Beyond Noise: The Role of Speaker Variability on Statistical Learning

Larissa S. Melville (lmelvill@student.ubc.ca)
School of Audiology and Speech Sciences, Friedman Building 443, 2177 Wesbrook Mall
Vancouver, BC, V6T1Z3, Canada

Alexis K. Black (alexis.black@audiospeech.ubc.ca)
School of Audiology and Speech Sciences, Friedman Building 443, 2177 Wesbrook Mall
Vancouver, BC, V6T1Z3, Canada

Abstract

Adult language learners have difficulty segmenting words from continuous speech when the phonology is unfamiliar. Since speaker variability is known to improve acquisition of novel language structures, it could be processed in ways that bootstrap phonological patterns and enhance learners' ability to segment words. To test this, the present experiment examined adult participants' learning of a stream of statistically determined tri-syllabic words that were spoken by one or multiple speakers. Syllables were constructed with either English phonology or non-English phonology. Two tasks (target detection and two-alternative forced choice) assessed the extent of listeners' sensitivity to language patterns and word segmentation. Results suggest speaker variability negatively impacted learners' ability to track the underlying statistics. 2AFC word segmentation performance was poor—independent of speaker number; it is hypothesized that attentional demands of the target detection task conflicted with statistical word segmentation mechanisms.

Keywords: statistical learning; multi-speaker variability; language learning; phonology

Introduction

Statistical learning (SL) explains in part how people form expectations about language structure (e.g., word segmentation; Saffran, 2001). Listeners are capable of extracting words when the SL conditions are highly stable or straightforward. For example, infants show evidence of segmenting tri-syllabic sequences from a stream of syllables which has no cues to word boundaries beyond the transitional probabilities (TPs) between the syllables (Saffran, Aslin, et al., 1996). Similar results are observed for adult listeners, both in-person (Arciuli et al., 2014; Saffran et al., 1999; Saffran, Newport, et al., 1996), and online (Craparotta et al., 2022). Successful word segmentation is evident in these contrived learning environments where the statistically defined words are isomorphic and there are few task-level demands.

When the SL stream becomes more complex, capacity for word segmentation appears to decrease. For example, infants have trouble segmenting streams that contain words of different syllable-lengths (Lew-Williams & Saffran, 2012), or segmenting words from two statistically balanced speech streams (presented sequentially) with different underlying statistics, even when pitch and accent cues mark word boundaries (Benitez et al., 2020). For adults, concurrent visual or auditory interference from a non-SL task leads to impairment of word segmentation ability (Toro et al., 2005).

Other examples concern the phonological characteristics of the SL stream. A statistical stream that contains a familiar phonology but simultaneously violates phonotactic constraints in a listeners' first language impairs word segmentation (Finn & Hudson Kam, 2008; Toro et al., 2011). Similarly, when the phonology of a SL stream is entirely unfamiliar to adult listeners, their word segmentation performance is compromised compared to a stream with familiar or even partially unfamiliar phonology, even after quadrupling the amount of exposure (Black, 2018; Black & Hudson Kam, 2024). Indeed, second language learners fail to segment words from unfamiliar-sounding streams even when the segmentation cues are more linguistically robust than just statistical probabilities (Snijders et al., 2007). These findings reveal a decreased capacity for word segmentation that may be attributed to participants' nescience of phonological forms.

This introduces a puzzle. That is, for infants, statistical word segmentation has been proposed as one of the earliest tools to break the speech code (Aslin, 2017), and has been demonstrated as early as within the first few hours after birth (Teinonen et al., 2009)—well before infants have established a phonological repertoire (c.f., Werker & Curtin, 2005), which adults *do* possess. Infants are successful at SL, even without an established phonological inventory, but the literature reviewed above suggests that adults may not have the same capacity. However, a critical component of the infant's early learning environment that is missing from these highly artificial learning tasks is *phonetic variability.*

Speaker Variability on Language Learning

There are multiple sources of variability in language input (see Quam & Creel, 2021, for a review). In this study, the focus is on one broad form: *between-speaker variability*. Between-speaker variability characterizes structured acoustic factors that signal speech differences from person-to-person. It does not naturally create differences in meaning, but it is by no means random (Kleinschmidt, 2019). Across individuals, there exist varied anatomical vocal tract characteristics (Johnson et al., 1993), vocal pathology (Kreiman et al., 2003), age and gender (Perry et al., 2001; Rojas et al., 2020), and unique socially indexed features like sociolect (Schultz, 2007), dialect (Labov et al., 2006), and accent (Yan et al., 2003), which all cause phonetic differences across different speakers.

