Stabilizing the production of nonnative consonant clusters with acoustic variability

Lisa Davidson[a] and Sean Martin
Department of Linguistics, New York University, 10 Washington Place, New York, New York 10003

Colin Wilson
Department of Cognitive Science, Johns Hopkins University, 237 Krieger Hall, 3400 North Charles Street, Baltimore, Maryland 21218

(Received 18 April 2014; revised 28 December 2014; accepted 7 January 2015)

Previous research on the perception, recognition, and learning of sounds and words has identified diverse effects of phonetic variation. The present study examined how variation affects cross-language production of consonant clusters. American English speakers shadowed words beginning with nonnative clusters in low- and high-variability conditions. Shadowing responses in the low-variability condition were quite sensitive to fine-grained phonetic properties that were manipulated across the stimuli. Notably, longer stop bursts led to increased rates of epenthesis, lower burst amplitudes resulted in more feature change and deletion, and intense periods of voicing at cluster onset elicited prothetic responses. Sensitivity to the acoustic manipulations was substantially attenuated in the high-variability condition, which combined stimuli from the first condition with baseline productions of the same items from two additional talkers. Detailed analyses of the response patterns indicate that more stable production targets in the high-variability condition resulted from integration, or blending, of the multiple talker stimuli. Implications of these findings for language-specific speech processing and the role of phonetic variability in second language acquisition are discussed.
[http://dx.doi.org/10.1121/1.4906264]

[CGC] Pages: 856–872

I. INTRODUCTION

Substantial phonetic variation is observed both within and across languages. Within a language, sources of variability include speaker physiology, speech context or register, speech rate, dialectal differences, sociolinguistic factors (e.g., Docherty, 2007), and speech affect (Barcroft and Sommers, 2005) among others. In the case of individual phonemes and phoneme sequences, which are the main focus of the present study, language internal phonetic variability can range from segmental changes (e.g., spirantization, glottalization, deletion) to lower-level phonetic variation, including durational shortening or lengthening, changes in amplitude, variability in the extent of phonetic voicing, and so on (Ernestus, 2012). Importantly, variation that falls within limits that are acceptable for a phoneme or sequence in one language can correspond to a phonological contrast in other languages. For example, voice onset time of voiceless stops in French can vary from at least 10 to 80 ms, but the same range spans two contrastive phonemes in Thai (unaspirated and aspirated voiceless stops) (Kessinger and Blumstein, 1997). Thus a central problem in acquiring the sound system of a language is to determine the range of permissible variation for each linguistic unit (such as a distinctive feature, phoneme, or word).

A proper understanding of phonetic variation may be especially challenging in the early stages of second language learning. It is widely accepted that adult learners' native phonological and phonetic systems strongly limit their ability to acquire new sound structures (e.g., Flege, 1995; Best and Tyler, 2007). Moreover, early L2 learners are unlikely to have sufficient orthographic, lexical, and syntactic knowledge of the new language for their interpretations of phonetic variability to be guided by appropriate higher-level structure (cf. well-known top-down effects in native language perception and production, e.g., Ganong, 1980). These considerations suggest that—at least upon first exposure to novel sound structures—the interpretation of fine-grained phonetic detail could pose particular problems. Listeners must somehow identify the same foreign category or sequence under diverse acoustic realizations (which furthermore can vary in their similarity to native structures).

Previous research, reviewed in the following, has established that adult listeners are highly sensitive to fine-grained phonetic detail in their native languages. Perhaps surprisingly, it has also been shown that acoustic variability can be beneficial for language learning by adults and children. The present study addresses the immediate effects of acoustic variation on cross-language speech production. In principle, within-category variation in the new language could result in highly *unstable* production patterns, with the nonnative speaker "over-interpreting" fine distinctions among stimuli as contrastive differences. Alternatively, it could *stabilize* productions around phonetic aspects of a nonnative structure that are systematic rather than idiosyncratic to particular realizations. The evidence presented here demonstrates that variability can lead to both outcomes for cross-language speech production, depending on the nature of the variation itself and the mode of exposure.

[a]Author to whom correspondence should be addressed. Electronic mail: lisa.davidson@nyu.edu

 0001-4966/2015/137(2)/856/17/$30.00

A. Sensitivity to non-contrastive detail in the native language

Research in a number of areas has demonstrated that listeners are sensitive to fine-grained phonetic detail and that such detail influences the representation and production of speech. Many perception studies have found processing costs associated with switching talkers, supporting the claim that talker-specific (indexical) properties are attended to and represented in some form by listeners (e.g., Mullennix *et al.*, 1989; Goldinger *et al.*, 1991; Sommers and Barcroft, 2006). For example, Mullennix *et al.* (1989) found that listeners were more accurate in identifying words presented in a single talker's voice than words produced by multiple talkers. Similar effects have been reported on recognition performance for vowel stimuli, with higher accuracy when successive vowels are produced by the same talker (Assmann *et al.*, 1982). Other research has shown that listeners perceptually adjust their phonetic categories to accommodate the idiosyncratic production patterns of particular speakers. For example, Kraljic and Samuel (2007) found that listeners shift their representation of /s/ to include phonetic detail that was more /ʃ/-like after exposure to a speaker who produced the alveolar fricative with more palatal characteristics.

In addition to these perceptual effects, the influence of phonetic detail has been found in research on phonetic convergence (or accommodation) in speech production. A number of shadowing tasks, in which participants must quickly repeat auditorily presented stimuli, demonstrate that repetitions are more perceptually similar to the shadowed stimulus than productions of the same word without a preceding auditory prompt (e.g., Goldinger, 1998; Nye and Fowler, 2003). Similar results have been found for speakers who engage in conversational tasks, such as navigating a map (Pardo, 2006). Some studies of imitation and shadowing have isolated the specific phonetic details of the stimulus that are mimicked, such as voice onset time (e.g., Shockley *et al.*, 2004; Nielsen, 2011).

The findings reviewed above highlight the fact that phonetic variability is salient enough that it influences listeners' perceptions and their subsequent productions. This suggests that phonetic variation could have a destabilizing effect on cross-language speech production. That is, nonnative listeners might be highly sensitive to fine-grained phonetic details in the input, attempt to mimic these details in their own productions, and perhaps even amplify variation by mapping diverse phonetic implementations of the same nonnative structure to categorically-different native representations. However, evidence from language learning discussed in the following section could lead to the opposite expectation: Namely, that phonetic variation should support greater stability and abstraction in nonnative perception and production.

B. Beneficial effects of variability in language learning

Across speakers, physiological, dialectal, and other idiosyncratic factors lead to individual differences in formant values and ratios, VOT durations, degree of vowel nasalization, and so on (e.g., Johnson and Mullennix, 1997). An important line of research has shown that the phonetic variability present in speech from multiple talkers is particularly effective in leading language learners to establish robust phonological representations, and we focus our review on this type of variation.

In a foundational study by Lively, Logan, and Pisoni (1993), Japanese speakers trained on stimuli produced by multiple talkers learned to make more accurate distinctions between English /ɹ/ and /l/ than speakers trained on stimuli from a single talker. The former set of participants also generalized better to utterances produced by a novel talker. In addition to more robust development of phonemes in L2 acquisition (see also Logan *et al.*, 1991; Bradlow *et al.*, 1997; Wang *et al.*, 1999; Iverson *et al.*, 2005), it has been shown that the acoustic variability present in multiple talker stimuli (sometimes referred to as high variability phonetic training) also leads to better L2 vocabulary learning (Barcroft and Sommers, 2005; Sommers and Barcroft, 2007, 2011). Benefits of acoustic variability have also been found in adult and infant phonetic and phonotactic learning (Krehm *et al.*, 2012; Seidl *et al.*, 2014), child lexical development (Richtsmeier *et al.*, 2009), and infant early word learning (Rost and McMurray, 2010).

Because the present experiment involves production of novel words, the findings of Barcroft and Sommers (2005) and Rost and McMurray (2010) are perhaps most relevant. In the Barcroft and Sommers study, simply increasing certain types of phonetic variability, while holding talker constant, did not lead to improved second language vocabulary learning by college-aged participants on measures such as latency and accuracy. However, presenting L2 learners with the same item produced by multiple talkers did improve vocabulary performance. A related pattern of results was found by Rost and McMurray in their study of early word learning by 14-month-old infants. Infants did not show evidence of recognizing single feature mismatches between novel objects and their labels (/buk/ and /puk/) when trained on productions from a single talker, even when the stimuli were manipulated to have considerable variation in the acoustic cues distinguishing the relevant sounds (i.e., VOT, burst amplitude, F0). In contrast, sensitivity to object-label mismatch was found for infants trained on productions from multiple talkers (naturally recorded, but acoustically manipulated to match the VOT ranges in the single talker condition). This suggests that the particular sort of phonetic variation that is characteristic of speech from multiple talkers may be more informative to learners than phonetic variability found within speech from a single talker (but cf. Galle *et al.*, 2015).

C. Sensitivity and stabilization in nonnative speech production

The main question addressed in this study is how phonetic variability affects adults' production of nonnative phonological structures, namely, initial obstruent-obstruent and obstruent-nasal consonant clusters, in a shadowing task. Previous work (Davidson, 2010; Wilson and Davidson, 2013) presented American English listeners with stimulus items (e.g., /bdafa/, /kpavo/, /gnatu/, and /zmasa/) produced

by one Russian speaker in each experiment. Close examination of the stimuli of Davidson (2010) showed that they contained within-talker phonetic variation that correlated with production accuracy and error type (Wilson and Davidson, 2013). Specifically, as stimulus stop burst duration or amplitude increased, speakers were significantly more likely to insert a vocalic transition (vocoid) between the two consonants of the sequence. Voicing also played a major role: Speakers produced an inserted vocoid significantly more often in voiced clusters than they did in voiceless clusters. Additionally, speakers were significantly more likely to produce a prothetic vocoid before the consonant sequence when there was evidence of pre-obstruent voicing (POV)—defined as higher-amplitude voicing at the onset of a stop or before the onset of the fricative, a natural phonetic implementation of voicing in obstruent clusters produced by Russian speakers.

The findings from these studies indicate that English participants, when presented with stimuli from one talker, are highly sensitive to phonetic details in their perception and production of nonnative structures. Given that the participants had no prior knowledge of Russian (or other languages with similar word-initial sequences), these findings illustrate the general problem identified at the outset: Early exposure to phonetic variation in a new language may lead to unstable and inaccurate productions. The variation found in the speech of one talker is evidently not sufficient to induce generalization (or abstraction) of consistent representations of the nonnative clusters on the part of the participants. Indeed, overall response accuracy and the proportions of various modification types remained essentially the same across blocks, indicating that little experiment-wide integration of phonetic information had occurred.

