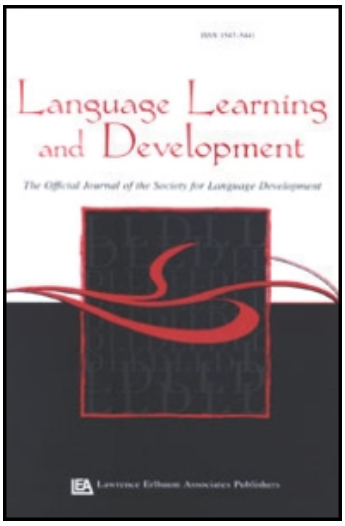

Language Learning and Development
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653671

Speech Segmentation in a Simulated Bilingual Environment: A Challenge for Statistical Learning?

Daniel J. Weiss [a]; Chip Gerfen [b]; Aaron D. Mitchel [a]
[a] Department of Psychology, Pennsylvania State University, [b] Department of Spanish and Program in Linguistics, Pennsylvania State University,

Online Publication Date: 01 January 2009

To cite this Article Weiss, Daniel J., Gerfen, Chip and Mitchel, Aaron D.(2009)'Speech Segmentation in a Simulated Bilingual Environment: A Challenge for Statistical Learning?',Language Learning and Development,5:1,30 — 49
To link to this Article: DOI: 10.1080/15475440802340101
URL: http://dx.doi.org/10.1080/15475440802340101

Language Learning and Development, 5: 30–49, 2009

ISSN: 1547-5441 print / 1547-3341 online
DOI: 10.1080/15475440802340101

Psychology Press
Taylor & Francis Group

Speech Segmentation in a Simulated Bilingual Environment: A Challenge for Statistical Learning?

Daniel J. Weiss
Department of Psychology, Pennsylvania State University

Chip Gerfen
Department of Spanish and Program in Linguistics, Pennsylvania State University

Aaron D. Mitchel
Department of Psychology, Pennsylvania State University

Studies using artificial language streams indicate that infants and adults can use statistics to correctly segment words. However, most studies have utilized only a single input language. Given the prevalence of bilingualism, how is multiple language input segmented? One particular problem may occur if learners combine input across languages: The statistics of particular units that overlap different languages may subsequently change and disrupt correct segmentation. Our study addresses this issue by employing artificial language streams to simulate the earliest stages of segmentation in adult L2-learners. In four experiments, participants tracked multiple sets of statistics for two artificial languages. Our results demonstrate that adult learners can track two sets of statistics simultaneously, suggesting that they form multiple representations when confronted with bilingual input. This work, along with planned infant experiments, informs a central issue in bilingualism research, namely, determining at what point listeners can form multiple representations when exposed to multiple languages.

A growing body of research indicates that statistical learning plays a central role in early language acquisition (e.g., Saffran et al., 1996a; Maye, Werker, & Gerken, 2002; Thiessen & Saffran, 2003). Of particular significance has been the application of statistical learning to the long-standing problem of speech segmentation. Unlike orthography, discontinuities in the speech stream do not reliably denote word boundaries (Cole & Jakimik, 1980); at the same time, there are no invariant, cross-linguistic acoustic cues to word boundaries (Miller & Eimas, 1995). Thus, the discovery that language learners of all ages can use statistical information (in the form

Correspondence should be addressed to Daniel J. Weiss, 643 Moore Building, University Park, PA 16802. E-mail: djw21@psu.edu

of transitional probabilities) to segment artificial speech streams affords critical insight as to how this problem may be solved (see below). To date, however, research in this area has focused on segmentation tasks involving the use of a single input language, an approach that models the segmentation problem in a monolingual setting. Given that more than half of the world's population learns to speak more than one language (Crystal, 1997), it is also crucial to consider how learners fare with input from multiple languages. In this paper, we begin to address this issue by extending the statistical learning paradigm to two input streams.

The logic of recent statistical learning segmentation studies is that sounds occurring together with high probability are more likely to represent words, whereas sounds co-occurring with low probability signal word boundaries. In the English stream *pretty baby*, for example, the syllable "pre" is followed by a restricted set of syllables, yielding a situation in which "pre" is followed by "ty" with high probability (close to 80% in speech to infants). By contrast, the "ty" in *pretty* is word-final and can be followed by almost any English syllable such that the probability of hearing "ba" following "ty" is thus very low (under 1% in speech to infants; see Saffran, 2003).

Although this well-known example (e.g., Saffran et al., 1996a,b; Johnson & Jusczyk, 2001; Saffran, 2003) neatly illustrates the central concept of statistical learning, the implicit assumption is that the statistics are being derived from only a single language (in this case English). However, one could imagine that input from two languages could present additional complications. For example, suppose a learner of both Hebrew and English encounters the syllable "zeh." In Hebrew, "zeh" often corresponds to a word meaning "this" and thus likely signals a word boundary (since "zeh" can be followed by many different words). However, in English "zeh" occurs within words (e.g., *zealous* and *zealot*). The challenge for learners is to realize that sounds in one language may pattern differently than sounds in a second language. If learners combine statistical information across languages, they will increase the noise in the statistical patterns of each, thereby increasing the risk of failure to segment either language correctly (or minimally, delaying success). The goal of our experiments is to use the logic from the statistical learning studies of word segmentation in order to determine whether learners are capable of forming multiple representations when confronted with two artificial speech streams (i.e., of performing separate computations for each stream), and in doing so, circumvent the statistical problem detailed above.

When learners encounter multiple languages, there are many potential cues that could facilitate discrimination, a prerequisite step for allowing them to perform separate computations. Some cues that learners might use come from the languages themselves. For example, stress cues, phonotactic patterns, as well as differences in allophonic and microphonetic detail can signal exposure to multiple languages. However, the challenge for the learner is converging on the right set of cues, as distinct language pairs will make use of different cues to varying degrees. For example, learners acquiring Spanish and Catalan may experience a greater degree of phonological overlap than those acquiring English and Spanish. Other cues to multiple language exposure may come from the context under which learners are exposed to the languages. For example, one parent may always produce a particular language, whereas a second parent may produce another. A crucial issue confronting the learner is to determine which of the potentially available cues, both contextual and within the language, can be used indexically to distinguish between the input languages. The central goal of these experiments is to develop a paradigm that allows us to test whether learners develop multiple representations when provided with an adequate cue. Future experiments will then focus on systematically

determining the types of cues that learners may be able to use indexically (e.g., Mitchel & Weiss, in review).

