Developmental Psychology
2015, Vol. 51, No. 11, 1517–1528

0012-1649/15/$12.00 http://dx.doi.org/10.1037/a0039725

Listening Through Voices: Infant Statistical Word Segmentation Across Multiple Speakers

Katharine Graf Estes
University of California, Davis

Casey Lew-Williams
Princeton University

To learn from their environments, infants must detect structure behind pervasive variation. This presents substantial and largely untested learning challenges in early language acquisition. The current experiments address whether infants can use statistical learning mechanisms to segment words when the speech signal contains acoustic variation produced by changes in speakers' voices. In Experiment 1, 8- and 10-month-old infants listened to a continuous stream of novel words produced by 8 different female voices. The voices alternated frequently, potentially interrupting infants' detection of transitional probability patterns that mark word boundaries. Infants at both ages successfully segmented words in the speech stream. In Experiment 2, 8-month-olds demonstrated the ability to generalize their learning about the speech stream when presented with a new, acoustically distinct voice during testing. However, in Experiments 3 and 4, when the same speech stream was produced by only 2 female voices, infants failed to segment the words. The results of these experiments indicate that low acoustic variation may interfere with infants' efficiency in segmenting words from continuous speech, but that infants successfully use statistical cues to segment words in conditions of high acoustic variation. These findings contribute to our understanding of whether statistical learning mechanisms can scale up to meet the demands of natural learning environments.

Keywords: language acquisition, statistical learning, variation, word segmentation

Learners face the daunting challenge of detecting structure amid wide variation in their environments. Exemplars of concepts, events, words, and grammatical patterns vary in many ways that obscure a clear delineation of structure. For example, items within an object category can differ in color and size, repetitions of events vary in actors and locations, and tokens of words differ across sentential contexts and superficial voice characteristics. Surface form variation occurs when the perceptual details of elements in the input are not identical even when they belong to the same category. To identify underlying structure, learners must process surface variation in ways that permit the detection of what is consistent, relevant, and meaningful across exemplars. The present research investigates how variation in voice—a salient source of variability in natural language environments—affects infants' detection of statistical regularities in fluent speech, which is a fundamental process in language acquisition. The experiments also address an outstanding issue in language acquisition research: whether statistical learning mechanisms can meet the kinds of processing demands that infants face in their natural language environments.

The demands of coping with surface variation are illustrated by infants' early attempts to recognize words. Each time a word is produced, its acoustic form differs depending on who says the word, the speaker's affect, speaking style, and speaking rate, as well as the sentence context. Importantly, infants do not begin learning language with a priori knowledge of which sound variants refer to a common concept and which variants distinguish between words. Early in development, infants have substantial difficulty recognizing words across surface variation. For example, in a word segmentation task conducted by Singh, Morgan, and White (2004), 7.5-month-olds failed to recognize native-language words across a change in affect (e.g., isolated words spoken with positive affect during familiarization but neutral affect during testing within fluent passages). The same pattern of results occurred when the words changed in voice pitch (Singh, White, & Morgan, 2008) or speaker gender (Houston & Jusczyk, 2000, 2003). Examining a comparable surface-level change, Schmale and colleagues reported that a change in dialect from familiarization to testing disrupted word recognition in 9-month-olds (Schmale, Cristia, Seidl, & Johnson, 2010; Schmale & Seidl, 2009). Varying speakers' voices has also been shown to interfere with lexical processing in adults (Goldinger, 1998; Goldinger, Pisoni, & Logan, 1991; Palmeri, Gold-

This article was published Online First September 21, 2015.

Katharine Graf Estes, Department of Psychology, University of California, Davis; Casey Lew-Williams, Department of Psychology, Princeton University.

This research was supported by grants to Katharine Graf Estes from the National Science Foundation (BCS0847379), the National Institute of Child Health and Human Development (HD062755), and the Hellman Foundation, and to Casey Lew-Williams from the National Institute of Child Health and Human Development (HD069094). We thank Carolina Bastos, Stephanie Chen-Wu Gluck, and the members of the Language Learning Lab at the University of California, Davis, for their assistance with this research. We would also like to thank Dylan Antovich, Erik Thiessen, and Lisa Oakes for their input and discussion regarding this work. We also thank the parents who generously contributed their time.

Correspondence concerning this article should be addressed to Katharine Graf Estes, Department of Psychology, University of California, Davis, 1 Shields Avenue, Davis, CA 95615. E-mail: kgrafestes@ucdavis.edu

inger, & Pisoni, 1993). These lines of work have been crucial for revealing the potential for surface variation to interfere with learning. If infants fail to detect recurrences of words across acoustic variations, they should have difficulty recognizing words as they occur across contexts, and these contexts are key for determining the full meanings and grammatical roles of words.

Statistical word segmentation presents an excellent test case for examining how variation affects learners' ability to detect structure in complex, novel patterns. Furthermore, investigating how acoustic variation affects statistical word segmentation is essential for understanding the utility of statistical learning mechanisms for supporting natural language processing. In statistical word segmentation, infants exploit patterns in syllable co-occurrences to detect words in fluent speech. In natural speech, the transitional probability from one syllable to the next (defined as the frequency of the sequence XY given the frequency of X) tends to be higher within words than across word boundaries (Harris, 1955; Swingley, 2005). Syllables that occur together consistently (i.e., with high transitional probability) indicate reliable sequences that (likely) form cohesive words in the language. In contrast, across word boundaries, the probabilities tend to be substantially lower because a given word can be surrounded by many other words, as in the phrases happy *baby*, crying *baby*, *baby* sister, and *baby* laughing. For example, in infant-directed speech, approximately 80% of the time infants hear the syllable "pre" it is followed by the syllable "ty," as in the word "pretty"; the probability that "ty" is followed by "ba," as in the phrase "pretty baby," is only around 0.03% (Saffran, 2003). Thus, syllable transitional probabilities are patterns that infants can use to initially detect words in the speech stream, which allows them to access other language-specific word boundary cues, such as lexical stress (Thiessen & Saffran, 2003).

Surface form variation compounds the complexity of detecting syllable sequence patterns. To use statistical regularities to segment words, infants must recognize acoustically distinct tokens of syllable sequences when they occur in different voices and utterances. For example, the sequence "baby" sounds different depending on who says it (mother, father, grandparent), their speaking style (happy, sad, infant-directed, adult-directed), and the sentential context in which it occurs (in isolation, in fluent speech, in utterance initial, final, or medial position). Such inter- and intraspeaker variability is a key feature of natural language learning environments. Infants must detect occurrences of disparate word tokens to take advantage of syllable transitional probability patterns that indicate word boundaries. The present experiments investigate this challenge by scrutinizing the conditions of voice variability that enable versus hinder infants' success in segmenting words.

This work explores an intriguing paradox in the effects of variation on information processing. Prior evidence indicates that surface form variation disrupts infant and adult lexical processing (e.g., Houston & Jusczyk, 2000; Mullennix, Pisoni, & Martin, 1989; Singh et al., 2004), but high variation can sometimes promote the detection of underlying commonalities. For example, in early word recognition, after hearing words produced by one speaker with high variation in affect (Singh, 2008) or produced by a wide range of speakers (Houston, 1999), infants can effectively generalize representations to recognize words in novel and distinct voices. In contrast, when infants hear words produced with low variation in affect or speakers' voices, they fail to perform this generalization. Similarly, providing infants with high variation in the voices producing minimal-pair object labels (i.e., labels that differ by a single phoneme) facilitates learning (Rost & McMurray, 2009), but when infants hear minimal-pair labels with low variation in the tokens of the labels, infants fail to learn them (Werker, Fennell, Corcoran, & Stager, 2002). There is also recent evidence that high variation in voices promotes infants' detection of phonotactic patterns (Seidl, Cristia, & Onishi, 2014). Furthermore, across linguistic and nonlinguistic domains and across infancy and adulthood, learning is more robust in conditions of high variation than in conditions of low variation (Gómez, 2002; Needham, Dueker, & Lockhead, 2005; see also Quinn & Bhatt, 2005). Whereas exposure to multiple tokens of category members promotes stronger generalization to new exemplars and stronger knowledge of underlying structure, low variation may promote attention to differences among exemplars.