In the current study, we examined whether a high-variability paradigm could induce performance that is more stable than that in the low(er)-variability paradigm used previously. Two conditions were compared. The low-variability condition in the current study was identical to that of Davidson (2010) except that the acoustic-phonetic properties identified in the preceding text were systematically manipulated across trials. In this condition, participants heard productions from a single Russian talker which had been acoustically manipulated in a way that affected the duration and amplitude of stop bursts and the presence or absence of POV. In the high-variability condition, each trial contained productions of the same word by three Russian talkers. The first two presentations within a trial were *baseline* tokens of the target word (i.e., tokens with relatively short burst durations, natural burst amplitudes, and no pre-obstruent voicing). The last presentation in a trial was identical (same talker and acoustic manipulations) to the corresponding trial in the low-variability condition.

Recall that prior studies have found contrasting results with between-trial talker and phonetic variation with detrimental effects on recognition memory and categorization but beneficial effects for word learning. The effect of within-trial variation on speech perception and production has not to our knowledge been previously studied. The comparison of shadowing responses in our low- and high-variability conditions both extends research on phonetic variability in cross-language speech production and identifies distinct effects of this novel type of manipulation.

We hypothesized that presenting the same stimulus item with multiple acoustic parameters, and in multiple voices, would provide participants with key information about the range of phonetic variability that is acceptable for the target sequence, leading to more stable and accurate productions (i.e., lower sensitivity to the fine-grained detail of particular utterances). However, a range of other possible outcomes are conceivable, especially in light of the different effects of variability found in other tasks. As a way of framing our theoretical interpretation of the results, we consider three possible effects of including baseline productions from two talkers along with manipulated productions in the same trials.

1. Selection

One possibility is that speakers would shadow the fine phonetic details of only one of the stimuli. If speakers select the first (or second) stimulus as their production model, response patterns should be statistically indistinguishable from those of manipulated tokens with baseline values. If speakers alternatively select the last stimulus item to guide their productions, responses should reflect the trial-by-trial modifications of phonetic properties just as in the low-variability condition (and previous single-talker experiments). The latter outcome would suggest that within-trial variability of the type studied here is not sufficient to stabilize production targets.

2. Abstraction

The presence of acoustic variation within a trial may lead the shadower to encode the stimuli at a fairly abstract level (e.g., in terms of their perceived phonemic content or discrete gestural organization) and hence to suppress the influence of fine-grained details that differ across the stimuli. Under this possibility, it would be expected that participants would show little effect of the acoustic manipulations. For example, speakers should produce stimulus items beginning with stop-stop sequences primarily with a single type of response—perhaps correctly, perhaps with an epenthetic or prothetic vocoid—regardless of what acoustic manipulations, if any, the stimuli contain.

3. Blending

A third possibility is that speakers' productions would reflect a blend of the properties of all of the stimuli in a trial, perhaps with a preference for preserving phonemes/gestures that are perceived in at least one of the stimuli. A preservation preference has previously been proposed as a central principle of loanword adaptation (Paradis and LaCharité, 1997). Crucially, if responses reflect a blend of the phonetic cues available in the multitalker stimuli, these responses should still contain some evidence of sensitivity to the phonetic manipulations of the third stimulus in the trial (e.g., longer stop bursts, higher burst amplitudes, or presence of

pre-obstruent voicing), although the degree of sensitivity may be attenuated.

These hypotheses depend on an account of cross-language production in which perceptual interpretation of acoustic properties plays a pivotal role. For detailed discussion of such an account in terms of the process of phonetic decoding, see Wilson *et al.* (2014). In the next section, we lay out the details of the study aimed at investigating which of these three possibilities provides the best account of how within-trial phonetic variation influences the production of nonnative consonant clusters.

II. METHODS

A. Participants

The participants were 48 New York University graduate and undergraduate students. All were native speakers of American English ranging in age from 19 to 26 yr. Responses to a demographic questionnaire indicated that the participants spoke neither Slavic languages nor any other languages with initial obstruent clusters, such as Hebrew. No speakers were bilingual in any other language, although they had studied languages such as Spanish, French, or Mandarin in high school and college. None of the participants reported any speech or hearing impairments that persisted beyond 4 yr of age (one person remarked that they stuttered between ages 3 and 4). They were compensated $10 for their participation.

B. Stimuli

The critical stimuli were nonce words of the form CCáCV. The initial consonant clusters were composed of fricative-nasal (FN), fricative-stop (FS), stop-nasal (SN), and stop-stop (SS) sequences. The individual clusters used in this study are shown in Table I. Stop-initial clusters contained both voiced and voiceless consonants. For fricative-initial clusters, only voiced fricatives were included. Each cluster appeared in four distinct stimulus items for a total of 96 CC-stimuli (see Wilson *et al.*, 2014 for a full list of stimuli). Filler items were words of the form CəCáCV (48 items) and əCCáCV (48 items). To create the fillers, two of the four words for each initial cluster were chosen at random and the –áCV ending from those items was assigned to CəC-, and the remaining two –áCV endings were used to form the əCC- stimuli (e.g., for /pn/: /pnabu/, /pənabu/, /pnata/, /pənata/, /pnaso/, /əpnaso/, /pnave/, /əpnave/). All of the stimuli were recorded by three Russian-English bilingual, phonetically trained linguists who had no trouble producing the words with the appropriate stress pattern and with reduced vowels (schwas) for the fillers. All three of the Russian-English talkers are females born in Russia. Talker A, a graduate student at the time of recording, came to the United States for graduate school. Talker B, an English instructor in the United States who completed her Ph.D. in cognitive science, came to the U.S. for college, and talker C, who was in her final year of her linguistics Ph.D. program, came to the U.S. in elementary school. The talkers are highly proficient English speakers, but all report that they continue to speak Russian regularly.

TABLE I. Target consonant clusters used in the CCáCV stimuli.

Cluster type	Voiceless C1	Voiced C1
Stop-nasal	pn tm km kn	bn dm gm gn
Stop-stop	pt tp kp kt	bd db gb gd
Fricative-nasal		vm vn zm zn
Fricative-stop		vd vg zb zg

For one of the speakers (talker C), the acoustic properties under investigation were manipulated in the following ways for each of the cluster-initial stimuli. These stimulus items were exactly the same in the low- and high-variability conditions.

1. Pre-obstruent voicing

The first modification was POV, which we define as an interval of voicing that appears before or at the beginning of the formation of an obstruent constriction and that has visibly higher amplitude than voicing (if any) present during the subsequent constriction. In the case of fricative-initial clusters, POV precedes the onset of frication, as shown in Fig. 1(a). In all cases, POV contains low-frequency periodic energy but does not have visible formant structure. POV for stop-initial stimuli is illustrated in Fig. 1(b).

Versions with and without POV were created for each stimulus item beginning with a voiced obstruent. If a token was naturally produced with POV by the Russian speaker, the initial voicing interval was spliced out to create the non-POV version. For stimuli that lacked POV in the original recording, an interval of POV from the waveform of a different utterance of the same consonant was spliced in. All splices were taken at zero-crossings to avoid acoustic artifacts. This manipulation affected voiced SN, SS, FN, and FS stimuli. Because the stimuli took advantage of the POV naturally produced by the Russian speaker, there was some variability in its duration. The mean duration of the POV was 53 ms for stop-initial sequences (range: 30–80 ms) and 44 ms for fricative-initial sequences (range: 30–90 ms).

2. Burst duration

The second manipulated property was the burst duration of the first consonant in stop-initial stimuli. Two levels of burst duration were used: 20 and 50 ms. The shorter duration was created by splicing 5–10 ms out of the burst of the initial stop as necessary. Longer durations were created by copying 10–20 ms of the middle section of the burst (after the initial transient) and splicing that material back in. The burst duration manipulation affected voiced and voiceless SN and SS stimuli. Spectrograms illustrating this manipulation are shown in Figs. 2(a) and 2(b).

3. Burst amplitude

The third modification targeted the relative burst amplitude of stimulus-initial stops. Using PRAAT, we first calculated the amplitudes of the bursts relative to the following consonant (stop or nasal) for each recording of the SN and

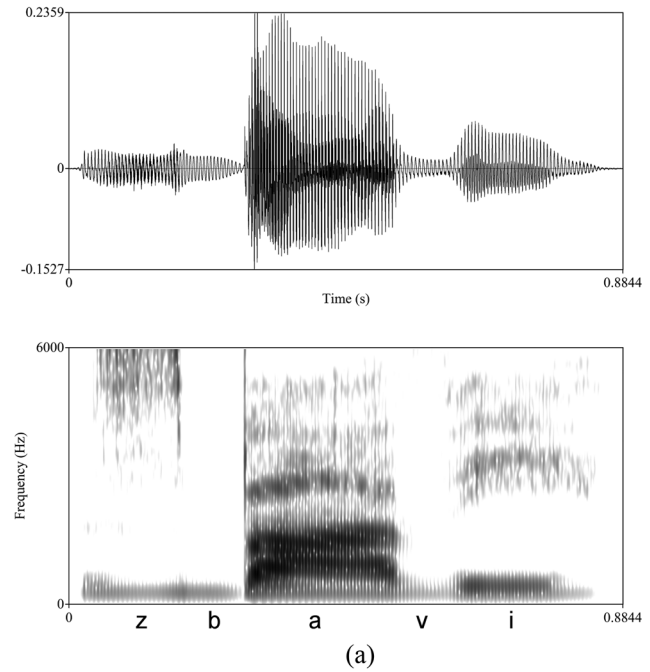

FIG. 1. (a) Example of fricative-initial stimulus /zbavi/ with pre-obstruent voicing (POV). (b) Example of stop-initial stimulus /gbake/ with POV.

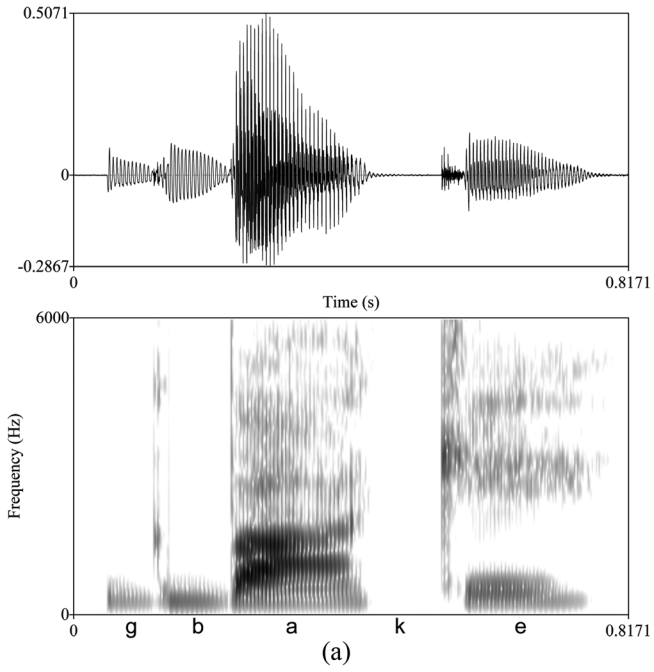

FIG. 2. Example of stop-initial stimulus /gbake/ with (a) short, low-amplitude burst and (b) long, high-amplitude burst.