Research with newborns and young infants attests to the consequences of trying to discriminate language pairs that vary in their degree of rhythmic similarity. Mehler and colleagues (1988) found that newborn infants are capable of discriminating nonnative language pairs that differed in rhythmic class. However, languages from the same rhythmic class were not discriminated (Nazzi, Bertoncini, & Mehler, 1998). By two months of age, infants no longer discriminate languages from different rhythmic classes unless one of the languages in the testing situation is the native language (Christophe & Morton, 1998). Interestingly, infants raised in a bilingual environment are capable of discriminating languages from the same rhythmic class as early as 4 months of age (Bosch & Sebastián-Gallés, 2001). Together, this evidence suggests that at an early stage of acquisition, infant language learners are forming representations about the features inherent in their native language (or languages) and are capable of using those representations to discriminate between the languages (at least in the context of the experimental stimuli used in the aforementioned studies). These abilities appear to rely, at least in part, on general features of the mammalian auditory system, as evidenced by similar rhythmic class discriminations being available to rats (Toro, Trobalon, & Sebastián-Gallés, 2003) and cotton-top tamarin monkeys (Ramus et al., 2000; Tincoff et al., 2005). Clearly, in the case of humans, this system is subsequently shaped by exposure early in development in order to facilitate the acquisition of a native language.

The underlying logic of this body of research asserts that language acquisition requires learners to extract regularities among utterances, and it follows that separating languages is prerequisite for successful language acquisition (see Mehler et al., 1988; Mehler & Christophe, 1995). If this logic holds, then learners acquiring multiple languages must be able to first discriminate among the individual languages, and subsequently perform separate computations across each language as a means of extracting the correct regularities for each individual language (see Bosch & Sebastián-Gallés, 2001).

With this in mind, the goal for the present work is to adopt a reductionist approach in order to test for whether learners do, in fact, form multiple representations when provided with an indexical cue. We chose speaker voice as the indexical cue, because voice cues are highly salient and may cue speaker-specific representations (see Eisner & McQueen, 2005; Kraljic & Samuel, 2006). More generally, this experiment constitutes the first stage of a larger project, the goal of which is to determine the conditions under which learners form multiple representations and to identify the types of cues that facilitate the process.

The outline of this study is as follows: In Experiment 1, we test whether learners can correctly segment one of three four-word artificial speech streams (produced by two different speakers) following 12 minutes of exposure in order to provide a baseline for the subsequent experimental manipulations. In Experiment 2, we test the effect of interleaving two of the streams from Experiment 1. In Condition 1, the streams, each produced by a different speaker (one male and one female), have compatible statistics (see below) and thus are predicted to be learnable—that is, successfully segmented even if learners combine information across streams. In Condition 2, the setup is identical except that the statistical properties of the streams are incongruent. By incongruent, we mean that learners should successfully segment both streams if they encapsulate the statistics of each and perform independent computations, as discussed below. Experiment 3 is identical in design to Experiment 2 except that the indexical cue of speaker voice is removed (i.e., the streams are all presented by the same speaker). The purpose

of this manipulation is to ensure that the results observed in Experiment 2 are only possible when an indexical cue is present to induce learners to encapsulate the statistical properties of each stream (i.e., without an indexical cue, the statistical structure of the incongruent streams should preclude correct segmentation of both of the interleaved languages). We conclude with Experiment 4, in which we interleave two congruent languages that contain a comparable number of overlapping syllables as the incongruent language pairs from Experiments 2 and 3.

EXPERIMENT 1—CONGRUENT AND INCONGRUENT LANGUAGES IN ISOLATION

Experiment 1 tested whether adult learners could successfully parse three artificial speech streams. Each stream consisted of four words with a similar underlying statistical structure. For clarity of discussion, we label each stream as an "artificial language" (hence L1, L2, L3) because in previous monolingual statistical learning research, these types of speech streams have been used to represent individual languages for the purposes of parsing (e.g., Saffran et al., 1996a,b; Newport & Aslin, 2004; Thiessen & Saffran, 2003; see Saffran, 2003 for review.). Our streams are created from the same sound inventory, using the same statistical structures. The indexical cue of voice (either male or female), along with combined statistical properties (see below), differ between the streams. Each subject heard one of the three languages for 12 minutes and was subsequently tested on what she or he had learned (see below). The purpose of Experiment 1 was to establish that all three languages were learnable at above chance levels when presented in isolation. One language (hereafter L1) was spoken in a male voice. The remaining languages (L2 and L3) were spoken in a female voice. Manipulations in Experiment 2 pair the languages to investigate whether cross-language statistical interactions affect learning.

Methods

Participants

39 monolingual English-speaking undergraduates (ages 18–23) were included in this study from the Introduction to Psychology subject pool and received class credit for participation. Criteria for bilingual classification included anyone with 5 or more years studying a second language or who self-classified as being bilingual. Due to the variability of the subject pool and the need for attention in statistical learning tasks (Toro et al., 2005), monolingual participants who marked a 6 or less on a scale of 10 rating their effort in the experiment were not included in the analysis (3). We excluded participants due to experimenter error (1), difficulty in matching the test trials with the correct numbers on the scoring sheet (4), or falling asleep (2), or failure to follow instructions (4; 14 total exclusions). None of the participants had previous experience with statistical learning experiments.

Stimuli

Each language consisted of 4 artificial trisyllabic (CV.CV.CV) words. One of the languages was spoken in a male voice, whereas the remaining two were spoken in a female voice. The CV

syllables comprising the words were created as follows. Two voices (one male and one female) were digitally recorded while producing multiple tokens of CVC syllables (e.g., [bab], [bad], [bag]). The CV sequences were recorded with coda consonants in order create labial, alveolar, palatoalveolar, and velar VC transitions. CV syllables were also recorded with no coda consonant in order to provide natural-sounding word and foil word endings for the trisyllabic strings presented in isolation in the testing phase. The tokens selected from the original recordings for stimuli creation were hand edited in Praat (Boersma, 2001) to control for duration and to remove the final consonant, leaving only the VC transition, thus allowing each CV to concatenate smoothly with a following CV sequence in the speech stream. For the languages used in the experiments here, the average word duration in the male voice was 800 ms (*SD* = .8 ms) per trisyllabic word, and was 805 ms (*SD* = .7 ms) per trisyllabic word in the female voice. Tokens were normalized in SoundForge to control for salient loudness differences and resynthesized in Praat. In the resynthesized tokens, for each speaker, a single, identical f0 contour was placed on each CV. To enhance the naturalness of the concatenated sequences of syllables, the intonation contour for each voice was modeled on the natural intonation contours of the medial syllable in CV.CV.CV strings produced by the respective speakers. In this way, possible pitch cues to segmentation were removed from the speech stream while retaining a more natural-sounding string.