In a series of experiments, we examined whether conditions of high and low variation in surface form details (i.e., speakers' voices) affect infants' ability to detect patterns of syllable transitional probabilities for word segmentation. We first tested whether infants could determine the statistical regularities that mark word boundaries in fluent speech when the speech stream incorporated several different speakers. In Experiment 1, infants heard an artificial language consisting of monotone, continuous speech in which transitional probability patterns provided the only reliable word segmentation cue (Aslin, Saffran, & Newport, 1998). However, the speech stream contained a high degree of acoustic variation. The language was produced by eight different voices that changed frequently, after every 10 to 20 syllables—somewhat analogous to conversational turn-taking.

We tested 7- to 8- and 10- to 11-month-olds because these ages are similar to those used in tests of infant statistical learning and word recognition (Houston & Jusczyk, 2000; Pelucchi, Hay, & Saffran, 2009; Saffran, Aslin, & Newport, 1996; Singh, 2008; Singh et al., 2004). Across previous studies, infants have displayed reliable statistical learning abilities, but these ages also represent points of change in how variation affects word recognition. At around 8 months of age, change in a speaker's voice disrupts word recognition for native-language words, but at around 10 to 11 months of age, recognition is more robust to changes in voice (Houston & Jusczyk, 2000; Schmale & Seidl, 2009; Singh et al., 2004; Singh, White, et al., 2008). Thus, the ages of participants in our experiments were selected to examine the development of sensitivity to high and low voice variation.

We predicted that infants' statistical learning abilities would be robust to variation across several voices. The key task in the experiment was to attend to syllable-level transitional probabilities to segment words, while disregarding variation in voice or integrating learning across multiple voices. Following from prior demonstrations of the facilitative role of voice variation, the high variation in speakers in Experiment 1 should support infants' attention to syllable patterns, prevent focus on voice features, and thus prevent comparisons across voices or separation of information presented by each voice.

By testing these predictions, the present study allows us to address an unresolved issue in research on language acquisition: whether or not statistical learning can scale up to meet the kinds of challenges that infants face in their natural language environments (Frank, Tenenbaum, & Gibson, 2013; Graf Estes, 2012; Johnson &

Tyler, 2010; Pelucchi et al., 2009). Learning from distributional patterns, or statistical learning, is hypothesized to contribute to the acquisition of many levels of linguistic structure, including phonemes (Maye, Werker, & Gerken, 2002; Yoshida, Pons, Maye, & Werker, 2010), words (Graf Estes, Evans, Alibali, & Saffran, 2007; Lany & Saffran, 2011; Lew-Williams, Pelucchi, & Saffran, 2011), and grammar (Gerken, Wilson, & Lewis, 2005; Mintz, 2003; Saffran & Wilson, 2003). It may also underlie infants' learning of the structure of visual stimuli (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002; Roseberry, Richie, Hirsh-Pasek, Golinkoff, & Shipley, 2011). To date, most of our knowledge of statistical learning derives from well-controlled experimental paradigms using input with relatively simplified structure. This input lacks certain dimensions of complexity and variation inherent in "real world" input. Although it is clear that infants are skilled at extracting structure across stimulus domains, we know relatively little about how infants deploy statistical learning mechanisms when they encounter the complexity-related challenges that are built into their natural environments.

Previous investigations have yielded evidence both supporting and questioning the scalability of statistical learning mechanisms, particularly in experiments that introduced word-length variability. When presented with an artificial language containing two- and three-syllable words, Johnson and Tyler (2010) found that infants failed to find word boundaries, suggesting that infants may not use transitional probability patterns effectively across words with varying lengths. However, other studies show that infants successfully track probability patterns across varying word lengths when highly familiar words appear intermittently in artificial languages (Mersad & Nazzi, 2012), and when listening to natural, highly variable sentences (Lew-Williams et al., 2011; Pelucchi et al., 2009). The findings regarding word length variability highlight the importance of incorporating natural language characteristics into statistical learning tasks. Here, we extend this question of scalability using a pervasive feature of natural language environments, variation in voices. Mothers, fathers, siblings, friends, other caregivers, and other family members all speak to infants, and infants must "listen through" the noise to determine the relevant structures in speech. By examining how infants integrate their learning of patterns across voices, this investigation represents a rigorous test of whether statistical learning supports language acquisition. In four experiments, we reveal both robustness and weakness in how infants cope with surface variation during statistical word segmentation.

Experiment 1

Infants participated in a statistical word segmentation task in which eight different female voices presented a fluent speech stream that contained transitional probability cues to word boundaries. If infants can track patterns of syllable occurrences despite changes in voice, they should successfully segment words from the speech stream and recognize them during testing. Word segmentation performance was measured by comparing the infants' listening duration (operationalized as looking time) to the statistically defined, high transitional probability words from the speech stream versus the low-probability sequences that crossed word boundaries in the stream, termed "part-words." If infants segment the speech, their listening times should differentiate between the relatively familiar words and novel part-words.

Method

Participants. Fifty infants participated in Experiment 1. Half were 7- to 8-month-olds (M = 8.3 months; range = 7.8–8.8; 12 females) and half were 10- to 11-month-olds (M = 10.8 months; range = 10.3–11.0 months; 14 females). All infants came from English-speaking homes; at least 95% of their language exposure was to English. Ten infants (n = 5 at each age) heard between 1 to 4 hr per week of a second language at home or in childcare. Before analyzing the data, 10 additional infants were excluded because of fussiness (i.e., crying, whining, hiding eyes; n = 4), excessive movement (n = 5), or experimenter error (n = 1). One additional infant in each age group was identified as an outlier (listening time difference to words vs. part-words greater than 2.5 *SD* from the full group mean) and was excluded from analyses.

Stimuli.

Segmentation phase. During segmentation, infants listened to an artificial language consisting of a continuous stream of four disyllabic novel words. To control for arbitrary listening preferences for test items, there were two counterbalanced versions of the language (originally designed by Graf Estes et al., 2007). The words in Language A were *timay*, *dobu*, *gapi*, and *moku.* The words in Language B were *pimo*, *kuga*, *buti*, and *maydo*. The counterbalancing allowed the same test items to be used for both language versions. As shown in Table 1, the test items that were words in Language A were part-words in Language B, and vice versa. Infants were randomly assigned to listen to Language A or B.

The artificial language had a frequency-balanced design to equate the frequency of the word and part-word test items during the segmentation phase, but maintain the difference in internal transitional probability (Aslin et al., 1998). In the language, two of the words occurred with high frequency, 180 times each, and two occurred with low frequency, 90 times each. All words had transitional probabilities of 1.0, because the syllables occurred together with perfect consistency. For example, in Language A, the low-frequency words were *timay* and *dobu*, and the high-frequency words were *gapi* and *moku.* An important component of the design is that half of the occurrences of each high frequency word preceded another high-frequency word (e.g., *gapi#moku* and *moku#gapi).* This formed two frequently occurring part-words (*pimo* and *kuga*) that crossed the word boundaries exactly 90 times.

Thus, the test items consisted of the low-frequency words and the part-words formed from the conjunction of the high-frequency words. Given this structure, part-word test items had internal transitional probabilities of .5; the word test items had internal transitional probabilities of 1.0, but all test items occurred with

Table 1
Word and Part-Word Test Items for Experiments 1 and 2

Language	Words	Part-words
A	*timay*, *dobu*	*pimo*, *kuga*
B	*pimo*, *kuga*	*timay*, *dobu*

equal frequency in the speech stream. This design provides a stringent test of whether infants attend to familiarity or probability during segmentation. The low probability of the part-words indicates that they occur across word boundaries, whereas the high probability of the words indicates that they are coherent units.