SS stimuli by talker C. Because stop bursts naturally have lower relative amplitude before nasals than before other oral stops (see further discussion in Wilson *et al.*, 2014) and because amplitude also varies with the voicing specification of the stop, the values for this manipulation were determined for each cluster type separately. For SN clusters, the low-amplitude versions had values based on the means of the corresponding natural productions (voiceless SN: −18 dB, voiced SN: −7 dB), and high-amplitude versions were raised several decibels above the means (voiceless SN: −10 dB, voiced SN: 0 dB). The direction of manipulation was reversed for SS clusters: The high-amplitude versions mirrored the natural means (voiceless SS: 23 dB, voiced SS: 0 dB), while the low-amplitude versions were reduced in amplitude (voiceless SS: +13 dB, voiced SS: −7 dB). The manipulated values were carefully chosen to ensure that all bursts (in particular, those of stops with lowered amplitude) were audible and sounded intelligible. Illustrations of high- and low-amplitude bursts in stop-initial stimuli are shown in Figs. 3(a) and 3(b).

These manipulations were crossed where possible as summarized in Table II. Together all of the manipulations (448) plus the fillers (96) (which were not modified except to normalize the amplitude of all of the stimuli to 67 dB) came

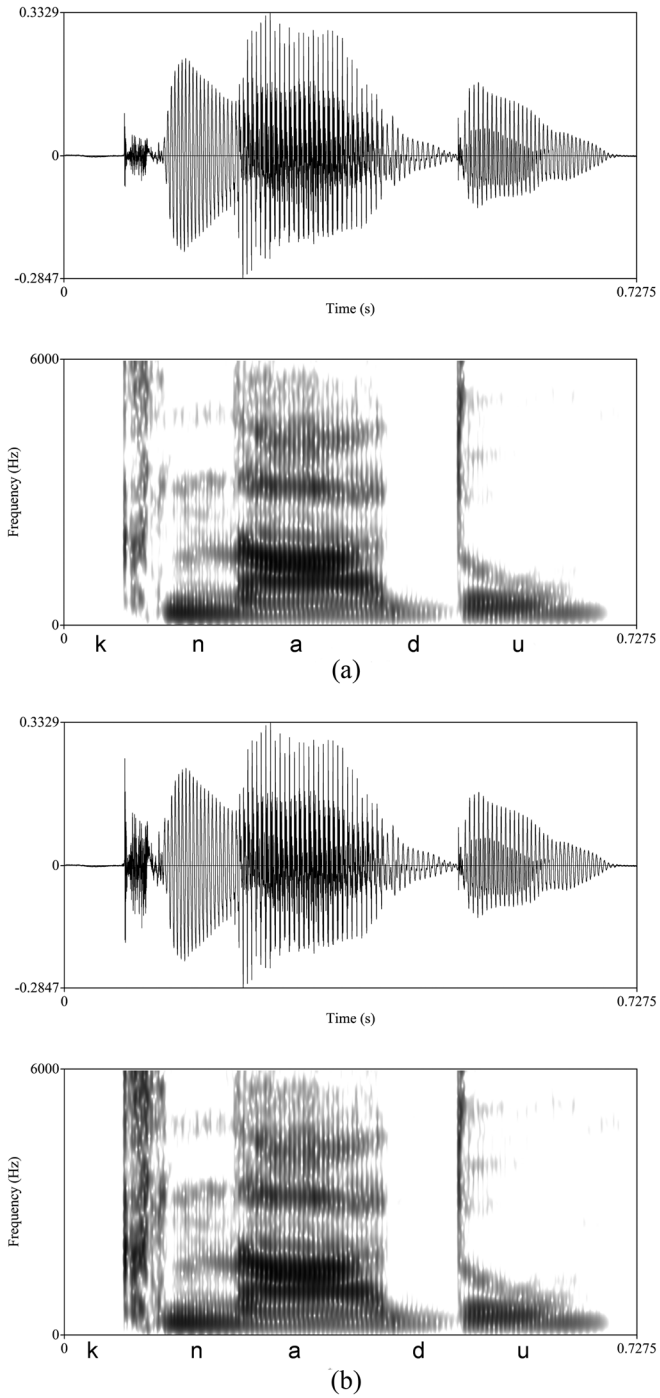

FIG. 3. Example of stop-initial stimulus /knadu/ with (a) long, low-amplitude burst and (b) long, high-amplitude burst.

to a total of 544 sound files. To create an experimental procedure that would not be too taxing for the participants, 12 counterbalanced lists were created containing 224 stimuli each. Each list was composed of 32 FN, 32 FS, 32 SN, 32 SS (half voiced, half voiceless), 48 CəC fillers, and 48 əCC fillers. The manipulations were distributed across the experimental lists so that each list contained approximately the same number of each manipulation type (within each cluster type), which occurred equally often across the lists. Two participants were assigned to each list.

TABLE II. Summary of acoustic manipulations in cluster-initial stimuli.

Cluster type	Crossed acoustic manipulations
Fricative-initial	POV (present vs absent)
Voiceless-stop initial	DUR (20 ms, 50 ms) × AMP (high vs low)
Voiced stop-initial	DUR (20 ms, 50 ms) × AMP (high vs low)
	DUR (20 ms, 50 ms) × POV (present vs absent)

The phonetic properties of stimuli for the other two talkers (talkers A and B) were set to baseline values. First, POV was absent from voiced stops and fricatives. Second, following the same splicing procedure described in the preceding text, stop burst durations were adjusted to match the means of each talker's original productions (talker A: 30 ms, talker B: 26 ms). Last, for the sake of simplicity, SN clusters were set to the lower relative amplitude values calculated for talker C, and SS clusters were set to the higher relative amplitude values. These characteristics were intended to reflect natural productions of the studied clusters. Note that when we refer to "manipulations" in the following discussion, we always mean those applied to stimuli from talker C not the normalization of stimuli from the other two talkers. The measurements for the average burst durations and amplitudes and proportion of POV in the stimuli as they were naturally recorded by the Russian speakers and before any acoustic manipulations are available on the first author's website (https://files.nyu.edu/ld43/public/publications.html).

In the high-variability condition, the stimulus from talker C was always presented last in each trial with the order of talkers A and B counterbalanced across trials. Given well-known limitations of auditory short term memory (e.g., Pisoni, 1973), presenting talker C's stimulus last should maximize the effect of the phonetic manipulations on participants' responses. Differences between the response patterns found for high- and low-variability conditions would therefore indicate abstraction or blending across the three stimuli in a trial.[1]

C. Procedure

The participants were seated in a sound-attenuated room with the computer that was used to present the stimuli using EPRIME 1.1 (Psychology Software Tools, Pittsburgh, PA). Half of the participants (24) participated in the low-variability condition.[2] These participants heard two consecutive repetitions of the acoustically manipulated item produced by talker C before giving their response. The other half of the participants (24) participated in the high-variability condition. For these participants, three unique versions of each stimulus item were presented before the response (either talker A ∼ talker B ∼ talker C or talker B ∼ talker A ∼ talker C).

Participants were told that they would hear either two (in the low-variability condition) or three (in the high-variability condition) repetitions of the same word and that they should repeat what they had heard into the microphone after all of the words had played.[3] The interstimulus intervals were 450 ms, and participants were given 1.5 s after the final stimulus to respond before the program automatically moved on to the next item. The 224 items were randomly divided into three blocks to provide the participants chances to rest. Item order within block was randomized separately for each

participant. The spoken responses were recorded with an Audio-Technica ATM-75 head-mounted condenser microphone onto a Zoom H4n digital recorder. The WAV files were recorded at 44.1 kHz, 16 bits. The experiment began with six practice trials containing different clusters than those used in the study.

D. Data analysis

Coding of the data followed the same procedure as in previous studies (e.g., Davidson, 2010). All production responses were analyzed by repeated listening and examination of waveforms and spectrograms in PRAAT. Modifications of a consonant cluster relative to the native Russian speakers' productions were labeled as shown in Table III. If multiple errors occurred, each error was labeled, and if none of the errors found in Table III occurred, the token was labeled as "no modification" (correct). A token was coded for epenthesis if it had vocalic material containing visible first and second formants, occurring after frication or stop release and before the onset of the following stop, fricative, or nasal consonant.[4] To be coded for prothesis, a response had to have a vocalic element containing first and second formants before the initial obstruent; voicing during stop closure, or voicing that started before, were not counted as errors because these properties are found in natural Russian productions of the target clusters. More generally, to be coded as correct, participants' utterances had to match the manner, place, and voice specifications of the input, and the consonants had to be produced in the correct linear order as determined using the spectrogram. Coding of errors was conservative: Small variations from the target stimulus, such as in the duration of a consonant or a burst, did not prevent the token from being coded as correct.

The responses were coded by three research assistants and two of the authors (LD and SM). All coding was done blindly; that is, the coders did not know what manipulations were present in talker C's utterances. All of the responses were then discussed by at least two different research assistants and the authors in regular lab meetings to ensure that coders agreed on the labels assigned to each of the responses.

III. RESULTS

The effects of phonetic and other factors on the distribution of coded responses were assessed with Bayesian generalized linear mixed-effects models (Gelman and Hill, 2006; Kruschke, 2011). In particular, multinomial (polytomous) logistic regression models were fit to the production responses as is appropriate when responses are drawn from an unordered set of categories (as in Table I, Raudenbush and Bryk, 2002). Responses to filler items, and responses to critical items that were coded as "other" (<3%), were removed prior to analysis. The small proportion of remaining responses with multiple modifications (<5%) were coded as such (i.e., the dependent variable was a matrix with one column for each modification type, with 1 indicating that the modification was present in a response and 0 indicating its absence). The correct (no modification) response type was treated as the baseline for the dependent variable.

The main factors of interest were variability condition (low vs high within-trial variability) and those associated with the acoustic manipulations in talker C stimuli (high vs low relative burst amplitude, long vs short duration, presence vs absence of POV). We also included factors for cluster type (e.g., SS and SN) and cluster voice (voiced vs voiceless) as appropriate for each data subset. Multinomial models generally include a response type factor with one level for each of the non-baseline response types (here *epen* = epenthesis, *proth* = prothesis, *chng* = C1 change, *del* = C1 deletion) that is crossed with the other predictors. For example, a significant two-way interaction of the form *epenthesis* × *variability* would indicate that the rate of epenthesis responses differed across variability conditions, and a significant three-way interaction of the form *epenthesis* × *variability* × *amplitude* would indicate that the effect of amplitude on epenthesis rate differed by condition. Note that variability condition is a between-participant factor, while the other factors varied within participant.