The four words were concatenated into a continuous stream consisting of random orderings of the four words. Each word was presented an equal number of times, and no words were repeated twice in a row. There were no pauses in between the syllables and no other acoustic word boundary cues to signal boundaries. Because recent research has shown that both syllable-based and segment transitional probabilities (defined as consonant to vowel to consonant, etc.) may play a crucial role for speech segmentation (Newport, Weiss, Wonnacott, & Aslin, 2004), the design of the experiments presented here takes both types of statistics into account. Figures 1a and 1b show the design details of the languages. Within each word, syllable-to-syllable transitional probabilities were perfect (1.0) and dipped to .33 at word boundaries (because no words were repeated, any word could be followed by one of three other words). The segment transitional probabilities were .5 within each word and .33 at the word boundary. This pattern of statistics was stable across all four words.

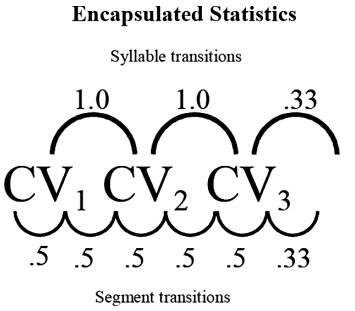

FIGURE 1a This figure shows the values of the transitional probabilities between adjacent syllables and adjacent segments for all three speech streams used in Experiment 1.

Language 1		Language 2		Language 3		Language 4	
Words	**Partwords**	**Words**	**Partwords**	**Words**	**Partwords**	**Words**	**Partwords**
bə tɪ gu	sæ to gʊ	po fɪ ku	dʒa dɪ dʒo	gu pæ tə	bo dʒi ga	po fɪ gu	tʃa dɪ dʒo
sɪ tʃə vi	gu vʊ bo	dɪ dʒo zi	ku zə pʊ	dʒi ga pʊ	sɪ gu pæ	dɪ dʒo zi	gu zə pʊ
vʊ bo sæ	tʃa sɪ tʃə	zə pʊ dæ	zi po fɪ	sæ dʒu bo	pʊ ta bi	zə pʊ dæ	zi po fɪ
to gʊ tʃa	vi bə tɪ	fʊ kə dʒa	dæ zə pʊ	ta bi sɪ	tə sæ dʒu	fʊ kə tʃa	dæ zə pʊ

FIGURE 1b This figure shows the design of languages L1, L2, L3, and L4 and the partwords used in the test.

Procedure

Participants were instructed to listen to a speech stream and told that there would be a test following the session. They were given no explicit directions about the task and had no information about the length or structure of the words or about how many words each language contained. Stimuli were presented in 4-minute blocks. Between blocks, there was a 30-second break. After 12 total minutes of listening (with a total of 864 words presented), participants received a two-alternative forced-choice test. Each test trial consisted of 2 trisyllabic strings, separated by one second of silence, with a 4-second pause in between test trials to allow participants to respond. On each trial, one string was a statistically defined word from one of the nonsense languages, whereas the other was a partword (consisting of the last syllable from one word concatenated with the first two syllables of a different word) in counterbalanced order. Tests always compared a word from a particular language with a partword from the same language (spoken in the same voice). Participants indicated which of the two strings sounded more like a word from the languages by circling either the "1" or the "2" on the answer sheet. The test was comprised of the four words, along with 4 partword foils. Each word was paired with two of the partwords twice (counterbalancing for order) yielding 16 total test trials across both languages. Following the test, participants filled out a questionnaire on language background (how many languages, number of years studied, and whether they would label themselves as being bilingual), on how hard the participants tried in the task, and on whether or not they became confused during test (being unable to track the trial number correctly since it was not explicitly mentioned before the word pairs were presented).

Results and Discussion

Performance on each individual language exceeded chance. Participants learning the L1 male voice averaged 10.6 (SD = 2.7) out of 16 (66%), $t(12) = 3.5$, $p = .001$ two-tailed, $d = 2.02$. The L2 female voice averaged 11.7 (SD = 2.1) out of 16 (73%), $t(12) = 6.5$, $p < .001$, $d = 3.75$. The L3 female voice averaged 9.9 (SD = 2.3) out 16 (62%), $t(12) = 3.02$, $p = .01$, $d = 1.74$. Overall, an analysis of variance using language as a factor found no significant differences across any of the conditions, $F(2,36) = 2.01$, $p = .14$, $\eta^2 = 0.10$.

The results of this experiment demonstrate that all three of the artificial languages were learnable in isolation and that learning across all languages did not differ significantly. The next experiment investigates whether these languages remain learnable when they are interleaved in pairs that share either compatible statistics or incompatible statistics.

EXPERIMENT 2—MIXED PAIR PRESENTATIONS

In Experiment 1, learners successfully parsed three miniature artificial languages. In Experiment 2, pairs of these languages were interleaved in 2-minute intervals (yielding 24 total minutes of exposure across two conditions). In Condition 1, the interleaved languages (L1 and L2) are statistically compatible such that the word boundaries remain stable regardless of whether the statistics are combined across both languages. In Condition 2, two of the interleaved languages (L1 and L3) have incongruent statistics. As a consequence, if participants attempt to combine statistics across this pairing of languages, the statistics that indicate word boundary information in each language will be combined, thereby obscuring the word boundaries (see below). The predictions are as follows: If participants can track and maintain separate sets of statistics when given an indexical cue of voice, then they should successfully learn both languages in both Conditions 1 and 2. By contrast, if participants attempt to combine statistics across the languages (i.e., by treating both streams as a single language), then only Condition 1 should facilitate successful learning of both input streams. This is because only Condition 1 preserves word boundary information unambiguously across the combined languages.

Methods

Participants

Thirty monolingual undergraduates were recruited from the Introduction to Psychology subject pool with the same conditions as noted above. Exclusion criteria were identical to those used in Experiment 1 above (18 total exclusions; monolingual participants who marked a 6 or less on a scale of 10 rating their effort in the experiment [6]; experimenter error [1]; difficulty in matching the test trials with the correct numbers on the scoring sheet [5]; those falling asleep [1]; or failure to follow instructions [5]).

Stimulus Materials

The stimulus languages (L1, L2, L3) for this experiment were identical to those used in Experiment 1. Crucially, however, by interleaving languages, we were able to manipulate their statistical properties, depending on the parsing strategy of the participants. If participants encapsulated each language (i.e., parsed each language independently, despite the interleaving during presentation), then the statistical properties of each language were identical to those reported in Experiment 1. However, if participants combined information across language pairs, then transitional probabilities varied as a function of the statistical interaction between language pairs. In Condition 1 (interleaving statistically compatible languages, L1 and L2), the syllable to syllable transitional probabilities remained the same as the encapsulated statistics (1.0 inside the word and .33 at word boundaries), and the segment to segment statistical properties bounced between .5 and .25 within word, dipping

to .18 at the word boundaries. This pattern of statistics was stable across all four words (see Fig. 2). In Condition 2 (interleaving statistically incompatible languages, L1 and L3), the statistical noise floor was raised. The syllable-to-syllable transitional probabilities now ranged between .5 and 1.0 within word but still dipped to between .17 and .33 at word boundaries. The segment-to-segment transitional probabilities ranged between .25 and 1.0 within word, with a dip to .17 at word boundaries. The reason for the range in these transitional probabilities is due to the overlap of some syllables and segments across streams, whereas others appeared only within one stream. Importantly, these statistics were also not stable between words (e.g., some words had a .5 syllable transitional probability occurring between syllables 1 and 2, whereas other words had that dip between syllables 2 and 3; see Figure 3). Although slight dips remain at the word boundaries, our prediction was that the instability of the statistics within words and the increased statistical variability overall should render these boundaries highly difficult to detect if listeners are combining the interleaved strings.