Eight female English speakers recorded speech samples to create the artificial language. The speakers were between the ages of 19 and 26. They were selected to present variation in female voices. All began learning English from birth or as young children. Although four speakers were bilingual (learning Mandarin Chinese, Portuguese, or Polish from infancy or Korean from childhood), they were fluent in English and did not have strong nonnative accents, as confirmed by adult ratings discussed below. Table 2 shows the mean speaking rate (measured by syllable length) and pitch (measured by fundamental frequency, F0) for each speaker.

To examine possible perceptual differences between the voices, we collected measurements from 10 adult listeners. We created fluent 5 second audio files of each speaker producing the artificial language (Language A). The listeners heard audio files from two different speakers, and then judged the similarity of the voices on a scale from 1 (*very similar*) to 5 (*very different*). Each listener heard all possible voice pairings. Table 3 shows the mean similarity ratings for each voice. The overall mean similarity rating was 3.35 (SD = 1.26), indicating that the voices sounded moderately different from one another. Across speakers, there was a relatively small range of similarity scores, suggesting that no voice stood out as sounding highly distinct from the other voices.

To compare the voices from the artificial language against the test items, which were presented in a novel female voice, the listeners also heard each audio file presented against the four test items from Language A (with 500 ms of silence between each item). The mean similarity rating for the test items was 3.54 (SD = 1.12), which was similar in magnitude to the comparison of voices within the artificial language (Table 3). Because four of the speakers were bilingual from childhood, we also asked the listeners to judge the degree to which each speaker sounded like a native English speaker. They used a scale from 1 (*definitely a native speaker*) to 5 (*definitely not a native speaker*). The bilingual speakers (speakers 1, 4, 5, 6) had a mean native speaker judgment score of 2.80 (SD = .74); the monolingual speakers had a mean score of 2.38 (SD = .62). The difference across groups was not significant, $t(9) = 1.87$, $p = .094$. Thus, the bilingual speakers were not rated as significantly less native-sounding than the monolingual English speakers.

To create the artificial language, each speaker recorded monotone three-syllable sequences that incorporated all of the possible coarticulation contexts present in the language. Each speaker was instructed to maintain a consistent pitch and speaking rate. All of the syllables were normalized for intensity using Praat, and the middle syllables were excised and spliced together to form a fluent speech stream. This recording and splicing method reduced the chance for the speakers to inadvertently introduce supplemental word boundary cues. The speech stream did not contain pauses or other reliable acoustic cues to word boundaries. Transitional probability patterns provided the only consistent cue.

To manipulate speaker variation, syllables from each speaker were spliced together in sequences that ranged from 10 to 20 syllables long, lasting between 3 and 7 s. The order of speakers was randomized. We created the stimuli using eight speakers to reduce the possibility that infants could segment the words by listening to a single voice. Across the eight speakers, the average number of complete repetitions of the two low-frequency words (i.e., the test items) was 10 per speaker; the average number of complete repetitions for the two high-frequency words was 21 per speaker. The remaining 40 word tokens occurred in midword speaker changes. Therefore, if infants segregate input by voice, they would hear few tokens of each item. But if they recognize words across voices, they would hear vastly more repetitions of each item.

Table 2
Mean (SD) Pitch (Fundamental Frequency, F0) and Syllable Length for Each Speaker in Languages A and B

Language	Pitch (F0)	Syllable length (ms)
A		
Speaker 1	219 (.17)	331 (2.46)
Speaker 2	223 (2.63)	**310 (9.15)**
Speaker 3	212 (2.25)	351 (2.88)
Speaker 4	215 (.79)	325 (8.61)
Speaker 5	202 (.70)	350 (17.2)
Speaker 6	**231 (1.46)**	**357 (3.61)**
Speaker 7	211 (3.46)	348 (5.43)
Speaker 8	**177 (.43)**	346 (3.08)
Overall	211 (14.88)	340 (17.35)
B		
Speaker 1	217 (.01)	346 (2.90)
Speaker 2	221 (1.79)	**297 (4.25)**
Speaker 3	210 (.36)	329 (2.13)
Speaker 4	215 (.49)	323 (3.61)
Speaker 5	208 (.32)	**352 (7.36)**
Speaker 6	**223 (.85)**	344 (5.28)
Speaker 7	216 (.58)	343 (14.23)
Speaker 8	**177 (.76)**	344 (2.90)
Overall	211 (14.82)	335 (18.09)

Note. Bolded items represent the highest and lowest values for each characteristic in a given language.

Table 3
Perceptual Ratings for Similarity to Other Voices in the Artificial Language, Similarity to the Test Voice, and Similarity to Native English Speakers

Speaker	M (SD) similarity to other voices	M (SD) similarity to test voice	M (SD) native-like rating
1	3.15 (1.12)	3.70 (1.25)	2.80 (1.22)
2	3.17 (1.22)	3.10 (1.45)	3.70 (1.06)
3	3.09 (1.14)	3.10 (1.10)	2.20 (1.23)
4	3.41 (.99)	4.10 (.88)	3.00 (.82)
5	3.24 (.98)	3.10 (1.10)	2.60 (1.26)
6	3.23 (1.29)	3.20 (1.48)	2.80 (1.03)
7	3.38 (1.07)	3.70 (.95)	2.40 (1.07)
8	4.11 (.89)	4.30 (.82)	1.20 (.42)
Overall	3.35 (1.13)	3.54 (1.12)	2.59 (1.01)

Note. The similarity ratings used a scale from 1 (*very similar*) to 5 (*very different*). The native English-speaker judgments were based on asking how much each speaker sounded like a native English speaker from 1 (*definitely a native speaker*) to 5 (*definitely not a native speaker*).

There were 71 voice changes across the duration of the 6-min speech stream (6:06 min for Language A, 6:03 min for Language B), which played at a level approximating conversational speech, around 65 dB. To prevent the changes in voice from introducing a word boundary cue, each individual sequence was randomly assigned to start at a word onset or at the second syllable; 31 (out of 71) voice changes occurred at word onsets and 40 changes occurred within words. This means that 7% of the full set of words in the language (40 of 540 total tokens) occurred with a midword speaker change. In natural speech, infants may encounter midword speaker changes when one person interrupts another, but such changes are likely to be uncommon in infants' input. However, we used this manipulation because it was important to ensure that the word boundary information was provided by transitional probabilities, not by the acoustic cue of the voice change. Changing speakers consistently at word boundaries would provide supplemental information to facilitate segmentation. While infants' use of voice as a segmentation cue is an interesting issue in itself, the goal here was to investigate infants' use of statistical cues to word boundaries above and beyond voice variation.

Test phase. The test items were recorded by a female speaker who did not produce syllables for the segmentation phase. She recorded the items in citation form (i.e., as isolated tokens) and she was instructed to use a monotone speaking style to mimic the speech in the segmentation phase. Her mean pitch was 234 Hz, which is slightly higher than the range of the female voices used during the segmentation phase. Repetitions of the test items were separated by 750 ms of silence.

For the visual stimulus during the test phase, the word and part-word test items were paired with a consistent visual animation of an orange oval rotating in a circle. Attention to the audio-visual stimuli (i.e., looking time) was used to measure infants' listening time to the test items.

Procedure. During the segmentation phase, the infant and his or her parent sat in a small sound-attenuated room and were allowed to play quietly with a small set of toys. The artificial language exposure was passive and had a fixed length (6 min). The parent was instructed to talk as little as possible and to refrain from discussing the artificial language playing over the loudspeakers. Following the segmentation phase, the infant and parent moved to the test booth. The infant sat on the parent's lap approximately 3 ft from a large TV screen. The visual stimuli appeared within a white rectangle at the center of the screen and the sounds played over integrated speakers. A camera mounted below the screen allowed the experimenter, located in a separate booth, to monitor the infant's behavior. To prevent bias, the parent listened to masking music on headphones and the experimenter was blind to the identity of the stimuli being presented.

When the parent and child entered the test booth, the parent heard brief reminder instructions for the test phase. Because of this delay, infants heard a 30-s refamiliarization with the artificial language before testing. The refamiliarization played during a soundless animated cartoon clip to encourage the infant's attention. The cartoon and language played continuously regardless of the infant's looking behavior.