All binary predictors were effect (sum-to-zero) coded and scaled to have a mean of zero and a difference in upper and lower values of one (Gelman *et al.*, 2013). Thus a change from one value of a binary factor to the other, holding all other terms constant, corresponds to an expected change in the log-odds of a response equal to the coefficient value. In addition to the fixed structure, the models also included crossed random effects allowing the probabilities of the response types to vary by participants and stimulus items. Random slopes were also included as permitted by the experimental design (e.g., the random effect for item included variability condition and the acoustic manipulations but not cluster type).

Analyses were performed with Markov chain Monte Carlo (MCMC) sampling as implemented by the MCMCglmm package (Hadfield, 2010) in R (R Development Core Team, 2012). This procedure samples coefficients from the posterior probability distribution conditioned on the data and the

TABLE III. Response codes for cluster-initial stimuli.

Response type	Definition	Example
Epenthesis	Target is produced with vocalic material between the consonants	/pkadi/ → [ᵊkadi]
C1 Deletion	Target is produced with the first consonant deleted	/pkadi/ → [kadi]
Prothesis	Target is produced with vocalic material before the cluster	/pkadi/ → [ᵊpkadi]
C1 Change	Target is produced with two segments, but C1 differs from the original	/pkadi/ → [skadi]
Other	Target is not produced or has an error other than the ones listed above or has more than two errors	/pkadi/ → Ø
		/pkadi/ → [kpadi]
		/pkadi/ → [spaga]

model's prior. We used a prior and other settings that are standard for mixed-effects multinomial models (Hadfield, 2010). Statistical significance was assessed with 95% highest posterior density (HPD) intervals as computed by applying the coda package (Plummer *et al.*, 2006) to the output of MCMCglmm. We report the mean value of each coefficient as a point estimate and an associated *p*-value (determined from the proportion of posterior samples that lie on the same side of zero as the point estimate). Coefficients are not reported for factors that do not reach significance. In cases where further investigation of significant effects was warranted, MCMCglmm was also used to carry out binomial logistic regressions with fixed factor coding and crossed random effect structures specified as in the multinomial analyses.

A. Analysis of fricative-initial stimuli

Because the fricative- and stop-initial clusters were subject to different phonetic manipulations, we analyzed these two stimulus types separately. For fricative-initial stimuli, the only manipulation under investigation was POV. Previous results indicated that the presence of POV increases the probability of prothesis responses in low-variability trials with modal voicing preceding the onset of the fricative being reinterpreted as vocalic (Wilson *et al.*, 2014). If presenting nonnative speakers with multiple renditions of a stimulus item results in more faithful encoding of the intended consonant cluster, this would be reflected in an overall effect of variability condition such that all modifications are rarer in the high-variability condition. If the high-variability condition simply suppresses the effect of phonetic manipulations (because the other two stimuli in a trial are always baseline tokens), a significant interaction between condition and POV would be the expected outcome.

The model for fricative-initial clusters included cluster type (FS vs FN) and POV (present vs absent) as fixed factors that were independently crossed with response type and variability condition (low- vs high-variability) [i.e., the entire fixed structure was response type × variability condition × (cluster type + POV)].

Figure 4 compares response proportions across levels of the fixed factors. It is evident that no-modification is the most prevalent response, and accordingly there were significant negative coefficients for all modifications ($epen = -3.17$, $proth = -1.90$, $del = -4.98$, $chng = -1.95$, all p's < 0.001). Cluster type had a significant overall effect on the probability of prothesis: Prothesis was less likely for FN clusters than for FS clusters ($proth \times clus.type = -0.54$, $p < 0.05$). Presence of POV significantly increased the probability of prothesis and reduced the probability of deletion, relative to trials in which POV was absent ($proth \times pov = 0.68$, $del \times pov = -0.72$, both p's < 0.01).

Variability condition had two main types of significant effect. First, all modifications were less probable overall in the high-variability condition than in the low-variability condition ($epen \times variability = -2.23$, $proth \times variability = -1.45$, $del \times variability = -2.06$, $chng \times variability = -1.31$, all p's < 0.05). Indeed, while 50% of the responses in the low-variability condition contained some modification, this

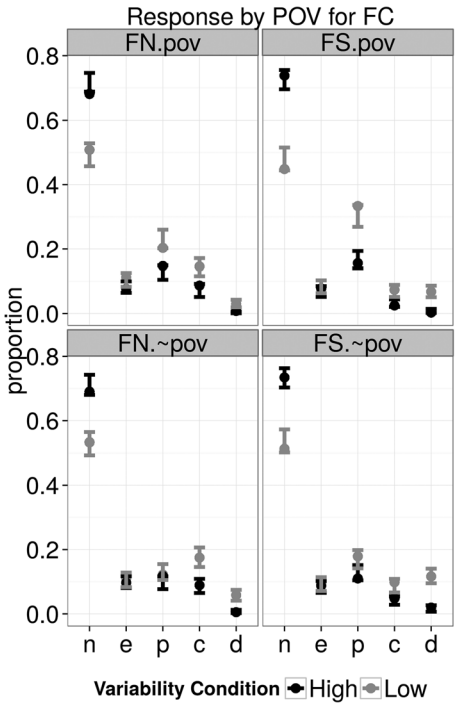

FIG. 4. Results for the POV manipulation for fricative-initial (FC) clusters. FN, fricative-nasal; FS, fricative stop; pov, POV is present; ~pov, POV is not present. For this and following figures, n, no modification; e, epenthesis; p, prothesis; c, C1 change; d, deletion.

decreased to just 29% in the high-variability condition. Second, there was a three-way interaction indicating that the effect of cluster type on prothesis probability varies across variability conditions ($proth \times variability \times clus.type = 0.81$, $p < 0.05$). A mixed-effects binary logistic regression of prothesis against cluster type was performed within each variability condition to better understand this interaction. The effect of cluster type on prothesis was significant in the low-variability condition (FN vs FS $= -0.30$, $p < 0.05$) but not in the high-variability condition (FN vs FS $= -0.04$ n.s.).

Because it is evident from Fig. 4 that prothesis remains the most common modification when POV is present even in the high-variability condition, we performed a mixed-effects binary logistic regression of prothesis against POV within each variability condition to confirm that the rates of prothesis were still significantly affected by the presence of POV. POV significantly increased the probability of prothesis both in the low-variability condition ($proth = 0.95$, $p < 0.01$) and, critically, in the high-variability condition ($proth = 0.45$, $p < 0.01$).

1. Interim summary

The results for the POV manipulation in fricative-initial clusters demonstrate that prothesis responses are more probable when POV is present, the expected effect of this phonetic manipulation. Moreover, the lower rate of deletion can be attributed to the fact that POV served as a cue to the presence of the fricative at the beginning of the cluster. The comparison between the variability conditions indicates that in the high-variability condition, the POV manipulation in the third stimulus has a weakened but nevertheless significant influence on the modification in the response.

B. Analysis of stop-initial stimuli

Stop-initial clusters were subject to POV, amplitude, and burst duration manipulations. Recall that POV and the amplitude manipulation were separately crossed with two levels of burst duration. We fit an initial model that included the phonetic manipulations, cluster type, and cluster voice as separate fixed factors, and otherwise had the same fixed and random structure as the model for fricative-initial clusters. This model revealed a number of significant interactions between variability condition and cluster type. In light of these interactions, and because of differences in the direction of the amplitude manipulation for SN and SS clusters (with amplitude raised from baseline for SN and lowered for SS), the two types of cluster are analyzed separately in the following text.

1. Stop-nasal stimuli

In both variability conditions, deletion and prothesis modifications each accounted for fewer than 5% of responses. These response types were excluded from the analysis to avoid well-known instabilities in logistic regression due to very small (or zero) counts for some factor combinations. They are, however, retained in the figures in the following text for completeness. The analysis proceeded as in the preceding text for the remaining response types (no-modification, epenthesis, change), with *nomod* as the reference level of the response type factor.

Overall, change responses were less probable than no-modification ($chng = -2.99$, $p < 0.01$), but epenthesis and no-modification did not differ significantly ($epen = 0.18$ n.s.). As is apparent from Fig. 5, epenthesis was particularly prevalent for clusters beginning with voiced (as opposed to voiceless) stops ($epen \times voice = 1.84$, $p < 0.01$). In addition, longer burst durations decreased the probability of C1 change in comparison to those with lower duration ($chng \times dur = -0.75$, $p < 0.05$).

Variability condition had two related effects. Epenthesis was less probable overall in the high-variability condition than in the low-variability condition ($epen \times variability = -1.15$, $p < 0.05$). Furthermore, variability condition modulated the effect of burst duration on epenthesis rate ($epen \times variability \times dur = -0.67$, $p < 0.05$). Mixed-effects binary logistic regressions of epenthesis against burst duration established that higher burst duration increased the probability of epenthesis in the low-variability condition (higher vs lower duration $= 0.34$, $p < 0.01$) but not in the high-variability condition (higher vs lower duration $= 0.00$ n.s.). Indeed the response percentages (and raw counts) of epenthesis differed substantially across burst duration levels in the low-variability condition—61% (230) for longer duration vs 49% (187) for shorter duration—but were nearly identical across burst duration levels in the high-variability condition—37.5% (141) for high duration and 37% (140) for low duration.

2. Stop-stop stimuli

In the SS subset, all modification types (and no modification) exceeded 5% of total responses in at least one of the variability conditions and were retained in the analysis.