Between-speaker variability has been shown to *boost* the acquisition of linguistic structures. Learners exposed to

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

multiple voices at training are more successful at generalizing phonetic contrasts, such as the /l/ and /ɹ/ English contrast for first language Japanese speakers (Lively et al., 1993) or differentiating between the phonological contrasts involved in minimal pairs (e.g., /buk/ vs /puk/; Rost & McMurray, 2009). A multi-speaker benefit has also been indicated at the word level. Infants more successfully segment words from a stream of speech (Graf Estes & Lew-Williams, 2015), and recognize phonotactic patterns (Seidl et al., 2014), when they are presented in multiple voices. There is also evidence of improved generalization for second-language vocabulary in multi-speaker conditions for adults (Barcroft & Sommers, 2005). Speaker variability has even shown improvement in the acquisition and generalization of new grammatical rules (Gonzales et al., 2018).

It was thus hypothesized that speaker variability would aid in statistical word segmentation when the phonology was unfamiliar. In this study, English-speaking adult learners were exposed to statistically defined syllable sequences composed of either English or non-English sounds, which were produced by a single or six different speakers (i.e., four conditions). The single speaker and multi-speaker English sound conditions were control conditions: it was anticipated that learners would be successful in both, and sufficiently proficient in their English phonology that there would be no difference between the single speaker and multi-speaker conditions. The two experimental conditions were therefore the single-speaker non-English and multi-speaker non-English. Worse performance was predicted for the single-speaker non-English (i.e., less familiar) condition in comparison to the two control conditions, and relatively improved performance in the multi-speaker non-English condition. If learners are successful in this task, it suggests that phonetic variability can enhance sensitivity to the statistical patterns in the input, perhaps by bootstrapping novel phonological forms. On the other hand, if learners fail, it suggests that variability is insufficient to improve acquisition of the unfamiliar phonetic SL input.

Methodology

One hundred adult participants were recruited using Prolific (www.prolific.co; Palan & Schitter, 2018). Seventeen participants were excluded due to server issues corrupting their data. Five additional participants met exclusion criteria (i.e., 100% miss rate on any of the experimental blocks) and were removed. In total, 78 participants (mean age = 25.3; SD = 8.8) were included in the analyses. All 100 participants were compensated 3.55£ for their time; median completion time of the study was 24 minutes and 57 seconds.

All participants met the screening requirements of (1) being aged between 18 and 100, (2) speaking English fluently though not necessarily as a first language, and (3) having not participated in a previous pilot study. No restrictions were put on participant location; therefore, responses were collected from individuals residing in a large variety of countries. Command of English was recorded: 59 participants labeled themselves as 'Proficient,' 17 as 'Intermediate,' and two participants did not indicate their proficiency.

Materials

Syllables were recorded by seven English-speaking, phonetically trained individuals in a sound-proofed booth. They were produced in isolation, to prevent coarticulation effects. To ensure consistency of target for non-English syllables, one speaker (Speaker 1) was used as a model for all other speakers. Speakers were instructed to produce as many tokens of each syllable as necessary to mimic the model syllables and/or approximate the IPA dictation as precisely as possible. English syllables were produced with reference to the IPA symbols only.

The syllables were then lengthened or shortened to 400ms, without any alterations to pitch. An artificial silence of 100ms was added to each syllable—50ms at the beginning and 50ms at the end. Intensity was normalized across tokens. Overall, there were a total of 84 syllables per phonology condition. Speaker 7's productions were used only for the 2AFC test items. Twelve non-linguistic stimuli were also selected for the training task. These were selected from a previously created set of nonsense sounds and underwent the same alterations as the experimental stimuli.

Syllable Phonology Two sets of twelve syllables were created to test the impact of unfamiliar phonology: English and non-English. They were derived from the syllable structure of the original speech segmentation studies (Saffran, Aslin, et al., 1996; Saffran, Newport, et al., 1996). The familiar set of syllables contained a phonetic inventory that corresponded to sounds in English: /bi, dʌ, ku, go, lʌ, bu, tu, pi, ɹo, pʌ, do, ti/. The unfamiliar set contained consonants and vowels that do not occur in the English phonetic inventory but do appear in other languages: /βy, ɖɒ, q'ɯ, ɠœ, ʎɒ, βɯ, ʈɯ, ɸy, Rœ, ɸɒ, ɖœ, ʈy/.

2AFC Test Items Participants were tested with a two-alternative forced choice (2AFC) paradigm, which exposed them to three different kinds of test items (Table 1): words, part-words, and non-words. To create these stimuli, Speaker 7's tokens (excluded from the exposure streams) were spliced into tri-syllabic sequences.

Table 1: 2AFC task contrast types per phonology.