Procedure

As noted above, we presented the exact same languages as those presented in Experiment 1. However, rather than presenting languages in isolation for 12 minutes, we presented the two

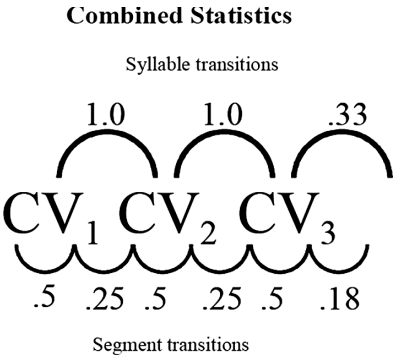

FIGURE 2 This figure shows the combined statistics of L1 and L2.

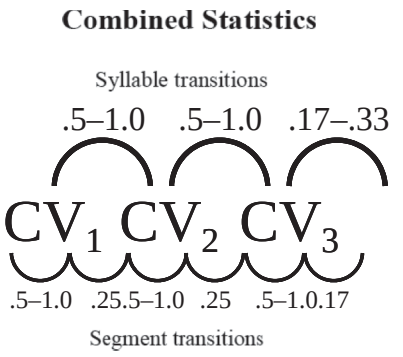

FIGURE 3 This figure shows the combined statistics of L1 and L3.

languages interleaved in three blocks of 8 minutes each. Within each block, the two languages were presented in alternating 2-minute strings (yielding 4 minutes total for each language within a block). Between blocks, participants were given a 1-minute break. After 24 total minutes of listening, participants completed a two-alternative forced-choice test. Each test trial consisted of the same test stimuli as used in Experiment 1. In this case, however, participants responded to 32 total trials rather than 16 in order to accommodate the presentation of two languages instead of one (in Experiment 1). Test trials for each language were presented in the same voice heard during familiarization. Thus, if a word (or partword) was produced in a male voice during familiarization, the same male voice was used during the test phase. The test was created using pseudo-randomized ordering with constraints such that no words or partwords occurred in adjacent trials. Essentially, there were two blocks of 16 trials (unbeknownst to the participants). All word-partword pairings occurred in both blocks and were counterbalanced such that each pairing occurred twice (balancing for order). Tests always compared a word from a particular language with a partword from the same language (spoken in the same voice). The rest of the procedure was identical to Experiment 1. Participants were randomly assigned to Condition 1 (compatible statistics) or Condition 2 (incongruent statistics).

Results and Discussion

Participants in each condition successfully learned above chance. In Condition 1, participants averaged 21.9 (4.48) out of 32 (68%) correct ($t(14) = 5.13$, $p = .001$, $d = 2.74$; see Fig. 4), with an average score of 11.2 (3.34) out of 16 (70%) correct in L1, $t(14) = 3.71$, $p = .002$, $d = 1.98$, and an average score of 10.7 (2.69) in L2 (67%), $t(14) = 3.94$, $p = .001$, $d = 2.11$. In Condition

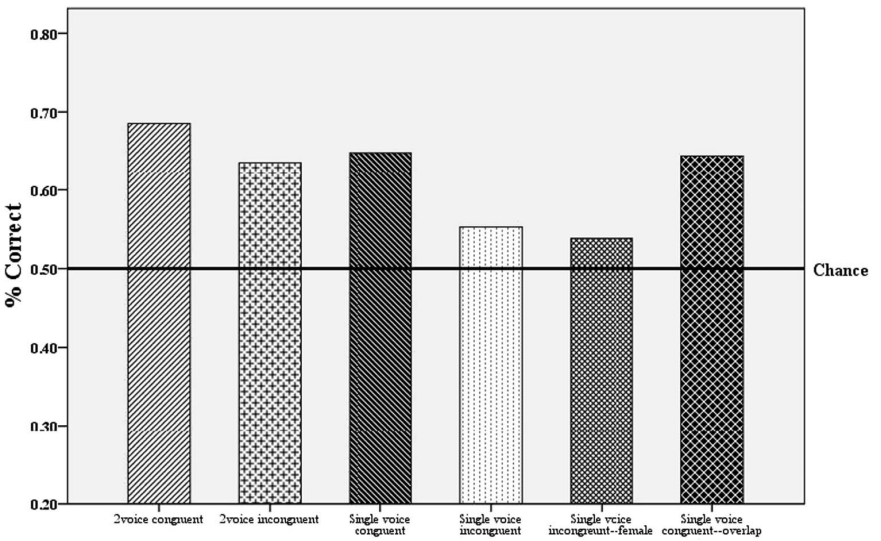

FIGURE 4 This figure shows the results of Experiments 2, 3, and 4. These scores reflect performance across both interleaved streams on a 32-item two-alternative forced choice task.

2, participants averaged 20.3 (2.92)out of 32 (63%) correct, $t(14) = 5.75$, $p = .001$, $d = 3.07$, with an average score of 10.4 (2.23) out of 16 (65%) correct in L1, $t(14) = 4.17$, $p = .001$, $d = 2.23$, and an average score of 9.9 (2.74) in L3 (62%), $t(14) = 2.74$, $p = .016$, $d = 1.46$. The overall scores in these conditions did not significantly differ from each other, $t(28) = 1.16$, $p = .25$. A repeated measures ANOVA with a between-subjects factor of condition revealed no differences in performance across individual languages, $F(1,28) = .40$, $p = .535$, $\eta^2 = .01$, nor were there any significant between subject differences across condition, $F(1, 28) = 1.34$, $p = .256$, $\eta^2 = .05$, the interaction of these two variables was also nonsignificant, $F(1, 28) = .00$, $p = 1.00$, $\eta^2 = .00$. Broken down by individual languages, there were no significant differences when languages were presented in isolation (Experiment 1) or interleaved as in Experiment 2—L1: $F(2,40) = .33$, $p = .723$, $\eta^2 = .02$; L2: $t(26) = 1.08$, $p = .29$, $d = .42$; L3: $t(26) = -.01$, $p = .92$, $d < .01$.