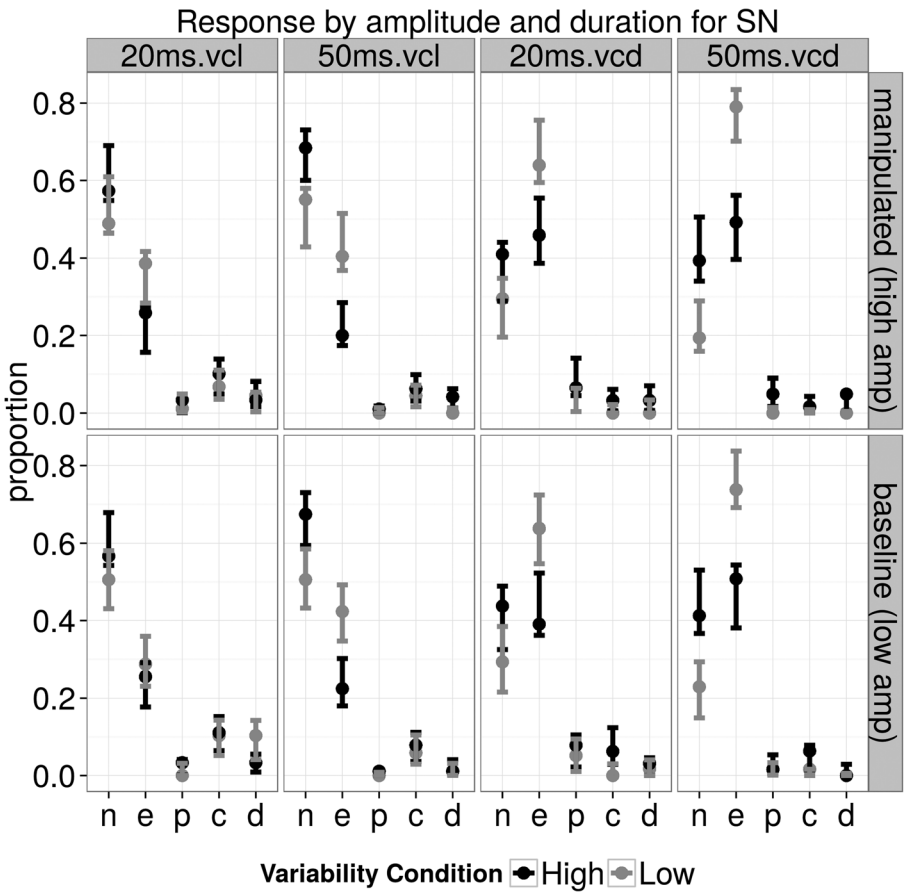

FIG. 5. Results for the amplitude and duration manipulations for stop-nasal clusters. 20 ms, short duration; 50 ms, long duration; vcl, voiceless; vcd, voiced. See Fig. 4 for key to modifications.

Apart from this difference, the present model was identical in structure to that for the SN data.

As shown in Fig. 6, overall, modifications except epenthesis were significantly less probable than no-modification ($proth = -4.24$, $del = -3.90$, $chng = -3.00$, all p's < 0.001). Voiced clusters had significantly higher probability of undergoing epenthesis, prothesis, and change modifications than voiceless clusters ($epen \times voice = 2.46$, $proth \times voice = 2.14$, $chng \times voice = 1.24$, all p's < 0.01). The $chng \times voice$ interaction accounts for the finding that voiced clusters are more likely to undergo change than deletion in comparison to voiceless stops. Presumably this is because voiceless stop clusters have only C1's burst as a cue to a cluster-initial stop, whereas in voiced stop clusters, voicing during C1's closure can serve as a cue to the presence of a cluster-initial stop; this difference is most evident at low amplitude, where voiced stops delete at a lower rate than voiceless.

There were three significant overall effects of the acoustic manipulations. Higher burst amplitude significantly increased the probability of epenthesis ($epen \times amp = 0.45$, $p < 0.05$) and also significantly lowered the probability of deletion ($del \times amp = -2.13$, $p < 0.01$). Prothesis modifications were more probable when POV was present ($proth \times pov = 1.59$, $p < 0.05$). (There were additional marginal effects indicating that higher burst duration may increase the probability of epenthesis, $epen \times dur = 0.32$ and that higher burst amplitude reduces the probability of change, $chng \times amp = -0.70$, both p's $= 0.07$.) Although there was no significant interaction for $epen \times amp \times variability$, Fig. 6 indicates that epenthesis responses in the high-variability condition outnumber epenthesis responses in the low-variability condition for stimuli with low burst amplitude. This suggest that participants can use the information in the burst from talkers A and B to "fill in" the degraded information in the low-amplitude bursts; the prevalence of the epenthesis response is further discussed in Sec. IV B.

As was found for fricative-initial and SN clusters, variability condition significantly modulated the probabilities of certain modifications. In particular, deletion and change were more probable in the high-variability than in the low-variability condition ($del \times variability = -1.99$, $chng \times variability = -1.97$, both p's < 0.001). Understanding these significant effects requires consideration of the conditions under which deletion and change are found in the low-variability condition. As indicated by the significant $del \times amp$ effect in the preceding text, 76% of the deletions in the low-variability condition are responses to low-amplitude bursts; the significant $del \times variability$ effect makes deletion responses highly improbable overall in the high-variability condition, effectively nullifying the effect of low burst amplitude found in the other condition. Essentially the same points apply to change modifications, which were also concentrated on stimuli with low-amplitude bursts in the low-variability condition (in particular, change was found in 13.8% of responses to low-amplitude bursts, but in only 7.3% of responses to high-amplitude bursts, in the low-variability condition). The significant $chng \times variability$ effect

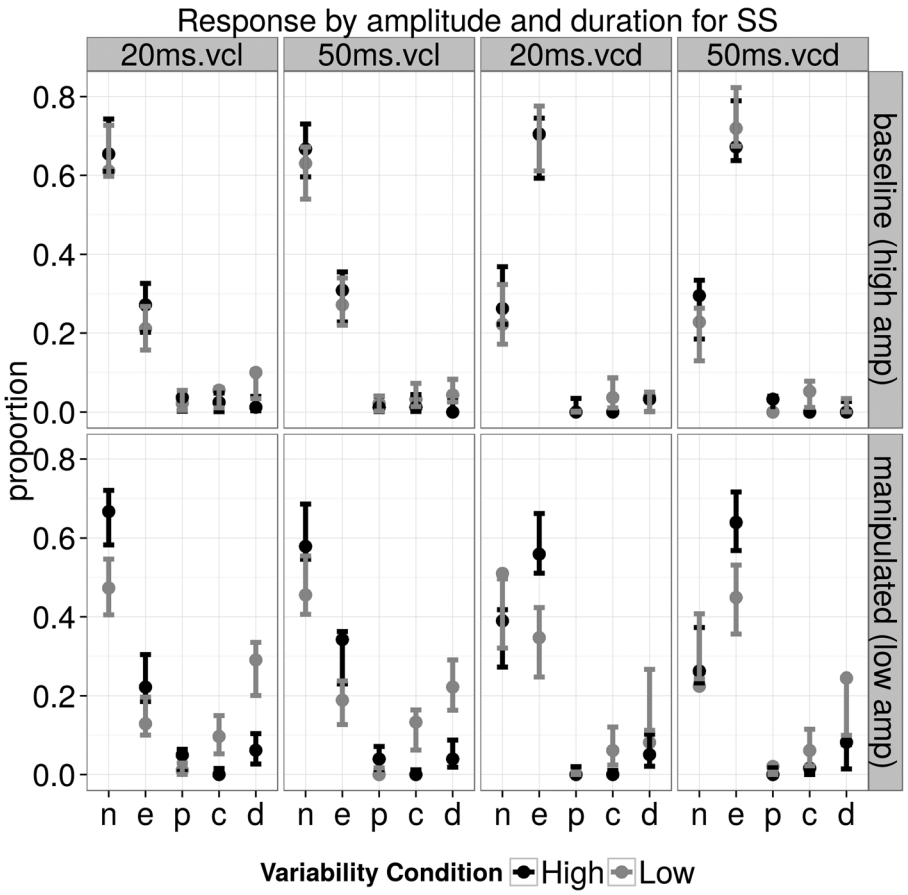

FIG. 6. Results for the amplitude and duration manipulations for stop-stop clusters. See Fig. 4 for key to modifications.

therefore serves to (partly) counteract the amplitude manipulation in the high-variability condition.

To further explore this pattern, we performed separate binary logistic regressions for deletion and change with the fixed effect of variability condition crossed with burst amplitude. Deletion was significantly more probable in the low-variability condition ($variability = 0.92$, $p < 0.001$), and less probable when burst amplitude was higher ($amp = -1.09$, $p < 0.001$), but these effects did not interact significantly. Change was significantly more probable in the low-variability condition ($variability = 0.84$, $p < 0.001$) but did not show strong effects of amplitude or any interaction.

Variability condition also participated in a significant three-way interaction involving prothesis. Specifically, the effect of cluster voice on prothesis was modulated by variability condition ($proth \times variability \times voice = -2.35$, $p < 0.05$). This interaction was investigated with separate mixed-effects binary logistic regressions, one for each variability condition, of prothesis against cluster voice. Whereas the probability of prothesis increased for voiced clusters in the low-variability condition ($voice = 1.12$, $p < 0.05$), no such effect was found in the high-variability condition ($voice = -0.25$, $p > 0.4$). Because the POV manipulation applies only to voiced clusters, this interaction accounts for the observation that prothesis was rarer for tokens with POV in the high-variability condition (4% of responses) in comparison to the low-variability condition (16% of responses) (see Fig. 7).

3. Interim summary

The results for SN clusters reflect a number of findings. First, there were more epenthesis responses for voiced clusters as opposed to voiceless clusters. This was true for both low- and high-variability conditions, although while the difference between no modification and epenthesis responses is significant for voiced stops in the low-variability condition, it was not significant in the high-variability condition.

Second, whereas there was a significant effect of the duration manipulation in the low-variability condition with respect to epenthesis for SN sequences, this effect disappeared in the high-variability condition. That is, only in the low-variability condition did longer bursts lead to more epenthesis. The results for duration provide clear support for the hypothesis, raised in the introduction, that the high-variability mode of presentation *stabilizes* production patterns in the face of acoustic manipulations. The fact that epenthesis rate was lower in the high-variability condition (with no commensurate increase in another modification types) converges with the finding from fricative-initial clusters that this condition results in not only more stable but also more accurate productions.

The other two manipulations, POV and amplitude, did not lead to substantial changes in speakers' responses relative to the baseline tokens in either the low- or high-variability conditions. Amplitude was not significant either as a main factor or in interaction with other factors for SN clusters. As for POV, given that prothesis was such an infrequent response type in both variability conditions, it is evident that this manipulation had little effect on speakers' productions.

The effect of voicing on production of SS clusters is similar to that for SN clusters, as epenthesis modifications were more frequent for the voiced clusters and correct responses (no-modification) were more common for the voiceless clusters. The duration manipulation had a marginal effect for the SS cluster type with epenthesis increasing slightly for 50 ms bursts in comparison to 20 ms bursts. There was a small effect of POV, with slightly more prothetic responses occurring when POV was present, but this was limited to the low-variability condition. However, as was the case for SN clusters, the overall proportion of prothesis responses is low, suggesting that speakers are especially sensitive to the acoustic characteristics of the stop burst, which is made salient by the surrounding closures.