	Words	Part-Words	Non-Words
English	bidʌk^{h}u	dʌk^{h}ugo	gok^{h}udʌ
	golʌbu	lʌbut^{h}u	t^{h}ubulʌ
	t^{h}up^{h}iɹo	p^{h}iɹop^{h}ʌ	p^{h}ʌɹop^{h}i
	p^{h}ʌdot^{h}i	dot^{h}ibi	bit^{h}ido
Non-English	βyɖɒq'ɯ	ɖɒq'ɯɠœ	ɠœq'ɯɖɒ
	ɠœʎɒβɯ	ʎɒβɯʈɯ	ʈɯβɯʎɒ
	ʈɯɸyRœ	ɸyRœɸɒ	ɸɒRœɸy
	ɸɒɖœʈy	ɖœʈyβy	βyʈyɖœ

To define these contrast types, first, words had high TPs (1.0) between syllables within the word; boundaries on either side of the words had TPs that were low (0.33). Second, "part-words" were tri-syllabic sequences that spanned over a word border. These contained the final two syllables of one of the words (TP = 1.0) combined with the first syllable of another word (TP = 0.33). Finally, non-words were triplets that did not occur in the exposure streams; both within-word boundaries had TPs of 0.0. In total, there were 12 tri-syllabic combinations per phonology, with four words, four part-words, and four non-words.

Procedures

This experiment was conducted online. Participants first completed a short introduction which included a consent form, language questionnaire, and audio testing before being sorted into the non-English or English conditions, and further, into single-speaker or multi-speaker conditions.

This paradigm follows the general procedure described by Lukics and Lukács (2021). That is, following a short practice block, participants began the experimental *target detection* trials. They were told to listen to a stream of sounds and to press the 'Z' key every time they heard a target (familiarized beforehand). Instructions said to answer as fast, but as accurately as possible. This section allowed participants to listen to the target syllable before interacting with the stream. Only the four syllables in the final position of the tri-syllabic words were possible targets, because the predictability of word-final syllables in a SL paradigm outweighs that of the first and even second syllable in a tri-syllabic sequence (Batterink et al., 2015; Franco, Eberlen, et al., 2015). A single target was randomly selected for each participant.

The target detection task contained five blocks: three training blocks, one random block, and one recovery block, (Figure 1). In each block, the participant listened to the stream of syllables. The single-speaker condition randomly selected one of the six possible voices per participant, to be heard across all the blocks. The multi-speaker conditions switched the voice to a new one (i.e., Speakers 1 to 6) every 10 to 20 syllables (Graf Estes & Lew-Williams, 2015), irrespective of the word boundaries in the stream to avoid extraneous cues to their position. Untimed participant-controlled breaks were permitted at the end of each block; the participants could re-listen to their target at this time.

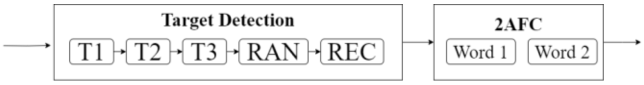

Figure 1: Target detection and 2AFC block structure.

Across the structured blocks, each of the four words was repeated 60 times in total. The words were played in a balanced order where each word followed the others equally as often (but the same word did not appear twice in a row), so that TPs within words were 1.0 and TPs across words were 0.33. The random stream was created such that TPs between syllables fluctuated unpredictably and did not create words.

Following the target detection task, pairs of tri-syllabic sequences were heard by participants in the 2AFC task, one after another, with a 500ms pause in-between. They were instructed to press 'A' if the first word occurred in the stream they heard, or 'L' if it was the second word. Three individual tri-syllabic sequence presentations are indicated to be the "sweet-spot" for data that show above-chance performance without sacrificing reliability and validity (Siegelman et al., 2017). For the current experiment, 24 trials were required to balance contrasts across the three word-types (i.e., words vs non-words, words vs. part-words, and part-words vs. non-words), meaning a total of four presentations of each tri-syllabic sequence was necessary. Presentation order was randomized.

Results

The 78 participants included in the final analysis were counterbalanced across the four conditions for a total of 20 participants in the single-speaker English group, 20 in the multi-speaker English group, 19 in the single-speaker non-English group, and 19 in the multi-speaker non-English group. Participants in the English conditions were randomly selected for the /ku/ target syllable 14 times, /bu/ 11 times, /ɹo/ 7 times, and /ti/ 8 times. For the non-English condition, 6 participants heard /ʈy/, 12 heard /Rœ/, 7 heard /βɯ/, and 13 heard /q'ɯ/. In the single-speaker conditions (i.e., regardless of phonology), four participants heard Speaker 1, seven heard Speaker 2, seven heard Speaker 3, seven heard Speaker 4, ten heard Speaker 5, and four heard Speaker 6.

Target Detection

Response Time Participant responses were considered successful (i.e., hits) when they were made within 1200ms from stimulus onset, following previous studies (Batterink et al., 2015; Lukics & Lukács, 2021). Responses outside of this time window were discarded. Mean response times (RTs) by block are reported in Table 2; median RTs and interquartile ranges are shown in Figure 2.

Table 2: Mean response times (ms) per condition. 'E' stands for "English", 'NE' for "Non-English". The 1 refers to single-speaker conditions, and the 6 refers to multi-speaker.