Experiment 2 demonstrates that learners can successfully parse these artificial languages both when they are presented in mixed fashion, interleaved over 24 minutes, or in isolated 12-minute blocks (as in Experiment 1). These results establish that learners are quite adept at the parsing task in both conditions. Given the compatible structure of the languages presented in Condition 1, however, we cannot conclude how learners were able to parse at above chance levels. As noted above, learners can parse successfully either by encapsulating the statistics separately for each of the two streams (i.e., forming multiple representations) or by combining statistics into a unified representation. Crucially, however, learners in Condition 2 should have had difficulty in correctly parsing both L1 and L3 if they attempted to combine statistics across the two languages. Because they did not exhibit a deficit in performance, we reason that, at least in this condition, learners maintained separate sets of statistics (i.e., that they formed multiple representations). This conclusion may be confounded, however, by the degree to which we manipulated the statistical properties of the combined languages in the incongruent Condition 2. Therefore, we designed Experiment 3 as a means of testing for whether the incongruent languages both remained learnable in the absence of the voice cue present in Experiment 2.

EXPERIMENT 3A—SINGLE VOICE CONDITIONS

Two fundamental features characterized Experiment 2: (a) language streams whose combination resulted in either compatible or incongruent statistics and (b) an indexical voice cue which may have facilitated the encapsulation of each stream, resulting in successful learning irrespective of condition. There are two possible alternative accounts that must be ruled out before we can conclude that learners encapsulated the statistics of each of the interleaved languages. One alternative hypothesis is that the combined statistics of the streams in Condition 2 were not sufficiently degraded so as to preclude successful parsing. A related alternative hypothesis is that the structure of the presentation (interleaving 2-minute intervals of consistent statistical information) sufficed to facilitate learning, rendering the issue of incongruent statistics moot. For example, because the blocks alternated within a session, learners who detected a difference between the blocks could have capitalized on the alternating blocks to further segregate the languages. The purpose of Experiment 3 is to test the conclusions we have drawn from Experiment 2 by replicating that experiment with the indexical cue of voice removed. This manipulation allows us to test both alternative accounts against our encapsulation hypothesis. If our hypothesis is correct, then learners should be capable of parsing both

streams *only* when the interleaved streams have compatible statistics. Without an indexical cue to prompt encapsulation, the incompatible statistical structures used in Experiment 2 Condition 2 should preclude correct parsing. However, if correct parsing persists in the incongruent statistical condition, then we cannot rule out that either the statistical structure of the combined streams or the structure of presentation was responsible for the findings reported in Experiment 2.

Methods

Participants

Thirty-one monolingual undergraduates were included from the Introduction to Psychology subject pool with the conditions noted above. In addition, we excluded individuals who fell below the effort criteria (7), those who felt confused on the test (5), or did not follow instructions (5), or slept during the experiment (1).

Stimuli and Procedures

The stimuli were structurally identical to those used in Experiment 2. In Experiment 3, however, both languages were presented by a single voice rather than by the male and female voices used in the experiments above. In this experiment, in both conditions, both streams were produced by the male speaker alone. As in the other experiments, participants listened to 24 minutes of the speech streams and then answered a 32-item test. In the single-voice incongruent condition, the order of language presentation was counterbalanced to explore whether learners would acquire the initial stream, but not the subsequent incompatible stream (i.e., to test for order effects). The results of these two subgroups were compared in order to determine whether order influenced performance.

Results and Discussion

Participants in the single voice congruent condition averaged 20.7 (4.17) out of 32 (65%) possible correct answers on the alternative forced choice task, a level of performance significantly above chance, $t(14) = 4.40$, $p = .001$, $d = 2.35$. Broken down by language, participants averaged 10.3 (2.60) out of 16 (64%) on L1 and 10.5 (2.10) out of 16 (66%) on L2, both significantly above chance—L1: $t(14) = 3.37$, $p = .005$, $d = 1.80$; L2: $t(14) = 4.55$, $p < .001$, $d = 2.43$. The results from this condition, though slightly lower, were not significantly different from those reported in Experiment 2 in which the same languages were presented with an indexical voice cue, $t(28) = .76$, $p = .454$, $d = .29$.

Participants in the single voice incongruent languages averaged 17.7 (2.80) out of a possible 32 (55%) correct choices, a level of performance significantly above chance, $t(15) = 2.41$, $p = .029$, $d = 1.24$ (see Fig. 4). Broken down by individual language, participants averaged 8.63 (1.75) out of 16 (54%) on the incongruent L1 and 9.06 (1.84) out of 16 (57%) on the incongruent L3. Performance on L1 was not significantly above chance, $t(15) = 1.43$, $p = .173$, $d = .74$. However, performance on L3 was significantly above chance, $t(15) = 2.31$, $p = .036$, $d = 1.19$. When directly compared using a paired *t* test, there was no difference in performance on L1 and

L3, $t(15) = -.778$, $p = .45$. There was no order effect overall, $t(14) = 1.46$, $p = .167$, $d = .78$, or for either language individually—L1: $t(14) = 1.29$, $p = .218$, $d = .69$; L3: $t(14) = .94$, $p = .365$, $d = .50$. In comparisons across experiments, there was a significant difference between the results from Experiment 2, Condition 2 (two-voice incongruent) and the single voice incongruent condition reported here, $t(29) = 2.58$, $p = .015$, $d = .96$. Likewise, there was a significant difference in overall scores between the single voice congruent and single voice incongruent conditions, $t(29) = 2.40$, $p = .023$, $d = .89$.

As seen in Experiment 2, Condition 2 above, learners were able to successfully parse each of two speech streams with incongruent statistical properties. That experiment provided listeners with an indexical cue that arguably facilitated the encapsulation of the respective statistics of the incompatible streams, thus allowing for the successful learning of both L1 and L3 by participants. The rationale for Experiment 3a was to test our conclusion by removing the indexical cue. In this context, Experiment 3a yielded two central findings. First, Condition 1 demonstrated that the removal of the indexical voice cue does not itself preclude the successful parsing of two streams. In this condition, the two streams had compatible statistics, thus allowing for successful learning irrespective of whether learners attempted to combine or encapsulate statistics. In Condition 2, however, the streams were incongruent. The logic of our hypothesis is that successful learning in the incongruent condition requires encapsulation of the particular statistical properties of each of the input streams. However, given the lack of an indexical voice cue, there is nothing in the task that would lead participants to learn via encapsulation. Our results support the hypothesis in that removal of the indexical cue resulted in significantly worse performance than in the single voice, congruent condition. Specifically, although we observed a pattern of results in which performance summed across both languages was at levels higher than chance, the overall performance was still lower than when the indexical cue was present. In sum, with congruent statistics, the lack of an indexical cue did not significantly change performance. By contrast, in the incongruent condition, the lack of the indexical cue led to a significant decrement in performance.