Finally, the effects of the amplitude manipulation, which involved lowering burst amplitude for SS clusters, were mainly evident in the increased proportion of deletion and C1 change modifications in the low-variability condition. As expected, a burst that is atypically low in amplitude can lead listeners to misperceive the place of the stop or even completely fail to encode the stop. The effects for deletion and C1 change are strongest in the voiceless sequences. For stop initial sequences, the patterns seen for the C1 change and deletion responses illustrate a general issue in (nonnative) cue parsing that was also observed for fricative-initial clusters: As the strength of a cue such as burst amplitude or POV increases, segment deletion and change become less likely, but interpretation of the cue as an independent segment or gesture becomes more probable, especially when there is low phonetic variability in the stimulus set.

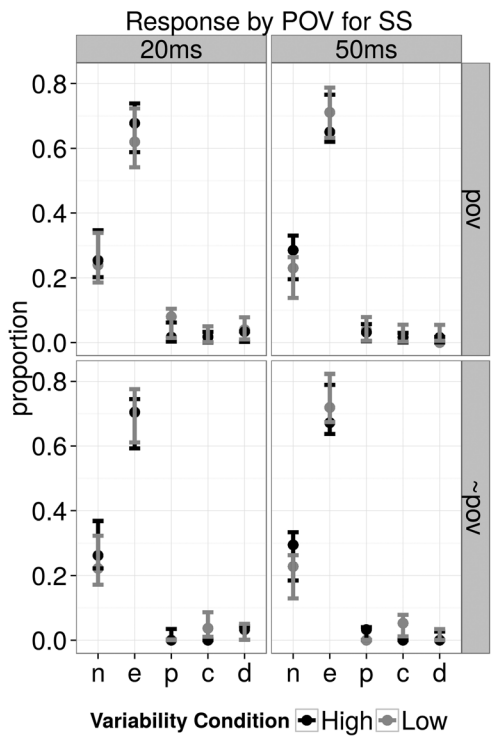

FIG. 7. Results for the POV manipulation for stop-stop clusters. pov, POV is present; ~pov, POV is not present. See Fig. 4 for key to modifications.

C. Implementation of burst duration

In addition to the categorical modifications reported in Sec. II B, we also examined burst durations in no-modification responses to stop-initial clusters. This purpose of this analysis was to determine whether the acoustic manipulations were reflected in phonetic implementation even in the absence of categorical modifications. Descriptive statistics for the relevant stop bursts in both conditions are given in Table IV. For voiceless stops, speakers produced longer burst durations for the 50 ms stimuli than for the 20 ms stimuli in both conditions; note that the stimulus durations were not perfectly mimicked (see also Cole and Shattuck-Hufnagel, 2011; Nielsen, 2011 on imperfect mimicry). Voiced stops did not vary according to stimulus burst duration, presumably because the range of acceptable variation in burst duration for voiced stops is much smaller than for voiceless stops in English.

To assess these results statistically, and to compare them with the results of the low-variability condition, a linear mixed-effects model with maximal random effect structure for participants and items was fit to the response burst durations with MCMCglmm. Fixed effects of variability condition, cluster voice, and stimulus burst duration were effect coded and fully interacted in the analysis. There were significant main effects of cluster voice ($voice = -8.8$, $p < 0.001$), with voiced stops having shorter bursts overall, and of stimulus burst duration ($dur = 1.49$, $p < 0.001$), indicating a significant positive effect of the acoustic manipulation. Cluster voice and stimulus burst duration interacted significantly ($voice \times dur = -0.83$, $p < 0.05$), and the three-way interaction of all fixed factors was also significant ($variability \times voice \times dur = 0.80$, $p < 0.05$). The former interaction accounts for the absence of a strong stimulus-driven effect on burst duration for voiced stops. The latter interaction suggests that the phonetic imitation of the manipulation in the final stimulus is weaker, though not absent, in the high-variability condition relative to the low-variability condition. This in turn provides further support for the blending hypothesis, which predicts that phonetic targets should be a (weighted) average of all stimuli within a trial.

IV. DISCUSSION

The aim of this study was to investigate whether providing participants with phonetic variability (here, mainly in the form of multiple talkers) would decrease their sensitivity to non-contrastive phonetic detail in nonnative phonotactic sequences. Results showed that this goal was largely achieved. In the low-variability condition, speakers' responses changed across the acoustic manipulations. When pre-obstruent voicing was present, speakers produced more prothesis, though the effect is stronger for fricative-initial sequences than for stop-initial ones. Longer stimulus burst duration led to greater rates of epenthesis responses for SN clusters and marginally for SS clusters. The amplitude manipulation had different forms and effects for SN and SS sequences. For SN sequences, the burst amplitude was raised from the baseline to determine if a greater intensity contrast between the burst and the following nasal results in more accurate overall performance [cf. perception results in Davidson and Shaw (2012) showing that English listeners tended to confuse SN sequences with C1 change and deletion alternatives]. This manipulation had essentially no effect. For SS sequences, the burst amplitude was lowered to provide a counterpart to the naturally lower relative amplitude of SN. Under this manipulation, the epenthesis modification decreased and the deletion and C1 change (for voiceless bursts) increased.

The effects of the manipulations were attenuated or even eliminated in the high-variability condition. Presence of POV still did lead to elevated rates of prothesis for the fricative-initial sequences, but effects of the duration and amplitude manipulations were mainly eliminated for stop-initial sequences. These findings suggest that speakers were integrating information from all three stimuli in the high-variability condition to determine their production targets. This within-trial integration had a stabilizing effect on responses, which were overall less sensitive to the acoustic manipulations in the final talker's stimuli. We now return to the specific nature of the integration observed in the high-variability condition.

A. Selection, abstraction, or blending?

In Sec. I, three possible outcomes for the high-variability condition were presented. The first potential outcome was *selection*: Speakers could shadow the fine phonetic details of only one of the stimuli. The second was *abstraction*: The presence of multiple acoustic stimuli could lead the participant to encode the stimuli at only an abstract (e.g., phonemic or gestural) level and therefore to consistently produce one type of modification (or possibly no modification) for a particular nonnative cluster. The final possibility was *blending*: Speakers' productions could reflect a combination of the acoustic properties of all of the stimuli within a trial with the result that sensitivity to the phonetic manipulations would be present but weaker than in the low-variability condition.

Taken together, the patterns in the data are best accounted for by the blending hypothesis. The particular type of blending we observe, as in previous studies of loanword adaptation, has a preference for preserving all of the

TABLE IV. Burst durations for English participants for responses coded as "no-modification" in the categorical coding.

Variability condition	Voicing	Duration for 20 ms stimuli	Duration for 50 ms stimuli
Low	Voiceless	36.5 ms (17.8)	50.2 ms (23.8)
	Voiced	25.8 ms (13.6)	29.8 ms (15.6)
High	Voiceless	37.8 ms (15.9)	45.3 ms (18.4)
	Voiced	25.6 ms (11.8)	25.6 ms (11.5)

phonemes (or gestures) that are perceived in the stimuli. The clearest indication of blending is found for the POV manipulation: Presence of POV resulted in greater prothesis rates in both the low- and high-variability condition. A potentially problematic finding is that the effects of the burst duration and amplitude manipulations mainly seem to have been eliminated, rather than simply attenuated, with respect the distribution of categorical response types. At first glance, this seems to support abstraction (or selection) because speakers produced relatively uniform proportions of correct utterances and essentially only one modification—epenthesis—regardless of the manipulations. However, analysis of the English speakers' burst durations for responses coded as "no modification" were found to reflect the acoustic manipulation of the third stimulus item in the high-variability condition. Minimally, this establishes that participants in the present experiment did not systematically ignore the final item of each trial, ruling out selection of the first (or second) item as a viable account. It converges with the hypothesis that participants determined their phonetic targets through a process of blending (i.e., weighted averaging) of within-trial stimulus properties.

Thus the effect of stimulus burst duration on speakers' phonetic realizations, in combination with the attenuation (but not disappearance) of the POV effect, is most consistent with the blending account. Speakers still show some sensitivity to the phonetic details of the manipulations, but either the proportion of tokens that are produced with a modification decreases or the speaker's modification was realized as a low-level, variable implementation of the burst that does not involve the introduction of qualitative properties not present in the stimuli, such as formant structure. As for the selection account, to the extent that the manipulations present only on the third stimulus still cause some variation in the speakers' productions, it does not appear that they are consistently selecting either the first or second stimulus, which always reflected the baseline productions, and discounting the third stimulus. Nor are speakers selecting the last (manipulated) stimulus because doing so would lead to response patterns indistinguishable from those in the low-variability condition.

The abstraction account is also a less adequate explanation for similar reasons; it predicts that speakers would exhibit little sensitivity to the manipulations or any other fine-grained phonetic detail of the stimulus, as this level of detail should have been discarded prior to production. Again, the attenuation of the POV effect and the mimicry of the burst duration should not be present if speakers were simply "factoring out" acoustic detail when planning and executing their responses.

B. Language-specific and language-independent interpretation of acoustic cues

Several aspects of the data implicate language-specific interpretation of the acoustic cues in the data. While space precludes in-depth discussion of cue interpretation (see Wilson *et al.*, 2014 for further discussion), some issues that are germane to the low- versus high-variability conditions are addressed here.

First, because SS and SN sequences are not possible at the lexical level in word-initial position in English, English listeners are likely to interpret characteristics of the burst as evidence of a vowel. This interpretation is made more probable by the fact that English speakers often produce highly reduced vowels in words like "potato" or "tomato" and may even completely overlap the vocalic portion with the preceding stop release, giving rise to a lengthened period of aspiration after the stop and before the following consonant (Davidson, 2006). When listeners heard similar acoustic properties in the present stop-initial stimuli, they may have interpreted them as evidence of an unstressed reduced vowel. Voiced stop-initial sequences furthermore feature phonetic voicing during the burst, yet another cue that could indicate the presence of a reduced vowel between the members of a cluster. This could account for the voicing effect for both SS and SN sequences in this study, where the epenthesis modification is more frequent for voiced stops and more common than the no-modification response. This is consistently true for both of the variability conditions. Because the simple open transition in Russian stop-initial clusters is not contrastive for English speakers in word-initial position, we hypothesize that English speakers may be choosing a native articulatory configuration/acoustic output that matches as closely as possible the properties of the burst in the Russian stimuli (see further discussion of possible articulatory configurations in Davidson, 2006, 2010).