	E1	E6	NE1	NE6
T1	604.1	677.8	632.2	691.6
T2	595.9	688.5	670.7	694.5
T3	614.6	677.5	606.6	667.9
Random	719.0	731.5	754.1	675.6
Recovery	609.9	663.6	667.9	696.1

To gauge participants' sensitivity to the statistics of the streams, RT (and d′) changes through the blocks were analyzed in two distinct ways: (1) to evaluate learning across the first three structured blocks (i.e., T1 – T3) by looking for a decrease in RT across blocks, and (2) to assess the impact of the random block compared to the two structured blocks immediately before and after it (i.e., T3 and REC). Because the normality of residuals assumption was violated for a parametric ANOVA as revealed by a Shapiro-Wilk test, a non-parametric Friedman test was utilized for each condition. The tests showed there was a statistically significant effect of block within three conditions (single-speaker English: $\chi^2(4) = 28.40$, $p < .001$; multi-speaker English: $\chi^2(4) = 11.96$, $p = .018$; single-speaker non-English: $\chi^2(4) = 24.30$, $p < .001$), but was not significant for the multi-speaker non-English condition.

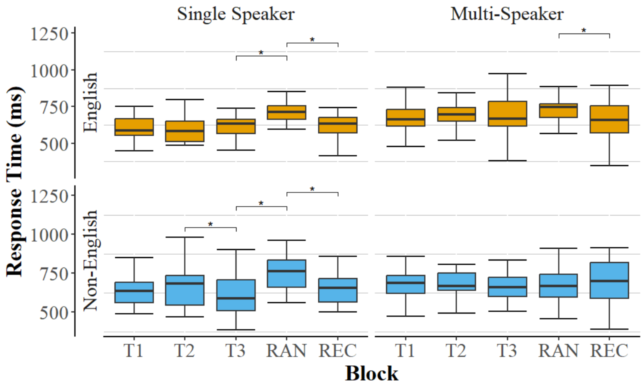

Figure 2: Median RTs per block by phonology and speaker conditions. Box edges indicate interquartile range, and whiskers indicate maximum and minimum data (excluding outliers).

To identify significant RT changes between sequential blocks (within each condition), post-hoc Wilcoxon signed-rank tests were conducted on the by-subject aggregated RT data. Four comparisons were completed per condition. The single-speaker English condition showed significant differences between two block pairs (T3 – RAN: Z = -3.74, $p < .001$, r = -0.84; RAN – REC: Z = -3.25, $p = .003$, r = -0.73), the multi-speaker English indicated one (RAN – REC: Z = -3.03, $p = .008$, r = -0.68), and the single-speaker non-English showed three (T2 – T3: Z = -2.79, $p = .01$, r = -0.64; T3 – RAN: Z = -3.86, $p < .001$, r = -0.88; RAN – REC: Z = -3.13, $p = .006$, r = -0.72). One-tailed p-values were adjusted with sequential Holm-Bonferroni corrections, the direction of which depended on the predicted outcome of the comparison.

If participants are gaining sensitivity to the underlying statistics across blocks T1 to T3, they should be impeded by the unpredictable sequences in the random block. The comparisons reveal that RTs are indeed slower for the random, unpredictable block than for structured blocks in the single-speaker conditions (both English and non-English). However, neither of the multi-speaker conditions robustly demonstrate this learning effect (evidence only between RAN and REC for the multi-speaker English condition). Additionally, as participants become more sensitive to the stream statistics, their RTs are expected to decrease—only the single-speaker non-English condition showed minimal evidence of this expected learning effect.

D-Prime D-prime (d′) is a measure of listeners' signal sensitivity and discriminability. It is calculated as the difference between the means of the target-present and target-absent distributions, as a composite of hits (i.e., response made within 1200ms) and false alarms (i.e., responses made outside the 1200ms window).

D-prime changes across the blocks were analyzed per condition (Figure 3). Again, since the normality of residuals assumption was violated for a parametric ANOVA (as indicated by the Shapiro-Wilk test), the Friedman statistic was utilized—it showed there was a statistically significant effect of d′ on block for the control conditions: single-speaker English ($\chi^2(4) = 17.61$, $p = .001$) and multi-speaker English ($\chi^2(4) = 20.07$, $p < .001$). There was also a significant effect for the multi-speaker non-English condition ($\chi^2(4) = 19.07$, $p < .001$). The single-speaker non-English condition did not show a significant effect of block.

