, across two sets of lab-recorded stimuli (see the Supporting Information) and our real-world input, normalized acoustic distance, and harmonics-to-noise ratio did not always vary more within talker than between talker. This suggests that there may be limits to the representativeness of lab-stimuli, which could be due to factors outside of the speaker's control (e.g., not actually talking to and/or around an infant), or due to recording parameters (e.g., length of the recording session or mic placement). Future research will need to directly explore whether any of these acoustic properties are particularly relevant for word learning, together or separately.

Broadly, these comparisons are important for understanding how results from lab experiments might translate to the real world: our results suggest generalizability for patterns of variability of pitch-based measurements from lab stimuli to naturalistic tokens from infants' daily lives. Further, despite hearing primarily speech from their mothers, our infants experienced more pitch-based variability from their other talkers relative to their top talker. This suggests that part of the advantage of experiencing multiple talkers may be the opportunity to learn about the limits of pitch variability across them. The analyses presented here were the necessary first step for describing this dimension of infants' everyday input and lay the groundwork for subsequent work aimed at testing whether talker variability based input properties shape (word) learning in the real world as well as in the lab. Further, this future research will need to examine whether variability from different sources plays distinct roles in learning, or whether hearing minimally variable speech is sufficient to support acquisition. That said, this snapshot of talker variability reflects a fairly homogeneous, U.S. population; the patterns we report may indeed differ cross-culturally and cross-linguistically. While a recent study in German builds on existing work in English, showing that talker variability helps infants learn minimal pairs in the lab (Hohle et al., 2020), this certainly does not mean that talker-based acoustic variability is similarly available in infants' daily environments around the world. In fact, prior research suggests that various voice-quality parameters (i.e., fundamental frequency, harmonics-to-noise ratio) vary as a function of the speaker's native language (Wagner & Braun, 2003).

Relatedly, previous work suggests that across communities children hear very different quantities of language input directed to them (Lieven, 1994). For example, Mayan children (Shneidman & Goldin-Meadow, 2012) and Tsimane infants (Casillas, Brown, & Levinson, 2019) and children (Cristia, Dupoux, Gurven, & Stieglitz, 2019) have been shown to hear roughly half as many utterances directed to them compared to U.S. children (Shneidman & Goldin-Meadow, 2012). The prevalence of child-directed speech also varies cross-linguistically for urban compared to rural or indigenous samples and by age (e.g., Casillas et al., 2019). While our results revealed that token counts and acoustic measurements showed modest but robust correlations, this may be due at least in part to the relatively high numbers of tokens infants heard on average, both from other-talkers (mean = 23) and from their

top-talker (mean = 48). It is not yet clear how this link between acoustic variability and token count would play out in a context where infants heard many fewer words in general; in principle, this could lead to either lower or higher levels of acoustic variability. To fully understand the robustness of patterns of talker variability and its role in learning much more cross-cultural and cross-linguistic work is needed (Bunce et al., under review in prep; e.g., Cychosz et al., 2021; Frank et al., 2017).

It is worth noting that the acoustic measurements used here are not lexically contrastive (i.e. phonemic) in English: talkers do not vary these properties to convey differences in word meaning. For instance, changes in pitch do not (in English) change the meaning of a word, as the word /baby/ produced at 200 Hz and at 500 Hz are still instances of the word baby. In fact, the measurements we used were selected deliberately by Galle et al. (2015) (and by us) to capture variability in the speech signal that should *not* be relevant for distinguishing words. Other acoustic measurements that are phonemic in English, like voice onset time, have also been shown to vary within and across talkers, for controlled and naturalistic speech (Chodroff & Wilson, 2017). As noted above, VOT variability has been shown not to facilitate novel word learning in the same minimal pair task where talker variability facilitates it (Rost & McMurray, 2010). While measuring voice onset time is beyond the scope of the present analysis, it remains an interesting open question whether this type of contrastive measurement would exhibit differential levels of variability for tokens from the top- or other-talkers, or whether its systematic use to differentiate meaning (e.g., pin vs. bin) would result in more similar variability across these talker-types.

Lastly, beyond thinking about how variability may vary cross-culturally and cross-linguistically, how acoustic variability patterns in multidialectal and multilingual environments is another important question for future research. For example, accented speech could further increase the acoustic variability infants hear but may also intersect with contrastive dimensions of the input. While previous research suggests that infants exposed to multiple accents in their daily lives perform differently on word-recognition tasks relative to age-matched peers who only have experience with a single accent (van Heugten & Johnson, 2017), this could be a result of having to sort out when the acoustic variability signals invariance, and when it signals a contrastive dimension of the input that they should attend to. This problem is further highlighted in multilingual language acquisition. How acoustic variability stemming from talker differences interacts with structural variability from different languages remains to be explored, and likely varies for any individual infant. The range of variability within and across languages could vary widely based on how many speakers of each language a child is exposed to (e.g., infants may not receive much "other-talker" variability for a minority language).