Test trials began immediately after the refamiliarization. The program Habit X (Cohen, Atkinson, & Chaput, 2004) was used to control the presentation of the test items in an infant-controlled auditory preference procedure. Each trial began with an attention-getting cartoon. When the infant looked at the screen, the experimenter triggered the presentation of a test trial consisting of repeated tokens of a word or part-word paired with a visual animation. The trial continued to play until the infant looked away for at least 1 s or looked for a maximum of 20 s. Each of the four test items (two words, two part-words) was presented four times in a pseudorandom order for each participant.

Habit X recorded the duration of the infant's attention to each test item. The measure of listening time to word and part-word test items used in the auditory preference procedure is similar to the measure used in other tasks with a central fixation point (Shi & Werker, 2001; Shi, Werker, & Cutler, 2006), as well as the headturn preference procedure used in many studies of word segmentation and statistical learning (e.g., Houston & Jusczyk, 2000; Saffran et al., 1996; Singh et al., 2004).

Preliminary analyses revealed no significant effects of gender or language version (A or B). Therefore, subsequent analyses collapsed across these variables. However, there was a significant effect of test block for the first eight trials (Block 1) versus the last eight trials (Block 2), with four word and four part-word trials per block. Test block was included in statistical analyses (for similar block effects see Gerken et al., 2005; Sahni, Seidenberg, & Saffran, 2010). Importantly, each test item occurred twice within each block, enabling a direct comparison between infants' performance in the first versus second half of the test phase.

Results and Discussion

A 2 (Age Group: 7–8 months vs. 10–11 months; between subjects) × 2 (Test Item: words vs. part-words; within subjects) × 2 (Test Block: 1 vs. 2; within subjects) mixed-design ANOVA of listening time was performed to examine learning. There were two significant effects. First, there was a significant main effect of test block, $F(1, 48) = 84.5$, $p < .001$, $\eta_p^2 = .64$. Infants listened significantly longer during Block 1 ($M = 10.36$ s, $SD = 3.16$) than during Block 2 ($M = 6.74$ s, $SD = 2.98$). Second, there was a significant interaction of test item and test block, $F(1, 48) = 14.21$, $p < .001$, $\eta_p^2 = .23$. The remaining main effects and interactions were not significant: test item ($p = .130$), age group ($p = .320$), Test Item × Age Group ($p = .924$), Test Block × Age Group ($p = .092$), Test Item × Test Block × Age Group ($p = .562$).

To explore the significant interaction of Test Item × Test Block, we analyzed performance separately for Blocks 1 and 2. In Block 1, infants listened significantly longer to words ($M = 11.05$ s, $SD = 3.65$) than to part-words ($M = 9.66$ s, $SD = 3.16$), paired samples $t(49) = 3.68$, $p < .001$, $d = .40$. In Block 2, the difference in listening time to words ($M = 6.45$ s, $SD = 2.81$) versus part-words ($M = 7.03$ s, $SD = 3.61$) was not statistically significant, $t(49) = -1.66$, $p = .103$, $d = .18$. Figure 1 shows the pattern of performance across ages and blocks. The graph illustrates the significant familiarity (word) preference that infants exhibited during the first block of trials. Follow-up analyses confirmed that the word preference held when each age group was analyzed separately: 10- to 11-month-olds listened longer to the word test items in Block 1, $t(24) = 2.22$, $p = .036$, $d = .41$. Seven- to 8-month-olds showed the same pattern, $t(24) = 2.96$, $p = .007$, $d = .40$. Figure 1 also illustrates that infants in both age groups showed no reliable preference during the second block, but the preference tended to shift toward a novelty (part-word) preference.

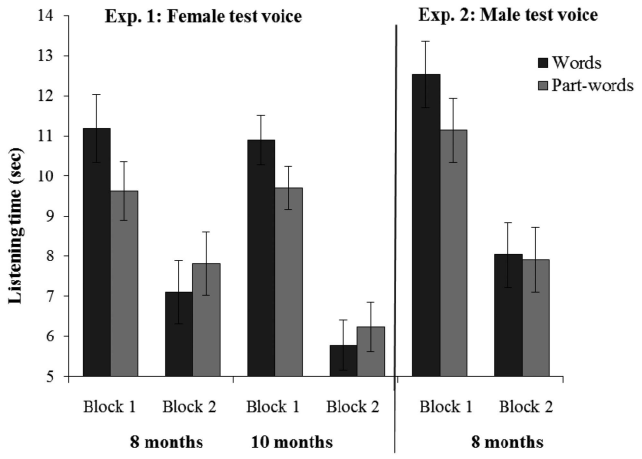

Figure 1. Mean listening time to word and part-word test trials for Experiments 1 and 2, separated by test block. Error bars indicate *SE*s.

This pattern was also consistent across ages: 10- to 11-month-olds showed no reliable preference in Block 2, $t(24) = -.973$, $p = .340$, $d = .16$, neither did 7- to 8-month-olds, $t(24) = -1.34$, $p = .194$, $d = .21$.

In sum, 7- to 8- and 10- to 11- month-old infants reliably differentiated the word and part-word test items when the language was presented by eight voices. The effect was strongest during the first block of test trials. Based on Hunter and Ames's (1988) model of infant attention, the familiarity preference exhibited in Block 1 is consistent with a pattern that infants display when a task is difficult or when they are still mastering new information (see also Houston-Price & Nakai, 2004; Hunter, Ames, & Koopman, 1983). The learning task in Experiment 1 was designed to be challenging, in that infants had to detect the consistent syllable patterns despite substantial acoustic variation. They then had to recognize the segmented words in a new voice and differentiate them from other highly frequent (but low probability) syllable sequences. Given the complexity of the task, it is not surprising that infants displayed learning via a familiarity (vs. novelty) preference for the statistically defined words.

There was also a tendency for preference patterns to change over the course of testing, as seen in previous work by Sahni et al. (2010) and Gerken et al. (2005). The reliable familiarity preference in the first block did not persist in the second test block. Based on Hunter and Ames's (1988) model, we propose that during testing, infants may have first recognized the word-like test items as the units they heard during segmentation, and these items initially held their attention. With additional exposure, infants had the opportunity to process the words repeatedly, reducing their interest in these words over time. Importantly, "training" and "testing" are investigator-applied labels; learning does not halt at the end of training. Infants presumably learned from exposure to each test item, which in turn affected their attention incrementally on subsequent test trials. The observed shift away from a familiarity preference showcases the importance of analyzing attention-based experimental measures across time.

The results of Experiment 1 demonstrate that by 7 to 8 months of age, infants' statistical segmentation abilities are robust to acoustic variation produced by changes in voice. To succeed in the task, infants also had to recognize new tokens of segmented words when they were produced by a novel voice. The voice was an additional female voice and was acoustically similar to the voices from the segmentation phase, particularly in fundamental frequency. To use words that are discovered via statistical learning, it is necessary for infants to generalize the representations of words across wider variation. Graf Estes (2012) reported that after segmenting words from a speech stream produced by a single consistent female voice, 10-month-olds recognized the words when they were produced by a male voice (see also Finley, 2013; Vouloumanos, Brosseau-Liard, Balaban, & Hager, 2012, for related findings with adults). Can infants also generalize representations of newly segmented words when *learning* takes place under conditions of surface variation? The combined challenges of learning amid acoustic variation plus recognizing words in a novel distinct voice may be too difficult for infants. Furthermore, the consistencies in the female voices that presented the segmentation phase in Experiment 1 may preclude generalization to a more acoustically distant voice. Alternatively, statistical learning may be sufficiently robust that infants can track syllable transitional probability patterns despite voice changes and subsequently recognize segmented words when spoken in a highly distinct voice. Experiment 2 tested these possibilities.

Experiment 2

In Experiment 2, infants listened to the same segmentation speech stream as in Experiment 1, which was produced by eight female voices. During testing, infants heard test items presented in a male voice. Infants were required to segment the words despite voice variation and then recognize tokens of the words in isolation when presented in a novel voice with significant perceptual distance from their initial learning experience. Seven- to 8-month-olds participated in Experiment 2 because this is close to the age at which generalization across voices interferes with word recognition in infants' native languages (Houston & Jusczyk, 2000; Schmale & Seidl, 2009; Singh et al., 2004). This age group allowed for a conservative test of whether statistical word segmentation can withstand voice variation during learning and word recognition.