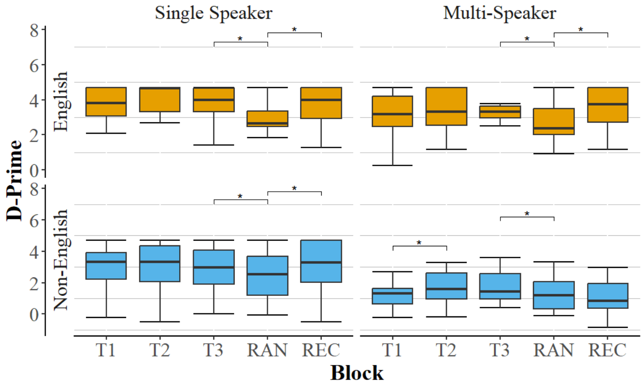

Figure 3: Median d′ scores per block by phonology and speaker conditions. Box edges indicate interquartile range, and whiskers indicate maximum and minimum data (excluding outliers).

Post-hoc Wilcoxon signed-rank tests were applied to the calculated d′ data. The single-speaker English condition showed significant differences between two block pairs (T3 – RAN: Z = -3.09, $p = .008$, r = -0.69; RAN – REC: Z = -2.79, $p = .015$, r = -0.62), the multi-speaker English indicated two (T3 – RAN: Z = -3.47, $p = .039$, r = -0.56; RAN – REC: Z = -2.79, $p = .004$, r = -0.78), the single-speaker non-English showed two (T3 – RAN: Z = -2.55, $p = .044$, r = -0.59; RAN – REC: Z = 2.55, $p = .044$, r = -0.59), as did the multi-speaker non-English condition (T1 – T2: Z = -2.53, $p = .033$, r = -0.58; T3 – RAN: Z = -3.20, $p = .004$, r = -0.73).

Differences in d′ values between the random block, and the structured blocks immediately before and after it are significant for the single- and multi-speaker English conditions as well as the single-speaker non-English condition. However, the multi-speaker non-English condition shows this effect only between T3 and RAN, but not RAN and REC. No condition showed evidence of improved d′ scores over the structured blocks, except for the multi-speaker non-English group—but only between T1 and T2.

Two-Alternative Forced Choice

Of the three contrasts used in this study (i.e., words, part-words, and non-words), part-words versus non-words were removed from 2AFC analyses, as they do not directly test word segmentation. Accuracy data are visualized in Figure 4.

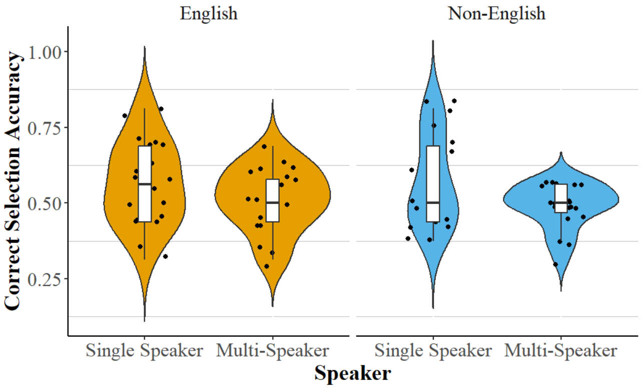

Figure 4: 2AFC median accuracy per block given phonology and speaker conditions. Wider areas of the coloured violin depict higher density of participant scores. On the inner box plot, black lines indicate median values, box edges indicate interquartile range, and whiskers indicate maximum and minimum values without outliers.

One sample t-tests comparing mean accuracy to a chance-level assumed mean of 50% did not indicate significantly above chance values for the single-speaker English condition (M = 56.3%; t(19) = 2.01; p = .058), multi-speaker English condition (M = 51.3%; t(19) = 0.53; p = .6), single-speaker non-English condition (M = 56.9%; t(18) = 2.01; p = .06), nor the multi-speaker non-English condition (M = 49.0%; t(18) = -0.59; p = .563). These findings suggest that no group chose word-forms over part- or non-word forms significantly more often than chance.

Discussion

This study proposed that between-speaker variability would boost statistical word segmentation of phonetically unfamiliar speech. Contrary to this hypothesis, the results of this experiment show that (1) speaker variability impedes sensitivity to the underlying statistics of both phonetically familiar *and* unfamiliar streams and that (2) participants do not show robust evidence of word segmentation and/or generalizability using this design. The latter point is discussed in the limitations and future directions section.

To address point (1), the results indicated that exposure to speaker variability decreased listeners' ability to track the statistical patterns in the stream, regardless of phonology condition. That is, participants exposed to multiple speakers showed less robust evidence for gaining sensitivity to the statistics of the stream as revealed by the relative uniformity of RTs and d′ scores across the blocks (i.e., the random stream did not disrupt learning). Evidently, the presence of underlying regularities in the structured streams was insufficient to garner sensitivity in the face of speaker variability. Speaker variability thus appears to actively interfere with processing TPs, hinting to decreased learning efficiency no matter the stream phonology.