Taken together, our frequency-based and acoustic analyses have highlighted wide-ranging and independent sources of variability in the word tokens in infants' environments, even within our relatively homogeneous longitudinal sample of infants. Despite this variability, similar types of acoustic properties that have been shown to shape infant word learning in the lab are also present in naturalistic settings, setting the stage for investigating whether and how talker variability shapes word learning in the real world.

Notes

1 Both twins were included as they do not receive identical input, for example, have different nap times and "high-talk" portions of their days (see details on annotation procedure below).
2 This choice was made in consultation with the original works' lead author (McMurray, personal communication). This occurred on February 10, 2020.
3 Conducting the analyses with the slope constraints revealed the same pattern of results reported below, but with a less representative sample of tokens.
4 All families produced both dog and doggy. For the non-contrastive acoustic measurements, means, ranges, and SDs were calculated from dog and doggy values pooled together. For the holistic measurement of sound similarity, we compared all tokens of dog to other tokens of dog, and all tokens of doggy to other tokens of doggy, then averaged these together to get a single value for dog(gy).
5 See the Supporting Information for analyses on a subset of data looking at these patterns across age.
6 For thoroughness, we provide means and ranges for each measurement by talker-type, see Supplemental Table 1.

Acknowledgments

This work was supported by grants to EB (NIH-OD, DP5 OD019812-01) and FB (NIH-NICHD, F32 HD101216). We wish to thank all of the research assistants at Duke University who aided in data coding. We also wish to thank Elliott Moreton and Bob McMurray for their feedback and suggestions in the design and execution of this project. The authors have no conflict of interest to disclose.

References

Albin, A. (2014). An architecture for controlling the phonetics software "Praat" with the R programming language. *Journal of the Acoustical Society of America*, *135*(4), 2198.

Benders, T., StGeorge, J., & Fletcher, R. (2021). Infant-directed speech by Dutch fathers: Increased pitch variability within and across utterances. *Language Learning and Development*, *17*(03), 292–325. https://doi.org/10.1080/15475441.2021.1876698

Bergelson, E. (2017). Bergelson Seedlings HomeBank Corpus. Retrieved from https://doi.org/10.21415/T5PK6D

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2018). Day by day, hour by hour : Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. https://doi.org/10.1111/desc.12715

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921. https://doi.org/10.1073/pnas.1712966114

Bergmann, C., & Cristia, A. (2018). Environmental influences on infants ' native vowel discrimination: The case of talker number in daily life. *Infancy*, *23*(4), 484–501. https://doi.org/10.1111/infa.12232

Boland, J. E., Kaan, E., Valdés Kroff, J., & Wulff, S. (2016). Psycholinguistics and variation in language processing. *Linguistics Vanguard*, *2*(s1), 3–12. https://doi.org/10.1515/lingvan-2016-0064

Braginsky, M., Sanchez, A. & Yurovsky, D. (2019). Childesr: Accessing the 'CHILDES" database. Retrieved from https://cran.r-project.org/web/packages/childesr/index.html

Bulgarelli, F. & Bergelson, E. (accepted). Talker variability shapes early word representations in English-learning 8-month-olds. Infancy.

Bulgarelli, F., & Bergelson, E. (2020). Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic day-long recordings. *Behavioral Research Methods*, *52*, 641–653. https://doi.org/10.3758/s13428-019-01265-7

Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Florencia, A., Migdalek, M. & Casillas, M. (under review). A cross-cultural examination of young children's everyday language experiences. Retrieved from https://psyarxiv.com/723pr/

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weis, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*, 159–199. https://doi.org/10.1016/0885-2014(95)90008-X

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early language experience in a Tseltal Mayan village. *Child Development*, *91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001

Choi, M., & Shukla, M. (2021). A new proposal for phoneme acquisition: Computing speaker-specific distribution. *Brain Sciences*, *11*(2), 177. https://doi.org/10.3390/brainsci11020177

Cooper, R. P., & Aslin, R. N. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, *65*(6), 1663–1677. https://doi.org/10.1111/j.1467-8624.1994.tb00841.x

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, *90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., … Seidl, A. (2021). Vocal development in a large-scale cross-linguistic corpus. *Developmental Science*, *e13090*, 1–48. https://doi.org/10.1111/desc.13090

Estes, K. G., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*(11), 1–12. https://doi.org/10.1037/a0039725

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, *81*(5), 1376–1383.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), 1–185. Retrieved from https://www.jstor.org/stable/pdf/1166093.pdf?refreqid=excelsior%7B/%7D3A28b49b2d69ee2a5cd880edbb428aad5b

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, *20*(1), 104–113. https://doi.org/10.1037/0012-1649.20.1.104

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Langauge*, *16*(3), 477–501.