Second, the effect of the burst amplitude manipulation is noticeable only in the low-variability condition. Like longer durations, higher amplitudes generally led to increased epenthesis responses, although the effect was weaker and was found mainly for the SS stimuli. Again speakers may be using higher amplitude as indicative of a vowel rather than a burst (Flemming, 2009). For SS stimuli, recall that amplitude was lowered in the manipulated sequences. In responses to the baseline, or high-intensity stimuli, epenthesis rates did not differ substantially between the variability conditions (see top four cells of Fig. 6). However, in responses to the manipulated, or low-intensity stimuli, there was significantly more epenthesis in the high-variability condition and significantly more C1 change and C1 deletion in the low-variability condition. The increase in C1 modification in the low-variability condition presumably reflects the well-known finding that bursts are critical cues for the detection and identification of stops (e.g., Blumstein and Stevens, 1979). A stop with a low-amplitude burst before an obstruent may be misperceived, and ultimately realized by the speaker with the wrong place of articulation, or it may be missed entirely in perception and thus not realized in production. In the high-variability condition, the deletion and C1 change modifications mainly disappear. This finding is compatible with the blending account because speakers seem to have "filled in" the degraded information in the last stimulus with acoustic detail from the other exemplars to accurately identify the low-amplitude burst.

The effect of POV on prothesis in fricative-initial clusters provides the clearest case in which the effect of an acoustic manipulation was attenuated, but did disappear, in the high-variability condition. This finding could be related

to the fact that even for fricative-initial stimuli in which POV is absent, the prothesis rate was typically as high as, or higher than, the other modifications (see bottom two cells of Fig. 4). Because prothesis is independently a more viable repair for fricative-initial sequences, it is unsurprising that POV reinforces it in both conditions. The comparatively higher rates of prothesis as a modification for fricative-initial sequences mirror findings from cross-linguistic phonological studies, which have shown that prothesis is frequently preferred over epenthesis for repairing phonotactically ill-formed FC sequences (Broselow, 1992; Fleischhacker, 2005). It has been argued that this cross-linguistic preference may be attributable to the greater perceptual similarity between FC and əFC in comparison to FəC (Fleischhacker, 2005), a similarity relation that is claimed to be language-independent rather than specific to particular sound systems

A rather different finding was obtained for the stop-initial sequences when POV is present. In this case, while there is some sensitivity to the presence of POV in the low-variability condition, it disappears in the high-variability condition. Recall that prothesis is rare to nonexistent for stop-initial clusters overall. Instead speakers were more sensitive to the presence of the burst as evidenced by the overall high rates of epenthesis for stop-initial clusters in both variability conditions (and regardless of whether POV is present or absent). This suggests that the presence of POV in stops in particular is a fragile cue that loses its effectiveness when burst cues are strengthened in the high-variability condition. There is a possible perceptual explanation of the contrasting effects of POV for stop- and fricative-initial clusters. The POV manipulation for stops involves an initially increased intensity in voicing during the stop closure, which decreases but continues throughout the remainder of the closure. In the case of fricatives, POV is characterized by a discontinuity—voicing precedes the onset of frication, which constitutes a marked change in the type of spectral information that participants are receiving—and it has been argued that spectral discontinuities are perceptually salient (Ohala and Kawasaki-Fukumori, 1997).

To conclude this section, it is worth pointing out that while acoustic manipulations meant to enhance the likelihood of some modifications over others had a strong effect in the low-variability condition and weakened or disappeared in the high-variability condition, it is not the case that the high-variability condition led to accurate performance (no-modification responses) across the board. Some sequences, such as the fricative-initial clusters, reached above 70% accuracy in the high-variability condition, and voiceless stops were typically over 60% accurate. However, even these sequences had relatively large proportions of modifications, and for the voiced stop sequences, the epenthesis modification was more frequent than correct productions. The high proportion of modifications even in the high-variability condition reflect the fact that the tested clusters are phonotactically illegal for English speakers, who were unable to fully overcome native language constraints to correctly produce these sequences. There are several possible factors that contribute to the modification patterns, including perceptual interpretation (e.g., the voiced burst may be most interpretable as a vowel), difficulties in gestural coordination of the obstruent-initial sequences (see Davidson, 2010), and higher-level phonotactic constraints on cluster well-formedness. Determining the full range and interaction of factors that contribute to speakers' specific modifications beyond the effect of the acoustic manipulations is an interesting question for future research.

C. Phonetic variability and multiple talker input

This study examined the hypothesis that introducing within-trial variability in the acoustic realization of a nonnative sequence would provide participants with information that is useful for encoding and production. This hypothesis was largely confirmed: While speakers still produced modifications when presented with high-variability input, it was mainly the same one across the board (i.e., epenthesis), and it was not nearly as sensitive to acoustic cues as in the low-variability condition. To introduce the relevant acoustic variability, stimuli from three different talkers were presented on the same trial. We chose this protocol mainly because previous studies have shown that input from multiple talkers is effective in helping learners establish stable representations of spoken stimuli (e.g., Lively *et al.*, 1993; Wang *et al.*, 1999; Barcroft and Sommers, 2005; Iverson *et al.*, 2005; Richtsmeier *et al.*, 2009; Sommers and Barcroft, 2011; Sadakata and McQueen, 2013).

While previous studies have employed increased variability to beneficial effect in learning, ours is the first to demonstrate that within-trial variation stabilizes the production of novel phonological sequences. Experiment-wide variability was present in both the high- and low-variability conditions. However, an analysis of the three blocks for the low-variability condition indicated that no-modification responses (collapsed over all of the acoustic manipulations) did not increase over time, and the average proportions of the modifications (also collapsed) did not change from block to block (e.g., no-modification, block 1–3: 58%, 60%, 58%; epenthesis, block 1–3: 22%, 22%, 24%). This strongly suggests that experiment-wide variation did not have a stabilizing influence and therefore that within-trial variation was the driving force behind the increase in correct productions in the high-variability condition. This in turn suggests that when attempting to establish non-native sound structures, participants (and by extension learners) may especially benefit from opportunities to directly compare different phonetic realizations of the same nonnative sound structure.

On the basis of this study alone, we cannot say, definitively whether exposure to any substantial acoustic variability in close proximity would have the same effects or whether it is crucial that renditions be heard from different talkers. If three utterances with varying acoustic characteristics were produced by the same speaker, would the same results have been obtained? This will have to be clarified in future research, but previous research hints that components of variability that are apparently unrelated to the particular contrast under examination (e.g., CC vs CəC and əCC) is an important part of the equation. For example, Rost and McMurray (2010) compared infants' discrimination of minimal pairs on the voice onset dimension when presented with

different types of variability in the training phase. In the first experiment, infants were trained on words beginning with /p/ and /b/ that were produced with variable stop burst durations. In the second experiment, variability was manipulated along a few dimensions (burst duration, burst amplitude, and F0), but notably, the words were spoken by only one speaker and the non-VOT portion of the words remained unchanged across stimuli. In the third experiment, the words varied along a continuum for burst duration, but they were spoken by 18 different talkers. Only in Rost and McMurray's third experiment were the infants able to discriminate between test presentations of the minimal pairs. In Galle *et al.* (2015), a similar experimental design is also carried out except that instead of being trained on multiple talkers, infants are presented with a single talker producing stimuli that contain variability throughout the entire item rather than only on the VOT contrast that was tested. The infants in this study showed the same results as they did for the third experiment of Rost and McMurray (2010). Galle *et al.* (2015) explained this pattern of results by hypothesizing that irrelevant variability helps infants to focus on which aspects of the stimuli are most likely to be contrastive, regardless of whether the experiment involves multiple talkers or only one.

Whether the participants in the present study particularly benefited from variability that is specific to indexical information in different speakers' voices, or whether variation within one speaker's stimuli would have yielded equivalent results, remains to be seen. The most important aspect of the current findings is that being presented with multiple utterances that vary in acoustic implementation can alert nonnative speakers to the range of natural phonetic variation of foreign sound sequences. In the present case, when an English speaker hears a particularly long or loud release after a stop, these cues may be perceived as more typical of reduced vowel realizations or they may be recognizable as a stop burst (albeit an imperfect one). The English speaker must decide on an encoding of the acoustic input (e.g., as a CC cluster or a CəC sequence). When only one realization of the nonnative cluster is presented and it contains particularly vowel-like cues (or the cues are imperceptible, as when the burst intensity is lowered), speakers are more likely to encode it in a way that conforms to their native phonotactics. However, when multiple complementary sources of evidence about the phonemic composition of the cluster are provided, the English participants may realize that the extreme acoustic properties of one stimulus simply reflect acceptable phonetic variation for the intended CC sequence. By blending together all of the bundles of cues with which they are presented in a trial, nonnative speakers can dilute such "outlier" acoustic realizations and achieve greater stability in their encoding and subsequent production.

V. CONCLUSION

This study demonstrates that while speakers rely heavily on fine acoustic details to determine which sounds and gestures are present in nonnative sequences, this sensitivity to low-level information is attenuated when speakers are presented with multiple acoustic realizations of the same structure. We have argued for a blending account: Speakers weight and integrate the information from multiple utterances in a trial to determine the structure of nonnative sequences. Blending is one way of developing (implicit) knowledge of the range of acceptable phonetic variation for each sequence and of increasing the rate of correct production. In contrast, the results of the low-variability condition suggest that in the absence of trial-level variability, speakers are highly sensitive to fine-grained details of particular stimulus items and are more likely to modify sequences to conform to their native sound system.

Within-trial variability was introduced in this study through stimuli produced by multiple talkers, as previous research has shown that the acoustic information present in multiple voices is particularly useful in helping learners to establish more stable phonemic and lexical representations. While this study does not definitively establish that multiple talker variability is more useful than variation internal to a single talker, it does establish that variability within individual trials has a significantly different effect than variability present across trials over the course of an experiment. If further research converges with the finding that experiencing multiple phonetic realizations in close proximity can stabilize novel phonological (and by extension lexical) representations, this could have relevance for the study of first and second language acquisition.

ACKNOWLEDGMENTS

The authors would like to thank Alice Hall, Francesca Himelman, Johnny Mkitarian, Elizabeth George, and Steven Foley for their assistance in coding the data. We would also like to thank the members of the NYU Phonetics and Experimental Phonology Lab and the JHU Phonology/Phonetics Lab for their questions and comments. This research program has benefited from discussions with audiences at the University of Pennsylvania, the University of Delaware, Stony Brook University, the JHU IGERT Workshop, the 2012 Laboratory Phonology meeting in Stuttgart, Germany, and the Acoustical Society of America meeting in Montreal in 2013. This research was supported by NSF Grant Nos. BCS-1052855 to L.D. and BCS-1052784 to C.W.

[1]A possible addition to the study design would have been to vary whether the acoustically manipulated token was presented first or last in the high-variability condition; however, given the length of the existing experiment, it was not feasible to further consider stimulus order.

[2]The low-variability condition constitutes a subpart of a larger study (Wilson *et al.*, 2014). The data from the low-variability study is included here as a baseline against which the high-variability condition, which is of main interest, can be compared. The close match in stimulus items, procedure, and coding protocol (though not participants) should make the differences between the two data sets of particular relevance for understanding the unique contribution of multiple talkers in nonnative speech production.