One explanation is that variability is processed in ways that are cognitively load-intensive. Indeed, processing costs associated with speaker variability are notable in several studies that concern various aspects of speech production (Goldinger, 1998; Martin et al., 1989; Mullennix et al., 1989). There is even non-behavioural neuroimaging evidence suggesting that the brain reacts to speaker variable input in different ways than when speaker is constant (Perrachione et al., 2016). One study found that any source of variability slowed RTs and concluded that variability (as a whole) imposes processing costs on speech perception (Kapadia et al., 2023). There is some evidence to suggest learning language structure amidst variable speech requires adjustment to additional *acoustic-phonetic disparities* (Luthra, 2023; Mullennix et al., 1989); costs are also consistent with load-intensive *attentional mechanisms* necessary to process task-irrelevant variability (Lim et al., 2019; Luthra, 2023).

However, not all studies analyzing variability show negative impacts on learning (e.g., Rost & McMurray, 2009)—and the current findings may still be compatible with research that shows multi-speaker benefits. That is, it is possible variability positively impacts some kinds of learning, but not others. For instance, it is postulated that variability improves the *generalization* of target items to novel forms (and voices; Barcroft & Sommers, 2005), while simultaneously slowing learners' efficiency and speed of acquiring those target structures (Raviv et al., 2022). Indeed, individuals can be less accurate and slower while interacting with linguistic structures that are produced by multiple different talkers (Kapadia et al., 2023; Martin et al., 1989; Mullennix et al., 1989; Stilp & Theodore, 2020), even with reduced ambiguity between tokens (Choi et al., 2018) or increased familiarity of speakers to the listener (Magnuson et al., 2021). In the present study, given the relatively poor performance in d′ and RT values across both multi-speaker conditions, these results are consistent with the hypothesis that speaker variability decreases the efficiency of learning (in this case, statistical patterns). Since listeners who encounter variability are often better able to generalize learning to novel structures (e.g., Barcroft & Sommers, 2005; Lively et al., 1993), it is possible the current multi-speaker listeners may have performed at the same level or better than those who did not. However, since learners were at chance in all four conditions, it was impossible to evaluate the impact of variability on generalization in this study.

One solution is to increase participants' length of exposure to the statistical learning streams. Indeed, when learning a complex system, greater exposure to more modulated stimuli streams is shown to optimize performance (Schiff et al., 2021). However, the authors tested this hypothesis, and it was clear that increasing length of exposure did not improve participants' learning outcomes (Melville, 2024).

Limitations and Future Directions

The second primary finding was that learners did not segment words as expected—the 2AFC scores were at chance for every condition. This is unexpected given the extensive literature on adult statistical word segmentation with 2AFC (e.g., Arciuli et al., 2014; Saffran et al., 1999).

The current most compelling explanation is that the target detection task itself actively interfered with word segmentation. Critically, the target detection task occurred immediately before the 2AFC task so that the participants were *reacting* to single syllable targets instead of listening *passively*. This explicit action may have prevented the typical encoding[1] of tri-syllabic sequences perhaps through depletion of cognitive resources or overburdened attentional mechanisms—limiting participants' ability to track the TPs. Indeed, divided attention is known to reduce performance on word segmentation tasks (Toro et al., 2005; Turk-Browne et al., 2005). Word segmentation ability was also hindered when listeners explicitly actively responded to a non-linguistic click in a statistically balanced syllable stream (Franco, Gaillard, et al., 2015). The current paradigm is in one sense more ecological than clicks, as it tests the statistically balanced stream in the same conditions it was learned in and has been shown to effectively target word segmentation in previous studies (Lukics & Lukács, 2022)—but even so, word segmentation was negligible. Overall, it is possible the act of detecting targets negatively impacted word segmentation. A future iteration of this experiment could reverse the order of the tasks: presenting the target detection task *after* the 2AFC task (see Batterink et al., 2015; Batterink & Paller, 2017; Franco, Eberlen, et al., 2015, for examples), to bypass possible interference.

Four other methodological issues may have additionally/alternatively impacted word segmentation performance in this task. First, the experiment was conducted online, and therefore introduced a lack of control over environmental conditions. Collecting data in-person may circumvent these environmental effects, such as by ensuring proper use of headphones, audio volume and quality, instruction delivery, and by preventing data collection inaccuracies. That being said, online studies programmed with jsPsych (the JavaScript plugin library utilized for the present experiment) have shown high validity and reliability (de Leeuw & Motz, 2016) and similar SL studies have been successfully replicated online (Craparotta et al., 2022).

Second, multilingualism and diverse linguistic experiences are known to impact several aspects of language acquisition (Selinker & Baumgartner-Cohen, 1995). Since the present experiment recruited multilingual participants, narrowing the population to a first-language English sample could prevent any unexpected influences from diverse linguistic backgrounds. A recent follow-up study, however, suggests that such a manipulation shows little difference from the present results (Melville, 2024).