Method

Participants. The participants were twenty-four 7- to 8-month-olds ($M = 8.1$ months; range = 7.6–8.5 months; 12 females). The infants met the same inclusion criteria as in Experiment 1. All were from English-speaking homes. Two infants heard a second language for 1 to 2 hr per week. Nine additional infants were excluded from the analyses because of fussiness ($n = 6$), excessive movement ($n = 2$), or experimenter error ($n = 1$). One additional infant was identified as an outlier using the same criteria as Experiment 1 and was excluded from the analyses.

Stimuli.

Segmentation phase. The infants listened to the same artificial languages as in Experiment 1.

Test items. The test items were the same words and part-words as in Experiment 1. In Experiment 2, the test items were produced by a male voice. The speaker was a monolingual English speaker 36 years of age. The mean pitch was 121 Hz, which is substantially

lower than the frequency of the female voice test items in Experiment 1 and the voices from the segmentation phase (Table 2).

Procedure. The procedure was identical to Experiment 1. Preliminary analyses revealed no differences in performance based on language version or gender. Subsequent analyses collapsed across these variables.

Results and Discussion

A 2 (Test Item: words vs. part-words) × 2 (Test Block: 1 vs. 2) repeated measures ANOVA of listening time was performed to examine learning. There was a significant effect of test block, $F(1, 23) = 30.1$, $p < .001$, $\eta_p^2 = .567$. Infants listened significantly longer during Block 1 ($M = 11.84$ s, $SD = 3.75$) than during Block 2 ($M = 7.97$ s, $SD = 3.50$). There was also a significant effect of test item, $F(1, 23) = 7.53$, $p = .012$, $\eta_p^2 = .247$. Infants listened significantly longer to the words ($M = 10.28$ s, $SD = 3.38$) than to the part-words ($M = 9.52$ s, $SD = 3.14$).

The interaction of test item and block was not significant, $F(1, 23) = 2.64$, $p = .118$, $\eta_p^2 = .103$. However, infants' performance in each block was examined separately to maintain consistency with Experiment 1. Figure 1 shows the infants' listening time separated by block. Similar to Experiment 1, the figure illustrates that the greater attention to words was primarily driven by Block 1, paired samples $t(23) = 2.52$, $p = .019$, $d = .35$. The difference between word and part-word trials was not significant in Block 2, $t(23) = .320$, $p = .752$, $d = .03$.

These findings indicate that 7- to 8-month-old infants recognized the statistically defined words when they were produced in a novel and acoustically distinct voice. They effectively confronted the added challenge of generalizing representations of words they had previously segmented from a varying speech stream.

Experiment 3

Experiments 1 and 2 showed that infants can statistically segment words amid acoustic variation and generalize representations of segmented words to a novel voice. The voice variation across eight speakers was high. However, a common language-learning scenario is for infants to hear two dominant voices (e.g., two parents). The results of previous experiments with infants and adults suggest that conditions of high variation should support learning when the important task is to extract invariant structure (e.g., Gómez, 2002; Lively, Logan, & Pisoni, 1993; Perry, Samuelson, Malloy, & Schiffer, 2010; Singh, 2008). But it is not yet clear whether the high surface variation present in Experiments 1 and 2 was essential for the infants to learn, or whether they would also be successful under conditions of lower variation. Low variation may encourage learners to focus on differences across exemplars, focusing attention on comparisons of features rather than underlying commonalities that signal relevant structure; we return to this idea in the General Discussion in reference to a recent exemplar-based model of memory. In some cases, low variation may promote learning, as in studies showing that low variation promotes learning of novel verbs (Maguire, Hirsh-Pasek, Golinkoff, & Brandone, 2008) and spatial categories (Casasola, 2005) via increased attention to relational patterns. But for statistical word segmentation, low variation in voice may boost the salience of the voices themselves, preventing infants from uniting the voices and tracking occurrences of syllables and words across them. That is, low variation in voice may promote infants' attention to irrelevant voice details at the expense of attending to the underlying syllable patterns. Thus, the resilience of statistical word segmentation to variation in voice may only hold under conditions of high surface form variation. Experiment 3 tested these possibilities by exposing infants to a speech stream spoken by just two speakers.

Method

Participants. The participants were twenty-five 7- to 8-month-olds ($M = 7.9$ months, range = 7.5–8.5 months; 12 females) and twenty-five 10- to 11-month-olds ($M = 11.2$ months, range = 10.7–11.6 months; 14 females). The infants met the same inclusion criteria as in Experiment 1. All came from English-speaking homes. Five infants heard a second language for 1 to 3 hr per week ($n = 2$ at 7 to 8 months; $n = 3$ at 10 to 11 months). Nine additional infants were excluded from the analyses because of fussiness ($n = 5$), excessive movement ($n = 1$), or experimenter or equipment error ($n = 3$).

Stimuli. Infants listened to an artificial language with the same structure as the language described in Experiment 1. In Experiment 3, we used speech recordings from two of the original speakers from Experiment 1, namely, speakers 1 and 4. We selected the two voices because they sounded distinct, yet were representative of the range of the voices in the full set (Table 3). To test infants in conditions of moderate—but not extreme—variability, we did not use the most distinct speakers in terms of pitch, speaking rate, or perceived similarity to other voices. For example, the speakers used in Experiment 3 ranked second and seventh in overall similarity to the other voices. They also ranked fifth and sixth in pitch, and second and third in speaking rate (Table 2, Language A). As shown in Table 3, the speakers had comparable ratings for their similarity to the test voice. The speakers were both bilingual from childhood (Speaker 1 spoke Portuguese, Speaker 4 spoke Mandarin Chinese), but neither spoke English with a notable non-native accent. Both speakers scored near the midpoint on the 1 (*definitely a native speaker of English*) to 5 (*definitely not a native speaker*) scale; they ranked as sounding more native-like than Speaker 2, who spoke only English. Overall, the two voices provided a balanced window into the effects of low (vs. high) voice variability on statistical word segmentation. In subsequent research, it will be important to replicate this task using other combinations of voices to ensure that the findings generalize beyond the specific voice pair presented here.

As in Experiment 1, the voice changed after every 10 to 20 syllables (3 to 7 s), but in Experiment 3, the voices alternated between two speakers. The speaker changes occurred as described in Experiment 1. The test items were identical to Experiment 1 (i.e., female voice).

Procedure. The procedure was identical to Experiments 1 and 2. There were no differences in performance between the counterbalanced languages or between genders. Therefore, we collapsed across language and gender in subsequent analyses.

Results and Discussion

A 2 (Age Group: 7–8 months vs. 10–11 months; between subjects) × 2 (Test Item: words vs. part-words; within subjects) ×

2 (Test Block: 1 vs. 2; within subjects) mixed-design ANOVA of listening time was performed to examine learning. The main effect of age group was significant, $F(1, 48) = 4.29$, $p = .04$, $\eta_p^2 = .082$. The younger infants listened significantly longer ($M = 9.60$ s, $SD = 2.95$) than the older infants ($M = 7.94$, $SD = 2.71$). There was also a main effect of test block, $F(1, 48) = 53.26$, $p < .001$, $\eta_p^2 = .53$. Infants listened significantly longer overall during Block 1 ($M = 10.60$ s, $SD = 3.82$) than during Block 2 ($M = 7.00$ s, $SD = 3.23$). The main effect of test item was not significant, $F(1, 48) = 1.23$, $p = .27$, $\eta_p^2 = .025$. Infants did not differentiate the word and part-word test items. None of the interactions (Test Item × Age, Test Block × Age, Test Item × Test Block, and Test Item × Test Block × Age) was significant, all Fs < 1.