[3]The associate editor and a reviewer point out two potential confounds in the presentation of the stimuli. First, there is not an equal number of presentations of the stimuli in the two conditions. One way to resolve this issue would have been to present three repetitions of the modified stimuli in the low-variability condition. However, it seems evident that further repetition of the acoustic manipulations would have the effect of reinforcing or enhancing the interpretation of the acoustic cues that is already

found in the participants' modifications in the low-variability condition (Goldinger, 1998). The other potential confound is that participants in the low-variability condition heard the modified stimulus from talker C twice, whereas the participants in the high-variability heard it only once. While it is possible that a single presentation of the modified stimulus might have led to diminished effects of the manipulations in the low-variability condition, we believe that results similar to the current findings would have been found even for a single presentation. First, as reported in Sec. III, we do still see some effects of the manipulated stimuli even in the high-variability condition. Second, findings from studies on shadowing and convergence have amply shown that imitation effects are present in perceptual evaluation (as in AXB tasks) even when a stimulus item is presented only once or speakers shadow continuously running speech (e.g., Namy *et al.*, 2002; Nye and Fowler, 2003; Pardo, 2006; Mitterer and Ernestus, 2008; Brouwer *et al.*, 2010). Thus despite the difference in the number of repetitions presented to the participants, we believe that the same outcome would have been obtained.

[4]The use of the term "epenthesis" is not meant to argue that speakers must be inserting a lexical vowel corresponding to English schwa. In previous studies, we have argued that the vocalic material present between the consonants in English speakers' productions does not necessarily correspond to the insertion of a lexical schwa but rather may arise from a gestural coordination pattern in which the consonant articulations do not overlap. It is not the purpose of this paper to further examine the precise nature of the epenthesis (or prothesis) modifications, so interested readers are referred to discussion in Davidson (2006, 2010) and Wilson *et al.* (2014). The main point about the epenthesis modification for this study is that when formants are present, English speakers are not achieving the Russian target.

Assmann, P., Nearey, T., and Hogan, J. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**(4), 975–989.

Barcroft, J., and Sommers, M. (**2005**). "Effects of acoustic variability on second language vocabulary learning," Stud. Second Lang. Acquis. **27**(3), 387–414.

Best, C., and Tyler, M. (**2007**). "Nonnative and second-language speech perception: Commonalities and complementarities," in *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*, edited by M. Munro and O.-S. Bohn (John Benjamins, Amsterdam), pp. 13–34.

Blumstein, S., and Stevens, K. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**(4), 1001–1017.

Bradlow, A., Pisoni, D., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/. IV. Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**(4), 2299–2310.

Broselow, E. (**1992**). "Transfer and universals in second language epenthesis," in *Language Transfer in Language Learning*, edited by S. Gass and L. Selinker, revised ed. (John Benjamins, Amsterdam), pp. 71–86.

Brouwer, S., Mitterer, H., and Huettig, F. (**2010**). "Shadowing reduced speech and alignment," J. Acoust. Soc. Am. **128**(1), EL32–EL37.

Cole, J., and Shattuck-Hufnagel, S. (**2011**). "The phonology and phonetics of perceived prosody: What do listeners imitate?," in *Proceedings of INTERSPEECH-2011*, Florence, Italy, pp. 969–972.

Davidson, L. (**2006**). "Schwa elision in fast speech: Segmental deletion or gestural overlap?," Phonetica **63**(2–3), 79–112.

Davidson, L. (**2010**). "Phonetic bases of similarities in cross-language production: Evidence from English and Catalan," J. Phon. **38**(2), 272–288.

Davidson, L., and Shaw, J. (**2012**). "Sources of illusion in consonant cluster perception," J. Phon. **40**(3), 234–248.

Docherty, G. (**2007**). "Speech in its natural habitat: Accounting for social factors in phonetic variability," in *Laboratory Phonology*, edited by J. Cole and J. I. Hualde (Mouton de Gruyter, Berlin), Vol. 9, pp. 1–36.

Ernestus, M. (**2012**). "Message related variation: Segmental within speaker variation," in *The Oxford Handbook of Laboratory Phonology*, edited by A. Cohn, C. Fougeron, and M. Huffman (Oxford University Press, Oxford, UK), pp. 92–102.

Flege, J. (**1995**). "Second-language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York, Timonium, MD), pp. 229–273.

Fleischhacker, H. (**2005**). "Similarity in phonology: Evidence from reduplication and loan adaptation," unpublished Ph.D. dissertation, UCLA, Los Angeles.

Flemming, E. (**2009**). "The phonetics of schwa vowels," in *Phonological Weakness in English: From Old to Present-Day English*, edited by D. Minkova (Palgrave, Macmillan, New York), pp. 78–95.

Galle, M. E., Apfelbaum, K. S., and McMurray, B. (**2015**). "The role of single talker acoustic variation in early word learning," Lang. Learn. Dev. **11**, 66–79.

Ganong, W. F. (**1980**). "Phonetic categorization in auditory word recognition," J. Exp. Psych. Hum. Percep. Perf. **6**(1), 110–125.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (**2013**). *Bayesian Data Analysis*, 3rd ed. (Chapman and Hall/CRC, Boca Raton, FL), Chap. 16, pp. 405–434.

Gelman, A., and Hill, J. (**2006**). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge, UK), Chap. 6, pp. 109–132.

Goldinger, S. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psych. Rev. **105**(2), 251–279.

Goldinger, S., Pisoni, D., and Logan, J. S. (**1991**). "On the nature of talker variability effects on recall of spoken word lists," J. Exp. Psych. Learn. Mem. Cognit. **17**(1), 152–162.

Hadfield, J. (**2010**). "MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package," J. Stat. Soft. **33**(2), 1–22; available at http://www.jstatsoft.org/v33/i02.

Iverson, P., Hazan, V., and Bannister, K. (**2005**). "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults," J. Acoust. Soc. Am. **118**(5), 3267–3278.

Johnson, K., and Mullennix, J. (**1997**). *Talker Variability in Speech Processing* (Academic, San Diego), 237 pp.

Kessinger, R., and Blumstein, S. (**1997**). "Effects of speaking rate on voice-onset time in Thai, French, and English," J. Phon. **25**(2), 143–168.

Kraljic, T., and Samuel, A. (**2007**). "Perceptual adjustments to multiple speakers," J. Mem. Lang. **56**(1), 1–15.

Krehm, M., Buchwald, A., and Vouloumanos, A. (**2012**). "The effect of variation on phonetic category learning," in *Supplement to the Proceedings of the 36th Annual Boston University Conference on Language Development*, edited by A. Biller, E. Chung, and A. Kimball (BUCLD Online, Boston, MA), pp. 1–12.

Kruschke, J. (**2011**). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS* (Academic, New York), 672 pp.

Lively, S., Logan, J. S., and Pisoni, D. (**1993**). "Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new perceptual categories," J. Acoust. Soc. Am. **94**(3), 1242–1255.

Logan, J. S., Lively, S., and Pisoni, D. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**(2), 2076–2087.

Mitterer, H., and Ernestus, M. (**2008**). "The link between speech perception and production is phonological and abstract: Evidence from the shadowing task," Cognition **109**(1), 168–173.

Mullennix, J., Pisoni, D., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.

Namy, L. L., Nygaard, L. C., and Sauerteig, D. (**2002**). "Gender differences in vocal ccommodation: The role of perception," J. Lang. Soc. Psych. **21**(4), 422–432.

Nielsen, K. (**2011**). "Specificity and abstractness of VOT imitation," J. Phon. **39**(2), 132–142.

Nye, P. W., and Fowler, C. A. (**2003**). "Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English," J. Phon. **31**(1), 63–79.

Ohala, J., and Kawasaki-Fukumori, H. (**1997**). "Alternatives to the sonority hierarchy for explaining segmental sequential constraints," in *Language and Its Ecology*, edited by S. Eliasson and E. H. Jahr (Mouton de Gruyter, Berlin), pp. 343–365.

Paradis, C., and LaCharité, D. (**1997**). "Preservation and minimality in loanword adaptation," J. Ling. **33**(2), 379–430.

Pardo, J. (**2006**). "On phonetic convergence during conversational interaction," J. Acoust. Soc. Am. **119**(4), 2382–2393.

Pisoni, D. (**1973**). "Auditory and phonetic codes in the discrimination of consonants and vowels," Percep. Psychophys. **13**, 253–260.

Plummer, M., Best, N., Cowles, K., and Vines, K. (**2006**). "CODA: Convergence diagnostics and output analysis for MCMC," R News **6**(1), 7–11; available at http://oro.open.ac.uk/id/eprint/22547.

Raudenbush, S., and Bryk, A. (**2002**). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Sage, Thousand Oaks, CA), 512 pp.
R Development Core Team. (**2012**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
Richtsmeier, P. T., Gerken, L., Goffman, L., and Hogan, T. (**2009**). "Statistical frequency in perception affects children's lexical production," Cognition **111**(3), 372–377.
Rost, G. C., and McMurray, B. (**2010**). "Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning," Infancy **15**(6), 608–635.
Sadakata, M., and McQueen, J. (**2013**). "High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates," J. Acoust. Soc. Am. **134**(2), 1324–1335.
Seidl, A., Onishi, K., and Cristia, A. (**2014**). "Talker variation aids young infants' phonotactic learning," Lang. Learn. Dev. **10**(4), 297–307.
Shockley, K., Sabadini, L., and Fowler, C. (**2004**). "Imitation in shadowing words," Percep. Psychophys. **66**(3), 422–429.
Sommers, M., and Barcroft, J. (**2006**). "Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification," J. Acoust. Soc. Am. **119**(4), 2406–2416.
Sommers, M., and Barcroft, J. (**2007**). "An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability," Appl. Psycholing. **28**(2), 231–249.
Sommers, M., and Barcroft, J. (**2011**). "Indexical information, encoding difficulty, and second language vocabulary learning," Appl. Psycholing. **32**(2), 417–434.
Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (**1999**). "Training American listeners to perceive Mandarin tones," J. Acoust. Soc. Am. **106**(6), 3649–3658.
Wilson, C., and Davidson, L. (**2013**). "Bayesian analysis of non-native cluster production," in *Proceedings of the Northeast Linguistics Society*, edited by S. Kan, C. Moore-Cantwell, and R. Staubs (Graduate Linguistic Student Association, Amherst, MA), Vol. 40, pp. 265–278.
Wilson, C., Davidson, L., and Martin, S. (**2014**). "Effects of acoustic-phonetic detail on cross-language speech production," J. Mem. Lang. **77**, 1–24.