A third possibility was explored: learning might be impacted by adding a *correct rejection requirement* to the target detection task. That is, a key difference between this experiment and the paradigm set by Lukics and Lukács (2021) is that they required participants to indicate the presentation of non-targets as well as targets (i.e., correct rejection criteria). Rate and distribution of decision types (i.e., positive or negative) and the impact of being correct or incorrect all weigh into the resulting participant behaviours (Green & Swets, 1966). It was predicted that the addition of a correct rejection criterion could therefore dramatically shift the learning trajectories and word segmentation capacity for this experiment, but in subsequent experiments by the authors, this was also shown to be false (Melville, 2024).

Finally, it is also possible that the adult listeners were tracking the frequent voice changes (i.e., every 10 to 20 syllables) while listening to the syllable streams, rather than the statistically determined syllable sequences (opposing what is observed in infant studies; Graf Estes & Lew-Williams, 2015). In other words, task-irrelevant speaker information may have unintentionally acted as a cue to segmentation, interfering with the tracking of the underlying statistical patterns. Additionally, if infants integrate speaker variability and word segmentation differently than adults, then successful findings with infants' interactions with rapidly changing voices (i.e., Graf Estes & Lew-Williams, 2015) may not apply to adult participants. Instead, voice changes could have interfered with encoding of the underlying statistics, because adults have learned to weigh voice changes as an absolute cue for category (e.g., word) boundaries. Similarly to the proposition that frequent voice changes interrupt attention, one solution could be to alternate voices by block, as opposed to within block.

Conclusion

In summary, adult listeners are less efficient at tracking TPs in multi-speaker conditions, regardless of phonological familiarity. Word segmentation accuracy was around chance for all conditions, possibly because learning mechanisms involved with target detection interfered with listeners' ability to segment words. Future studies implementing workarounds to mitigate this issue (i.e., presenting the target detection task after the 2AFC trials), and studies addressing other solutions (i.e., in-person data collection, and alternating speakers per block) may circumvent the word segmentation problem and answer the question of how speaker variability impacts word segmentation in the context of unfamiliar phonology.

Acknowledgments

This research was funded by the Social Sciences and Humanities Research Council of Canada Insight Development Grant 430-2020-00430 and Canada Graduate Scholarship Master's Award #6566.

[1] How statistically influenced sequences are encoded and stored in memory is still under investigation, though some theories suggest extraction of TPs between syllables, and extracting word-like chunks, among others (see Endress et al., 2020 for a review).

References

Arciuli, J., von Koss Torkildsen, J., Stevens, D. J., & Simpson, I. C. (2014). Statistical learning under incidental versus intentional conditions. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00747

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. In *Wiley Interdisciplinary Reviews: Cognitive Science* (Vol. 8, Issues 1–2). Wiley-Blackwell. https://doi.org/10.1002/wcs.1373

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387–414. https://doi.org/10.1017/S0272263105050175

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. https://doi.org/10.1016/j.cortex.2017.02.004

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, *83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Benitez, V. L., Bulgarelli, F., Byers-Heinlein, K., Saffran, J. R., & Weiss, D. J. (2020). Statistical learning of multiple speech streams: A challenge for monolingual infants. *Developmental Science*, *23*(2). https://doi.org/10.1111/desc.12896

Black, A. (2018). *How Perception Constrains Statistical Learning Across Development* [Doctoral Dissertation]. University of British Columbia.

Black, A., & Hudson Kam, C. (2024). The mechanisms of statistical learning: evidence for position-based encoding during word segmentation. In *[Manuscript in Preparation]*.

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, and Psychophysics*, *80*(3), 784–797. https://doi.org/10.3758/s13414-017-1395-5

Craparotta, A., Feinberg, L., Kocen, D., Koelling, R., Ricketts, W., Wong, A., Jennings, M., Mejia, M., & Hartshorne, J. K. (2022). Fifth replication of Saffran, Newport, & Aslin (1996) Exp. 1. In *Fifth replication of Saffran* (Issue 1). https://psyarxiv.com/cbu2m/.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10.3758/s13428-015-0567-2

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499. https://doi.org/10.1016/j.cognition.2008.04.002

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation: A new measure of statistical learning in speech segmentation. *Experimental Psychology*, *62*(5), 346–351. https://doi.org/10.1027/1618-3169/a000295

Franco, A., Gaillard, V., Cleeremans, A., & Destrebecqz, A. (2015). Assessing segmentation processes by click detection: online measure of statistical learning, or simple interference? *Behavior Research Methods*, *47*(4), 1393–1403. https://doi.org/10.3758/s13428-014-0548-x

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *105*(2), 251–279.

Gonzales, K., Gerken, L. A., & Gómez, R. L. (2018). How who is talking matters as much as what they say to infant language learners. *Cognitive Psychology*, *106*, 1–20. https://doi.org/10.1016/j.cogpsych.2018.04.003

Graf Estes, K., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*(11), 1517–1528. https://doi.org/10.1037/a0039725

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*(2), 701–714. http://acousticalsociety.org/content/terms.