It is interesting to note that the combined statistics of the incompatible condition did contain dips in the transitional probabilities at word boundaries (see Fig. 3). However, given that there was more variance in the transitional probabilities within the words (i.e., noise), this information was not sufficient for learners to segment both streams at above chance levels. To date, to the best of our knowledge, there has not been a systematic study of the magnitude of statistical differences required by learners in order to be able to detect word boundaries. Likewise, our results suggest that the magnitude may not be an absolute value, but may fluctuate as a function of the degree of variance within the rest of the stream. It is thus more likely that this value could be expressed as a ratio rather than an absolute magnitude. Future work must consider this issue, as the statistics within real languages likely contain a substantial level of variance.

Alternatively, we considered the hypothesis that presenting L3 in the male voice in the incongruent condition somehow hindered its learnability, despite our results demonstrating that L3 is learnable in the female voice, both in isolation in Experiment 1 and in the two-voice incompatible condition in Experiment 2. To follow up on this possibility, we ran Experiment 3b, the purpose of which was to replicate the single voice incongruent condition in the female voice that we used in Experiments 1 and 2.

EXPERIMENT 3B—SINGLE VOICE FEMALE INCONGRUENT STATISTICS

This experiment was identical to the incongruent single voice condition in Experiment 3a with two important exceptions. First, we used the female voice for both languages rather than the male voices. Second, because we did not find evidence of an order effect in Experiment 3a, this was not counterbalanced in Experiment 3b (which consequently made it more analogous to the congruent single voice condition in Experiment 3a).

Methods

Participants

Eighteen monolingual undergraduates were included from the introductory psychology subject pool with the conditions noted above. In addition, we excluded individuals who did not meet the aforementioned criteria (low effort (4), difficulty following test (2)).

Stimuli and Procedures

The stimuli were structurally identical to those used in Experiment 3 Condition 2, only the single voice was now the same female voice used above in Experiments 1 and 2. The procedure was identical to Experiment 3a.

Results and Discussion

Participants in Experiment 3b averaged 17.22 (4.29) out of a possible 32 correct (54%), a level of performance not significantly above chance, $t(17) = 1.21$, $p = .25$, $d = .59$. Broken down by individual language, participants averaged 9.22 (2.73) out of 16 (58%) on the incongruent L1 and 7.78 (3.07) out of 16 (49%) on the incongruent L3. The performance on L1 trends toward significance, $t(17) = 1.9$, $p = .07$. $d = .92$. Performance on L3 was not significantly above chance, $t(17) = -.31$, $p = .79$, $d = -.15$. Neither the overall results nor the results by individual language are significantly different than those found in Experiment 3 Condition 2—overall: $t(32) = -.37$, $p = .714$, $d = -.13$; L1: $t(32) = .75$, $p = .460$, $d = .27$; L2: $t(32) = -1.45$, $p = .156$, $d = -.51$. As was the case with Experiment 3a, there was a significant difference between the results from Experiment 2, Condition 2 (two-voice incongruent) and the single voice incongruent condition in Experiment 3b, $t(31) = 2.38$, $p = .024$, $d = .85$.

The results from Experiment 3b further confirm that presenting the statistically incongruent speech streams in a single voice precludes learners from correctly parsing both streams as evidenced in Experiment 3a. Participants trended toward above-chance performance on L1, but were right at chance for L3. This somewhat contrasts the findings of Experiment 3a in which performance on L1 did not significantly differ from chance and performance on L3 was just above chance levels. The reason for the fluctuation in performance on individual languages across experiments is unclear. However, the common finding between Experiments 3a and 3b is that when the incongruent languages are presented in only a single voice,

learners do not acquire both languages. Therefore, taking both sets of results into consideration, we conclude that the lack of learning evidenced in the incongruent statistics conditions is directly related to the statistical properties of the input. Future work must address the possible influences of the gender of presentation on these types of statistical learning experiments.

EXPERIMENT 4—CONGRUENT LANGUAGES WITH OVERLAP

We interpreted the results from Experiment 3a and 3b as supporting evidence for the theory that learners can form multiple representations when presented with two speech streams that contain an effective indexical cue. Further, in the absence of an appropriate indexical cue, learners combine statistics across the two streams. Thus, when the statistics of the streams are incompatible, learners fail to segment both languages at above chance levels. This conclusion, however, remains open to an alternative interpretation. One potentially confounding variable across the congruent and incongruent designs is that in the incongruent language pairing, there is overlap across syllables, whereas in the congruent language pairing there is no overlap. The overlapping syllables occur with twice the frequency of other syllables, in more varied contexts, and spoken by both voices. To address this issue, Experiment 4 presented learners with a *congruent* language pairing designed to contain a similar degree of overlap relative to the incongruent languages presented to learners in Experiments 3a and 3b. This was accomplished by creating streams in which the shared syllables played the same role (i.e., as a within-word syllable or a word boundary syllable) across both streams. This resulted in combined statistics that were more robust than nonoverlapping congruent languages, because the word boundary dips in transitional probabilities were somewhat amplified. If the previous findings were truly due to the statistical properties of our languages, and not a function of presence or absence of overlapping syllables, then we predicted that learners should successfully segment both streams in the current experiment at above chance levels. Further, because the dips at word boundaries were larger than in previous language pairings, we also predicted that learners might improve on the previous congruent language pairing.

Method

Participants

18 monolingual undergraduates were included from the introductory psychology subject pool with the conditions noted above. In addition, we excluded individuals who did not meet the aforementioned criteria (difficulty following answer sheet [5]; very low effort [5]; failure to fill out answer sheet correctly [1]).

Stimuli and Procedures

The stimuli consisted of the same L1 used in the previous experiments and a new stream (L4) whose statistics were compatible with L1. Both streams were presented using the male voice. Similar to L3, L4 shared two syllables with L1. The difference between L3 and L4, then, was whether the shared syllables with L1 had compatible positions within the steams (i.e., whether

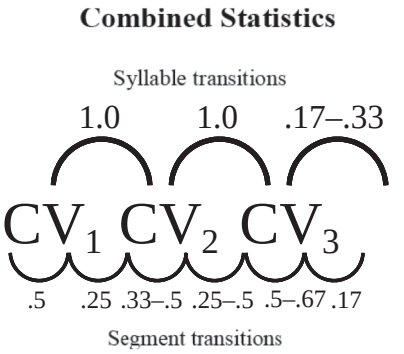

FIGURE 5 This figure shows the combined statistics of L1 and L4.

they were within-word elements or word boundary elements) or not. In the incongruent L1-L3 pairing, the shared syllables were located at different positions across the streams. However, in Experiment 4 we paired L1 and L4 in which the overlapping syllables across streams were located in the same positions across streams (i.e., shared within-word syllables in L1 were also within-word in L4, and shared word boundary syllables in L1 were located at word boundary locations in L4; see Fig. 5). This kept the word-internal statistics consistent and provided a larger dip at word boundaries. The rest of the procedure was identical to Experiments 3a and 3b.