Frank, M. C., Bergmann, C., Cristia, A., Floccia, C., Hamlin, J. K., Hannon, E. E., & Soderstrom, M. (2017). A collaborative approach to infant research : Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development*, *11*(1), 66–79. https://doi.org/10.1080/15475441.2014.895249

Gentner, D. (1982). Why nouns are learned before verbs. *Language Development*, *2*, 301–334.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, *31*(7), 1–24. https://doi.org/10.18637/jss.v031.i07

Gleason, J. B. (1975). Father and other strangers: Men's speech to young children. In D. P. Dato (Ed.). *Developmental psycholinguistics: Theory and applications* (pp. 289–297). Washington, DC: Georgetown University Press. http://hdl.handle.net/10822/555468

Hohle, B., Fritzsche, T., Meb, K., Philipp, M., & Gafos, A. (2020). Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Developmental Science*, *23*(5), e12950. https://doi.org/10.1111/desc.12950

Houston, D. M. (1999). *The role of talker variability in infant word representations (Unpublished doctoral dissertation)* (Ph.D. thesis). Johns Hopkins University, Baltimore, MD.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582. https://doi.org/10.1037/0096-1523.26.5.1570

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252. https://doi.org/10.1121/1.1288413

Lieven, E. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway & B. Richards (Eds.), *Input and interaction in language acquisition* (pp. 56–73). Cambridge, England: Cambridge University Press.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017-Augus (pp. 498–502). Lous Tourils, France: ISCA. https://doi.org/10.21437/Interspeech.2017-1386

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. https://doi.org/10.1016/j.lingua.2011.04.006

Moore, C. & Bergelson, E. (under review). Examining the roles of regularity and lexical class in 18—26-month-olds' representations of how words sound. Retrieved from https://osf.io/n3phk/

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. https://doi.org/10.1121/1.397688

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Quam, C., Knight, S., & Gerken, L. (2017). The distribution of talker variability impacts infants' word learning. *Laboratory Phonology*, *8*(1), 1.

Richtsmeier, P. T., Gerken, L., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition*, *111*(3), 372–377. https://doi.org/10.1016/j.cognition.2009.02.009

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x.Speaker

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Schmale, R. L., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: flexibility of early word representations. *Developmental Science*, *70*(1), 0718. https://doi.org/10.1111/j.1467-7687.2009.00809.x

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants' phonotactic learning. *Language Learning and Development*, *10*(4), 1–24. https://doi.org/10.1080/15475441.2013.858575

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*(2), 833–870. https://doi.org/10.1016/j.cognition.2007.05.002

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, *51*(2), 173–189. https://doi.org/10.1016/j.jml.2004.04.004

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–383.

Tripp, A., Feldman, N. H., & Idsardi, W. J. (2021). Social inference may guide early lexical learning. *Frontiers in Psychology*, *12*(May), 1–19. https://doi.org/10.3389/fpsyg.2021.645247

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the Switch task. *Developmental Psychology*, *55*(5), 934–950.

van Heugten, M., & Johnson, E. K. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *Journal of Speech, Language, and Hearing Research*, *55*(2), 554–560. https://doi.org/10.1044/1092-4388(2011/10-0347)

van Heugten, M., & Johnson, E. K. (2017). Input matters: Multi-accent language exposure affects word form recognition in infancy. *The Journal of the Acoustical Society of America*, *142*(2), EL196–EL200. https://doi.org/10.1121/1.4997604

Wagner, A. & Braun, A. (2003). Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages. In Proceedings of the 15th International Congress of Phonetic Sciences, (May) (pp. 651–654). London: International Phonetic Association. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_0651.pdf

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table 1: Mean and ranges for the acoustic measurements for top- and other-talker tokens in the current dataset

Table 2: By word means for the acoustic measurements for all tokens in the current dataset

Figure 1: Acoustic measurements for top-talker (red) and other-talkers (blue) tokens for each word.

Table 3: Standard Deviations of acoustic measurements for other-talkers and top-talker tokens in the current dataset split at 1 year of age (6-11months vs. 12-17 months) for non-contrastive acoustic measurements and Mean normalized acoustic distance for all tokens.

Table 4: Acoustic analyses comparing variability in tokens of 'ball' in the main dataset (i.e., the naturalistic corpus), and in variably-produced lab-recorded stimuli.

Table 5: Acoustic analyses comparing variability in tokens of 'dog' in the main dataset (i.e., the naturalistic corpus), and in variably-produced lab-recorded stimuli.