Figure 2 shows infants' listening times in Experiment 3. For consistency with Experiments 1 and 2, we have shown performance broken down by block, despite the nonsignificant block effects. The difference in listening times to words versus part-words was not significant in either block (ps > .4) or at either age (ps > .3). Infants' failure to differentiate the high transitional probability words and low probability part-words is consistent with the conclusion that the younger and older infants did not detect words in the language produced by two speakers rather than eight speakers. Note that we tested infants at ages at which they have shown vulnerability (around 7 to 8 months) and resilience (around 10 to 11 months) to voice changes during word recognition in prior research (e.g., Houston & Jusczyk, 2000; Singh et al., 2004), and we found no evidence of learning in either age group. It is always difficult to interpret the source of a null finding, and this experiment was not designed to uncover precisely what infants learned, if anything, in this condition. What is clear is that infants who heard only two speakers did not show clear evidence of learning the structure of the artificial language.

This conclusion is supported by comparing these results with those obtained in Experiments 1 and 2, in which infants *at both ages* successfully segmented the language when it was produced by eight different speakers. To directly test whether infants showed stronger evidence of learning from high speaker variation relative to low variation, we compared performance across Experiments 1 and 3. We performed a 2 (Age Group; between subjects) × 2 (Trial Type; within subjects) × 2 (Test Block; within subjects) × 2 (Experiment: 1 [8 speakers] vs. 3 [2 speakers]; between subjects) mixed-design ANOVA. The significant main effects were test block, $F(1, 96) = 133.28$, $p < .001$, $\eta_p^2 = .581$, revealing that infants listened longer during the first block than during the second block, and age group, $F(1, 96) = 4.79$, $p = .031$, $\eta_p^2 = .048$, revealing that younger infants listened longer than older infants. There was a significant interaction of Trial Type × Test Block, $F(1, 96) = 96.0$, $p = .023$, $\eta_p^2 = .053$, which was superseded by a significant three-way interaction of Trial Type × Test Block × Experiment, $F(1, 96), = 7.02$, $p = .009$, $\eta_p^2 = .068$. The remaining main effects and interactions were not significant (all ps > .06).

To interpret the three-way interaction, we considered the block effects observed previously and performed a two-way Trial type × Experiment ANOVA separately for Block 1. In Block 1, the main effects of trial type ($F(1, 98) = 3.60$, $p = .061$, $\eta_p^2 = .035$) and experiment were not significant ($F < 1$), but there was a significant interaction of Trial Type × Experiment, $F(1, 98) = 9.70$, $p = .002$, $\eta_p^2 = .090$. Thus, the difference in listening time to words versus part-words at test differed reliably depending on whether infants heard eight voices (Experiment 1) or two voices (Experiment 3) during segmentation. As shown previously, only infants who heard eight voices listened significantly longer to words than part-words (during Block 1), whereas infants who heard two voices did not.

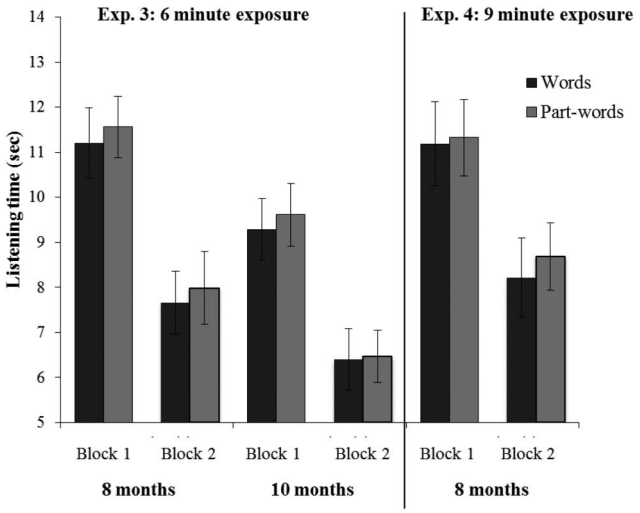

Figure 2. Mean listening time to word and part-word test trials for Experiments 3 and 4, separated by test block. Error bars indicate *SE*s.

Experiment 4

A consideration following from the results of Experiment 3 is that the 6-min exposure to two voices may have been insufficient to yield evidence of learning. Infants may have been learning and progressing toward a test item preference that would indicate successful segmentation of the speech stream. Thus, with additional exposure, infants could potentially demonstrate a consistent test item preference. We tested this possibility by extending the length of the artificial language exposure by 50% (to 9 min) for a group of 7- to 8-month-old infants. If infants were on the cusp of revealing evidence of learning, the additional exposure should move them toward reliable differentiation of the test items (either toward a familiarity or novelty preference; see Hunter & Ames, 1988). However, if infants again demonstrate no preference for the word or part-word test items (like the 7- to 8-month-olds and 10- to 11-month-old infants in Experiment 3), it would strengthen our proposal that infants do not segment speech as readily from two voices relative to eight voices.

Method

Participants. The participants were twenty-five 7- to 8-month-old infants ($M = 8.2$ months, range = 7.7–8.6 months; 13 females). The infants met the same inclusion criteria as in Experiment 1. All came from English-speaking homes. Four infants heard a second language for 1 to 4 hr per week. Four additional infants were excluded from analyses because of fussiness.

Stimuli. Infants listened to an artificial language with the same design as described in Experiment 3. The listening time was

extended by 50% by taking the first 3 min of the language and copying and splicing it to the end of the original 6-min sequence. The infants heard an additional 90 repetitions of the high-frequency words and an additional 45 repetitions of the low-frequency words (that served as test items in the test phase). The part-words used during the test phase also occurred 45 additional times relative to the exposure in Experiment 3. The test items were identical to the test items used in Experiments 1 and 3.

Procedure. The procedure was identical to the previous experiments. There were no differences in performance between the counterbalanced languages or between genders. Therefore, we collapsed across language and gender in subsequent analyses.

Results and Discussion

A 2 (Test Item: words vs. part-words) × 2 (Test Block: 1 vs. 2) within-subjects ANOVA of listening time was performed to examine learning. There was a main effect of test block, $F(1, 24) = 16.84$, $p < .001$, $\eta_p^2 = .41$. Infants listened significantly longer overall during Block 1 ($M = 11.26$ sec, $SD = 4.18$) than during Block 2 ($M = 8.45$ sec, $SD = 3.69$). The main effect of test item was not significant, $F < 1$. Infants did not differentiate the word and part-word test items. The interaction of Test Block × Test Item was not significant, $F < 1$.

Figure 2 shows infants' listening times separated by test block. The listening times to words versus part-words were not significant in either block ($ps > .5$). These findings replicate the null trial type effects observed in Experiment 3. Collectively, we have three demonstrations that infants fail to differentiate word and part-word test items after listening to an artificial language comprised of two voices: 8-month-olds and 10-month-olds with 6 min of language exposure, and 8-month-olds with 9 min of exposure. The replication across ages and across exposure durations supports the conclusion that infants do not demonstrate evidence of learning with low variation in voice. Importantly, we have also reported three demonstrations that infants *do* exhibit learning of the same artificial language when it is spoken in eight distinct voices: 8-month-olds and 10-month-olds in Experiment 1 (when tested with a novel female voice), and 8-month-olds in Experiment 2 (when tested with a novel male voice).

General Discussion

This research is the first to our knowledge to examine how voice variation affects infants' learning about the statistical structure of fluent speech. In Experiments 1 and 2, we found that infants successfully detected statistical regularities that marked word boundaries in a speech stream produced by eight different speakers. Seven- to 8-month-olds and 10- to 11-month-olds recognized statistically defined words produced by a novel female voice during testing. Moreover, 7- to 8-month-olds recognized the words produced by a novel male voice that was perceptually distinct from voices in the segmentation speech stream. The results of Experiments 1 and 2 suggest that infants' learning of the statistical regularities that determine word boundaries is resilient to surface form variation. However, when the same speech stream was presented by only two female voices in Experiments 3 and 4, infants failed to show evidence of successful word segmentation. Thus, the degree of variation in the speech stream affected infants' learning. Infants showed clear evidence of learning from the speech stream with high variation in voice, but no evidence of learning the same linguistic structure produced with low variation in voice. We propose that high variation across voices helps infants detect the invariant underlying structure of syllable patterns in fluent speech.