Results and Discussion

Participants in Experiment 4 averaged 20.61 (3.88) out of 32 (64%) possible correct responses. This was significantly above chance, $t(17) = 5.04$, $p < .001$, $d = 2.44$. The average number of correct responses, broken down by language, was 10.22 (1.96) out of 16 (64%) for L1 and 10.39 (2.93) out of 16 (65%) for L4. This level of performance was significantly above chance for both languages—L1: $t(17) = 4.82$, $p < .001$ $d = 2.34$; L4: $t(17) = 3.46$, $p = .003$, $d = 1.68$. Overall performance in Experiment 4 did not differ from that in Experiment 3a, Condition 1, $t(31) = .09$, $p = .932$, $d = .03$. Additionally, performance on L4 did not differ from L2 in Experiment 3a, Condition 1, $t(31) = .09$, $p = .931$, $d = .03$, nor was there a difference in L1 performance, $t(31) = .06$, $p = .956$, $d = .02$. Overall performance in Experiment 4 was significantly greater than performance in Experiment 3b, $t(32) = 2.49$, $p = .018$, $d = .88$, with a significant difference in L1 performance, $t(32) = 2.50$, $p = .018$, $d = .88$, but no significant difference between L4 and L3 performance, $t(32) = 1.55$, $p = .130$, $d = .55$.

The results from Experiment 4 confirmed that the issue of overlapping syllables cannot account for the results reported in Experiments 3a and 3b. In Experiment 4 learners were capable of acquiring both speech streams at above chance levels, despite the fact that the streams shared syllables. This finding reinforces our earlier conclusion that the results from the single-voice incongruent conditions were due to the underlying statistical properties of the streams.

Although learners were capable of acquiring both streams at above chance levels, there was no evidence of an improvement over the previous single-voice congruent language pairing (Experiment 3a) despite the more robust dips at word boundaries. As mentioned, to the best of our knowledge, there has been no systematic study exploring the effects of modifying word

boundary statistics on performance in a segmentation task. It is therefore not clear whether graded effects should be expected. Future work must address this issue, as it has important implications for our understanding of how models of statistical learning can be implemented by learners during the early stages of acquisition.

GENERAL DISCUSSION

The findings from this series of studies extend the statistical learning literature, which has typically focused on segmentation of a single input stream. Here we presented learners with two artificial speech streams in order to determine whether they can form multiple representations by encapsulating the statistics for each stream. In Experiment 1, our control condition, we demonstrated that the artificial speech streams are learnable given 12 minutes of passive exposure. In Experiment 2, we interleaved pairs of languages, each presented in a different voice (one male and one female). In Condition 1, the languages had congruent statistics in that they could be successfully segmented irrespective of whether learners combined statistics across languages or encapsulated the statistics in each language. In Condition 2, the statistics were incongruent, such that combining statistics across streams would likely present difficulties for segmentation (due to increased noise in the statistical structures of the streams). Arguably then, the successful route to segmentation entails encapsulating the statistics in each language. After finding that learners were capable of successfully segmenting streams in both conditions, we conducted Experiment 3. Crucially, this experiment replicated Experiment 2 but removed the indexical voice cue (both streams were presented in the male voice), with the goal of determining whether the underlying statistical structures were learnable in the absence of any cueing. Under these conditions, learners were only capable of successfully segmenting the congruent language pair, suggesting that the noisiness of the statistics in the incongruent condition blocked the successful formation of multiple representations. The order of presentation did not affect these results, thus demonstrating that this finding was not an order effect. In Experiment 3b, we presented the single voice incongruent condition in the female voice, and our findings were consistent. In 3b, participants again failed to correctly segment both of the incongruent streams. In Experiment 4, we demonstrated that these findings cannot be attributed to differences in the degree of overlap between language pairs. Overall, then, our results demonstrate that adult monolingual learners can form multiple representations when confronted with two artificial language inputs, provided that they are given a sufficient indexical cue.

Our findings thus represent the first empirical evidence that language learners can simultaneously compute separate sets of transitional probability statistics over multiple speech stream inputs. The results are compatible with findings from artificial grammar learning studies that demonstrate that multiple grammars can be learned provided that they vary in perceptual dimensions (such as pairing an artificial grammar comprised of colors with one comprised of tones; Conway & Christiansen, 2006). Aside from computational differences (i.e., our experiments require the use of transitional probabilities), our results also extend previous findings in that our input consisted entirely of speech streams, rather than a mix of speech and other auditory or visual stimuli. In fact, the use of multiple representations per se is certainly not unique to language and has been explored in other fields, such as in spatial memory (e.g., McNamara, 2003; see Wang et al., 2005). This general skill may have important applications during real-world language acquisition, be it an adult immersion experience or, perhaps more importantly, during bilingual acquisition in infancy.

Previous research suggests that young infants rely heavily on statistical learning in segmenting a language, particularly during the early stages of acquisition (Saffran et al., 1996a; Thiessen & Saffran, 2003). Indeed, a central issue in bilingual language acquisition research has been determining the point at which listeners can form multiple representations when exposed to more than one language (Genesee, 1989; Pettito et al., 2001; see Sebastián-Gallés & Bosch, 2005 and de Houwer, 2005). The controversy stems at least partly from methodological issues. Research on language production originally surmised that this ability emerges after the third year of life (e.g., Redlinger & Park, 1980; Vihman, 1985; Genesee, 1989). More recent work suggests that this ability may arise between 1 to 2 years of age (De Houwer, 1990; Genesee et al., 1995; Pettito, 2001). By contrast, perception studies suggest that the ability may appear much earlier in the learning process. As mentioned in the introduction, these studies typically focus on the infant's abilities to discriminate between two different languages or between sounds from the languages (e.g., Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998; Bosch & Sebastián-Gallés, 2001; see also Sebastián-Gallés & Bosch, 2005). The reasoning is that if infants discriminate between languages, they should be able to start forming representations that facilitate the acquisition of both systems.

Nevertheless, it is unclear whether discrimination alone suffices to establish that infants are forming multiple representations. Here we define the formation of multiple representations as the ability to perform separate computations over each input language individually. This ability may play a fundamental role during simultaneous bilingual acquisition by allowing infants to avoid confusing the properties of their L1 and L2. Our results provide a useful framework for studying this question. Even if other cues, such as fine-grained phonetic distinctions between sound inventories across languages, may cue infants that they are being exposed to more than one language, ultimately, the ability to perform separate computations on each input language will be necessary. Therefore, a logical extension of this work will be to conduct similar experiments with infants acquiring language. In addition, our research program will investigate how prosodic cues, such as stress, and other language particular differences such as distinct phonotactic patterns and/or different microphonetic segmental cues, interact with the basic statistical encapsulation abilities that our participants have exhibited in this study in order to yield an increasingly sophisticated understanding of how multilingual learners arrive at the creation of the multiple representations necessary for the acquisition of their languages.