Kapadia, A. M., Tin, J. A. A., & Perrachione, T. K. (2023). Multiple sources of acoustic variation affect speech processing efficiency. *The Journal of the Acoustical Society of America*, *153*(1), 209. https://doi.org/10.1121/10.0016611

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698

Kreiman, J., Gerratt, B. R., & Gabelman, B. (2003). Jitter, shimmer, and noise in pathological voice quality perception. *The Journal of the Acoustical Society of America*, *112*, 2446–2446. https://doi.org/10.1121/1.4780067

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change*. Mouton de Gruyter.

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*(2), 241–246. https://doi.org/10.1016/j.cognition.2011.10.007

Lim, S. J., Shinn-Cunningham, B. G., & Perrachione, T. K. (2019). Effects of talker continuity and speech rate on auditory working memory. *Attention, Perception, and Psychophysics*, *81*(4), 1167–1177. https://doi.org/10.3758/s13414-019-01684-w

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. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Lukics, K. S., & Lukács, Á. (2021). Tracking statistical learning online: Word segmentation in a target detection task. *Acta Psychologica*, *215*. https://doi.org/10.1016/j.actpsy.2021.103271

Lukics, K. S., & Lukács, Á. (2022). Modality, presentation, domain and training effects in statistical learning. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-24951-7

Luthra, S. (2023). Why are listeners hindered by talker variability? In *Psychonomic Bulletin and Review*. Springer. https://doi.org/10.3758/s13423-023-02355-6

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, *83*, 1842–1860. https://doi.org/10.3758/s13414-020-02203-y/Published

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of Talker Variability on Recall of Spoken Word Lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 676–684.

Melville, L. S. (2024). *Beyond noise: The role of speaker variability on statistical learning*. University of British Columbia.

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*(1), 365–378.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P., Halverson, K., Ghosh, S. S., Christodoulou, J. A., & Gabrieli, J. D. E. (2016). Dysfunction of Rapid Neural Adaptation in Dyslexia. *Neuron*, *92*(6), 1383–1397. https://doi.org/10.1016/j.neuron.2016.11.020

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America*, *109*(6), 2988–2998. https://doi.org/10.1121/1.1370525

Quam, C., & Creel, S. C. (2021). Impacts of acoustic-phonetic variability on perceptual development for spoken language: A review. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(5). https://doi.org/10.1002/wcs.1558

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Rojas, S., Kefalianos, E., & Vogel, A. (2020). How Does Our Voice Change as We Age? A Systematic Review and Meta-Analysis of Acoustic and Perceptual Voice Data From Healthy Adults Over 50 Years of Age. *Journal of Speech, Language, and Hearing Research: JSLHR*, *63*(2), 533–551.

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

Saffran, J. R. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition*, *81*, 149–169. www.elsevier.com/locate/cognit

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science, New Series*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, *35*, 606–621.

Schiff, R., Ashkenazi, P., Kahta, S., & Sasson, A. (2021). Stimulus variation-based training enhances artificial grammar learning. *Acta Psychologica*, *214*. https://doi.org/10.1016/j.actpsy.2021.103252

Schultz, T. (2007). *Speaker Classification I: Fundamentals, Features, and Methods*. http://www.cs.cmu.edu/∼tanja

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker Variation Aids Young Infants' Phonotactic Learning. *Language Learning and Development*, *10*(4), 297–307. https://doi.org/10.1080/15475441.2013.858575

Selinker, L., & Baumgartner-Cohen, B. (1995). Multiple language acquisition: 'damn it, why can't i keep these two languages apart?' *Language, Culture and Curriculum*, *8*(2), 115–121. https://doi.org/10.1080/07908319509525195

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Snijders, T. M., Kooijman, V., Cutler, A., & Hagoort, P. (2007). Neurophysiological evidence of delayed segmentation in a foreign language. *Brain Research*, *1178*(1), 106–113. https://doi.org/10.1016/j.brainres.2007.07.080

Stilp, C. E., & Theodore, R. M. (2020). Talker normalization is mediated by structured indexical information. *Attention, Perception, and Psychophysics*, *82*(5), 2237–2243. https://doi.org/10.3758/s13414-020-01971-x

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*. https://doi.org/10.1186/1471-2202-10-21

Toro, J. M., Pons, F., Bion, R. A. H., & Sebastián-Gallés, N. (2011). The contribution of language-specific knowledge in the selection of statistically-coherent word candidates. *Journal of Memory and Language*, *64*(2), 171–180. https://doi.org/10.1016/j.jml.2010.11.005

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2). https://doi.org/10.1016/j.cognition.2005.01.006

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Werker, J., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, *1*(2), 197–234. https://doi.org/10.1207/s15473341lld0102_4

Yan, Q., Vaseghi, S., Rentzos, D., Ho, T.-H., & Turajlic, E. (2003). *Analysis of Acoustic Correlates of British, Australian, and American Accents*.