Processing variation in voice is not a trivial problem for young language learners. They are still determining which sound variants make meaningful distinctions between words and which do not (Stager & Werker, 1997). But as infants gain experience with speech, either across development (Houston & Jusczyk, 2000; Singh et al., 2004) or with increased exposure to frequent words (Singh, Nestor, & Bortfeld, 2008), the ability to attend to relevant phonemic variation during word recognition and listen through irrelevant voice information becomes more robust. The present research demonstrated a mechanism that is even more demanding than identifying individual words across surface changes. In Experiments 1 and 2, infants detected sound sequence regularities being presented with a range of individual voices that differed in pitch, speaking rate, and articulation patterns, among other features. Infants had to track and store relations between syllables across surface form variation and across time—all in an unfamiliar, pause-free speech context lacking prosodic word boundary cues. At test, infants then recognized statistically defined words when they occurred in a new voice that was either similar (Experiment 1) *or* dissimilar (Experiment 2) to the voices they heard during exposure to the speech stream.

Infants' successful learning from the speech stream with high variation in voice does not mean that learning was unaffected by variation. Even adults experience degradation in lexical processing amid speaker variation (Goldinger, 1998; Goldinger et al., 1991; Mullennix et al., 1989; Palmeri et al., 1993). Patterns in infants' direction of listening preferences can provide insight about the effects of high variation on infants' learning. A recent experiment used the same artificial language stimuli as the present experiments, except that the full language and the test items occurred in a consistent voice (Graf Estes, 2012). Ten-month-olds segmented the speech, showing a reliable preference for part-words. The novelty (i.e., part-word) preference is consistent with numerous prior statistical word segmentation experiments (Aslin et al., 1998; Johnson & Jusczyk, 2001; Saffran et al., 1996; Thiessen, Hill, & Saffran, 2005; Experiment 2). However, the novelty preference found in prior research contrasts with the familiarity (word) preference that the infants exhibited in Experiments 1 and 2. What accounts for this difference? One key factor affecting the direction of infants' listening preferences is the demands of the experimental task. According to models of infant attention (Houston-Price & Nakai, 2004; Hunter & Ames, 1988), when infants are presented with stimuli that they can process readily (e.g., simple stimuli) or when infants have learned thoroughly during familiarization, they typically attend longer to novel test items that differ from the familiarization stimuli. In contrast, when learning is still in progress, infants attend longer to familiar test stimuli (Hunter & Ames, 1988; Thiessen et al., 2005). Thus, while the present experiments do not offer a direct comparison (and the interpretation must be taken cautiously), the contrast in direction of preference from a single consistent voice (novelty) versus eight voices (familiarity) suggests that processing eight voices was more challenging than a

single voice. The successful learning that occurred from eight voices was not thorough enough to induce a novelty preference.

Our findings suggest that high speaker variation in Experiments 1 and 2 facilitated learning relative to the low variation in Experiment 3. Given that variation introduces complexity, how does high variation promote learning during statistical word segmentation? Broadly, variation may support learning by highlighting commonalities in critical relevant features and reducing attention to irrelevant features. Variation can help learners to detect features that are not critical for category membership (Apfelbaum, Hazeltine, & McMurray, 2012; Apfelbaum & McMurray, 2011; Perry et al., 2010). During statistical word segmentation, experience with voice variation in language input may allow infants to extract word structures because it highlights what is constant (syllable probability patterns) across surface-level changes (e.g., pitch, breathiness, and articulation differences). The changes in speaker identity that occurred every 3 to 7 seconds and across eight different voices may have (a) encouraged infants to disregard voice changes as uninformative, and (b) enhanced attention to higher-level syllable regularities. This interpretation is consistent with previous evidence that high variation is more effective than low variation in supporting learning across a variety of tasks and ages (Gómez, 2002; Houston, 1999; Lively et al., 1993; Needham et al., 2005; Rost & McMurray, 2009; Singh, 2008). Our findings provide the first evidence that high acoustic variation supports rather than hinders the tracking of syllable probability patterns, strengthening the notion that statistical word segmentation supports learning in natural language environments.

The present findings can be further understood within the context of exemplar models of memory. Of particular relevance is Thiessen and Pavlik's (2013) Integrative Minerva (iMinerva) model, designed to reveal how long-term memory shapes learning in a variety of linguistic tasks. This exemplar model provides a useful framework for interpreting the observed effects of high and low variation in voice on word segmentation. It is based on the idea that when learners encounter a novel exemplar of a syllable, it activates prior exemplars based on similarity. Under conditions of high variation during learning, infants encounter novel syllable exemplars that do not closely match in voice characteristics with other recently presented syllable exemplars because the voice changes frequently. However, the novel exemplars consistently match the phonemic characteristics of other recent syllable exemplars. During the model's interpretation process, the representation integrates the novel and prior exemplars, emphasizes their common characteristics (phonemic information), and deemphasizes the variable characteristics (voice). Through an abstraction process, the variable voice characteristics wash out of the representation and the syllable patterns are strengthened and retained. During testing, the words produced in the novel test voice match closely with the phonemic information of the previously segmented words; the novel voice information will not disrupt recognition because it is not an emphasized component of the stored representations.

iMinerva also helps us understand why low variation does not support learning effectively. According to the model, when infants hear two voices during learning, they frequently encounter syllable exemplars that match with recent syllable exemplars that occurred in the same voice. Therefore, consistent voice and phonemic information occur frequently and are both encoded in the representation; neither washes out during abstraction. During testing, the words presented in the novel test voice match in phonemic information, but not voice information. This produces a mismatch between the stored representations and the novel test exemplar, thereby inhibiting infants' abilities to recognize the words.

An additional and related consideration is that low levels of voice variation may push infants to focus on differences between voices, masking the underlying syllable-level regularities and promoting segregation of the speech streams by speaker identity. That is, infants may be led to attend to *speaker*-level categories, as opposed to *word*-level categories. This type of overly narrow category formation when processing low levels of variation has been shown in previous research (Singh, 2008; see also Gómez, 2002). Future experiments will be necessary to tease apart what information infants store about speakers and words when tracking probabilistic syllable patterns.

The results of this series of experiments are important for understanding the processes of statistical learning, and for addressing the potential criticism that behaviors in simple lab settings do not generalize to the demands of learning words in everyday language. Previous research has suggested that statistical word segmentation based on transitional probabilities may not function in the presence of varying word length (Johnson & Tyler, 2010), and it may be a relatively ineffective learning process compared with other types of cues (Endress & Hauser, 2010; Johnson & Jusczyk, 2001; Johnson, Seidl, & Tyler, 2014; Seidl & Johnson, 2006; Yang, 2004). Incorporating naturalistic challenges into word segmentation tasks is essential for understanding the ultimate bounds of statistical learning. Evidence that statistical word segmentation is robust despite irrelevant (but naturally occurring) surface form variation is important for establishing whether statistical learning is a viable contributor to natural language acquisition. Learners must listen through variation that is ubiquitous and salient—but not highly informative—for linguistic processing, such as variation in speakers' voices, affect, volume, and speaking rate. But critically, this information must not be ignored entirely, as it is important for interpreting natural communicative contexts. Speakers' voices cue infants to speaker identity (e.g., mother vs. father), affect conveys critical emotional content, and volume helps infants fall asleep or avoid danger. The present experiments demonstrate that even communicatively immature infants can effectively listen through high variation in speakers' voices to detect structurally relevant information in syllable transitional probability patterns. Because infants benefited from the presence of high variation, statistical learning mechanisms are meeting a key challenge present in natural languages: that multiple speakers of the ambient language contribute to infants' speech input. This finding supports the notion that statistical learning can scale up to meet naturalistic learning challenges, and in turn, supports accounts of language acquisition that propose a foundational role of distributional learning in language acquisition.