Based on the pattern of results found in our experiments, we speculate that the ability to simultaneously acquire two languages may present additional challenges relative to monolingual acquisition. The challenges are twofold: First, learners must encapsulate the input (at least to some degree) in order to perform computations and discover regularities for each language. Second, learners must acquire enough statistical input across both languages in order to segment words from the running stream. These types of challenges are consistent with developmental studies on speech contrast discrimination in young bilinguals, such as reports of a longer developmental trajectory for discriminating infrequent language-specific contrasts in the course of bilingual language acquisition (e.g., Bosch & Sebastián-Gallés, 2003) and delays in the facilitation of native-language speech sound contrasts in 4-year old bilingual children (Sundara, Polka, & Genesee, 2006). However, the issue of how bilinguals represent their phonetic space is still under debate (e.g., Flege, 1995; Grosjean, 1997; Bosch, Costa, & Sebastián-Gallés, 2000) and there exists evidence from developing bilinguals suggesting they are not delayed in their development of phonetic representations relative to monolinguals (e.g., Burns et al., 2007). Together with the results of the present study, these findings suggest the need for developmental studies specifically focused on speech segmentation in developing bilingual children.

Future studies must also assess the degree to which learners form speaker-specific representations versus individual language representations. In our mixed experiments, these two variables were confounded as there were two speakers, one for each language. There is a growing body of research on perceptual categorization demonstrating that language learners incorporate information about particular speakers into their perceptual representations (e.g., Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Schacter & Church, 1992; Kraljic & Samuel, 2005; Creel, Aslin, & Tanenhaus, 2008). Given that memory for speakers can aid in the perception and interpretation of speech streams (e.g., Mullinex & Pisoni, 1990; see Kraljic & Samuels, 2005; Newman & Evers, 2007), it is possible that providing speaker voice as an indexical cue was particularly salient in our task. Future studies will manipulate the number of speakers per language to see whether learners combine across different speakers and if so, at what age this ability comes online. Our planned studies testing whether other acoustic or statistical features may be used as indexical cues when no speaker cue is present will also provide relevant data addressing this issue.

ACKNOWLEDGMENTS

The authors would like to thank Allison Allmon, Mark Basile, Jill Boelter, and Molly Jamison for conducting experiments. We would also like to thank Wendy Rizzo for help in synthesizing and assembling the artificial language stimuli. We are grateful to Dick Aslin and Judy Kroll for helpful comments on an earlier draft and to NIH R03 grant HD048996-01 for support during the preparation of this article.

Portions of this research were presented at the 30th annual Boston University Conference on Language Development and the 46th Annual Meeting of the Psychonomic Society.

REFERENCES

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bosch, L., Costa, A., & Sebastián-Gallés, N. (2000). First and second language vowel perception in early bilinguals. *European Journal of Cognitive Psychology*, *12*, 189–221.

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, *2*(1), 29–49.

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a Language specific vowel contrast in the first year of life. *Language and Speech*, *46*, 217–244.

Burns, T.C., Yoshida, K.A., Hill, K., & Werker, J.F. (2007). The development of phonetic representation in bilingual and monolingual infants. *Applied Psycholinguistics*, *28*, 455–474.

Cole, R., & Jakimik, J. (1980). *A model of speech perception*. Hillsdale, NJ: Erlbaum.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities. *Psychological Science*, *17*(10), 905–912.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *16*, 633–664.

Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge, England: Cambridge University Press.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

De Houwer, A. (2005). *Early bilingual acquisition: Focus on morphosyntax and the separate development hypothesis*. New York: Oxford University Press.

DeHouwer, A. (1990). *The acquisition of two languages from birth: A case study*. Cambridge, England: Cambridge University Press.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*, 224–238.

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 233–272). Baltimore, MD: York Press.

Genesee, F. (1989). Early bilingual development: One language or two? *Journal of Child Language, 16*, 161–179.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word indentification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*(5), 1166–1183.

Grosjean, F. (1997). Processing mixed language: Issues, findings and models. In A. de Groot & J. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives*. Mahwah, NJ: Erlbaum.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton-top tamarins. *Cognition, 78*, B53–B64.

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.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262–268.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), 101–111.

McNamara, T. P. (Ed.). (2003). *How are the locations of objects in the environment represented in memory?* Berlin: Springer.

Mehler, J., Jusczyk, P. W., Lambertz, G., Halsted, G., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143–178.

Miller, J. L., & Eimas, P. D. (1995). Speech perception: From signal to word. *Annual Review of Psychology, 46*, 467–492.

Mitchel, A., & Weiss, D. J. (in review). What's in a face? Visual contributions to speech segmentation. *Language and Cognitive Processes*.

Mullinex, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 756–766.

Newman, R.S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics, 35*, 85–103.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance: II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology, 49*, 85–117.

Newport, E. L., Weiss, D. J., Wonnacott, E., & Aslin, R. N. (2004). *Statistical learning in speech: Syllables or segments?* Presented at Boston University Conference on Language Development, Boston, MA.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*, 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Pettito, L. A., Katerelos, M., Levy, B., Gauna, K., Tétrault, K., & Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: Implications for mechanisms underlying bilingual language acquisition. *Journal of Child Language, 28*, 453–496.

Ramus, F., Hauser, M.D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and cotton-top tamarin monkeys. *Science, 288*, 349– 351.

Redlinger, W., & Park, T. Z. (1980). Language mixing in young bilingual children. *Journal of Child Language, 7*, 337–352.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Directions in Psychological Science, 12*, 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(5), 915–930.

Sebastián-Gallés, N., & Bosch, L. (Eds.). (2005). *Phonology and bilingualism*. New York: Oxford University Press.

Sundara, M., Polka, L., & Genesee, F. (2006). Language experience facilitates discrimination of /d–Δ/ in monolingual and bilingual acquisition of English. *Cognition*, *100*, 369–388.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- and 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716.

Tincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F., & Mehler, J. (2005). The role of speech rhythm in language discrimination: Further tests with a nonhuman primate. *Developmental Science*, *8*(1), 26–35.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34.

Toro, J. M., & Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics*, *67*(5), 867–875.

Toro, J.M., Trobalon, J.B., & Sebastián-Gallés, N. (2003). The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition*, *6*(2), 131–136.

Vihman, M. M. (1985). Language differentiation by the bilingual infant. *Journal of Child Language*, *12*, 297–324.

Wang, H., Johnson, T. R., Sun, Y., & Zhang, J. (2005). Object location memory: The interplay of multiple representations. *Memory & Cognition*, *33*(7), 1147–1160.