Conclusions

The ability to process speech in acoustic variation is essential for learning and processing language, and investigations of how infants overcome signal complexity are needed for evaluating the explanatory power of statistical learning. The experiments reported here indicate that infants successfully perform statistical word

segmentation when many speakers contribute to the speech signal. These findings support the notion that statistical learning can scale up to resolve a salient source of complexity inherent in infants' natural language environments.

References

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2012). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology*. Advance online publication.

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science, 35,* 1105–1138. http://dx.doi.org/10.1111/j.1551-6709.2011.01181.x

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324. http://dx.doi.org/10.1111/1467-9280.00063

Casasola, M. (2005). When less is more: How infants learn to form an abstract categorical representation of support. *Child Development, 76,* 279–290. http://dx.doi.org/10.1111/j.1467-8624.2005.00844.x

Cohen, L. B., Atkinson, D. J., & Chaput, H. J. (2004). *Habit X: A new program for obtaining and organizing data in infant perception and cognition studies* (Version 1.0). Austin, TX: University of Texas.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology, 61,* 177–199. http://dx.doi.org/10.1016/j.cogpsych.2010.05.001

Finley, S. (2013). Generalization to unfamiliar talkers in artificial language learning. *Psychonomic Bulletin & Review, 20,* 780–789. http://dx.doi.org/10.3758/s13423-013-0402-7

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 15822–15826. http://dx.doi.org/10.1073/pnas.232472899

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. *PLoS ONE, 8,* e52500. http://dx.doi.org/10.1371/journal.pone.0052500

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32,* 249–268. http://dx.doi.org/10.1017/S0305000904006786

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279. http://dx.doi.org/10.1037/0033-295X.105.2.251

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152–162. http://dx.doi.org/10.1037/0278-7393.17.1.152

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431–436. http://dx.doi.org/10.1111/1467-9280.00476

Graf Estes, K. (2012). Infants generalize representations of statistically segmented words. *Frontiers in Psychology, 3,* 447. http://dx.doi.org/10.3389/fpsyg.2012.00447

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18,* 254–260. http://dx.doi.org/10.1111/j.1467-9280.2007.01885.x

Harris, Z. S. (1955). From phoneme to morpheme. *Language, 31,* 190–222. http://dx.doi.org/10.2307/411036

Houston, D. M. (1999). *The role of talker-specific information in word segmentation by infants* (Unpublished doctoral dissertation). The 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,* 1570–1582. http://dx.doi.org/10.1037/0096-1523.26.5.1570

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 1143–1154. http://dx.doi.org/10.1037/0096-1523.29.6.1143

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infants' preference procedures. *Infant and Child Development, 13,* 341–348. http://dx.doi.org/10.1002/icd.364

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5,* 69–95.

Hunter, M. A., Ames, E. W., & Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology, 19,* 338–352. http://dx.doi.org/10.1037/0012-1649.19.3.338

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44,* 548–567. http://dx.doi.org/10.1006/jmla.2000.2755

Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds. *PLoS ONE, 9,* e83546. http://dx.doi.org/10.1371/journal.pone.0083546

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science, 13,* 339–345. http://dx.doi.org/10.1111/j.1467-7687.2009.00886.x

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* B35–B42. http://dx.doi.org/10.1016/S0010-0277(02)00004-5

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental Science, 14,* 1207–1219. http://dx.doi.org/10.1111/j.1467-7687.2011.01073.x

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science, 14,* 1323–1329. http://dx.doi.org/10.1111/j.1467-7687.2011.01079.x

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94,* 1242–1255. http://dx.doi.org/10.1121/1.408177

Maguire, M. J., Hirsh-Pasek, K., Golinkoff, R. M., & Brandone, A. C. (2008). Focusing on the relation: Fewer exemplars facilitate children's initial verb learning and extension. *Developmental Science, 11,* 628–634. http://dx.doi.org/10.1111/j.1467-7687.2008.00707.x

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111. http://dx.doi.org/10.1016/S0010-0277(01)00157-3

Mersad, K., & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development, 8,* 303–315. http://dx.doi.org/10.1080/15475441.2011.609106

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91–117. http://dx.doi.org/10.1016/S0010-0277(03)00140-9

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

Needham, A., Dueker, G., & Lockhead, G. (2005). Infants' formation and use of categories to segregate objects. *Cognition, 94,* 215–240. http://dx.doi.org/10.1016/j.cognition.2004.02.002

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 19, 309–328. http://dx.doi.org/10.1037/0278-7393.19.2.309

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80,* 674–685. http://dx.doi.org/10.1111/j.1467-8624.2009.01290.x

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science, 21,* 1894–1902. http://dx.doi.org/10.1177/0956797610389189

Quinn, P. C., & Bhatt, R. S. (2005). Learning perceptual organization in infancy. *Psychological Science, 16,* 511–515. http://dx.doi.org/10.1111/j.0956-7976.2005.01567.x

Roseberry, S., Richie, R., Hirsh-Pasek, K., Golinkoff, R. M., & Shipley, T. F. (2011). Babies catch a break: 7- to 9-month-olds track statistical probabilities in continuous dynamic events. *Psychological Science, 22,* 1422–1424. http://dx.doi.org/10.1177/0956797611422074

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

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12,* 110–114. http://dx.doi.org/10.1111/1467-8721.01243

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. http://dx.doi.org/10.1126/science.274.5294.1926

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4,* 273–284. http://dx.doi.org/10.1207/S15327078IN0402_07

Sahni, S. D., Seidenberg, M. S., & Saffran, J. R. (2010). Connecting cues: Overlapping regularities support cue discovery in infancy. *Child Development, 81,* 727–736. http://dx.doi.org/10.1111/j.1467-8624.2010.01430.x

Schmale, R., Cristia, A., Seidl, A., & Johnson, E. K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy, 15,* 650–662. http://dx.doi.org/10.1111/j.1532-7078.2010.00032.x

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibility of early word representations. *Developmental Science, 12,* 583–601. http://dx.doi.org/10.1111/j.1467-7687.2009.00809.x

Seidl, A., Cristia, A., & Onishi, K. H. (2014). Talker variation aids infants' phonotactic learning. *Language Learning and Development, 10,* 1–11. http://dx.doi.org/10.1080/15475441.2013.858575

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science, 9,* 565–573. http://dx.doi.org/10.1111/j.1467-7687.2006.00534.x

Shi, R., & Werker, J. F. (2001). Six-month-old infants' preference for lexical words. *Psychological Science, 12,* 70–75. http://dx.doi.org/10.1111/1467-9280.00312

Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy, 10,* 187–198. http://dx.doi.org/10.1207/s15327078in1002_5

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106,* 833–870. http://dx.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,* 173–189. http://dx.doi.org/10.1016/j.jml.2004.04.004

Singh, L., Nestor, S. S., & Bortfeld, H. (2008). Overcoming the effects of variation in infant speech segmentation: Influences of word familiarity. *Infancy, 13,* 57–74. http://dx.doi.org/10.1080/15250000701779386

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development, 4,* 157–178. http://dx.doi.org/10.1080/15475440801922131

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388,* 381–382. http://dx.doi.org/10.1038/41102

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50,* 86–132. http://dx.doi.org/10.1016/j.cogpsych.2004.06.001

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7,* 53–71. http://dx.doi.org/10.1207/s15327078in0701_5

Thiessen, E. D., & Pavlik, P. I., Jr. (2013). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science, 37,* 310–343. http://dx.doi.org/10.1111/cogs.12011

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39,* 706–716. http://dx.doi.org/10.1037/0012-1649.39.4.706

Vouloumanos, A., Brosseau-Liard, P. E., Balaban, E., & Hager, A. D. (2012). Are the products of statistical learning abstract or stimulus-specific? *Frontiers in Psychology, 3,* 70. http://dx.doi.org/10.3389/fpsyg.2012.00070

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3,* 1–30. http://dx.doi.org/10.1207/S15327078IN0301_1

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences, 8,* 451–456. http://dx.doi.org/10.1016/j.tics.2004.08.006

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy, 15,* 420–433. http://dx.doi.org/10.1111/j.1532-7078.2009.00024.x

Received April 18, 2014
Revision received May 26, 2015
Accepted July 6, 2015 ■