Journal of Memory and Language 77 (2014) 1–24

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Effects of acoustic–phonetic detail on cross-language speech production

Colin Wilson [a,*], Lisa Davidson [b], Sean Martin [b]

[a] Johns Hopkins University, 3400 N. Charles Street, Department of Cognitive Science, Baltimore, MD 21218, United States
[b] Department of Linguistics, 10 Washington Place, New York, NY 10003, United States

ARTICLE INFO

Article history:
Received 4 October 2013
revision received 8 August 2014

Keywords:
Phonotactics
Cross-language speech production
Phonetic detail
Phonetic decoding

ABSTRACT

Nonnative sounds and sequences are systematically adapted in both perception and production. For example, American English speakers often modify illegal word-initial clusters by inserting a vocalic transition between the two consonants (e.g., (/bdagu/ → [bᵊdagu]). Previous work on such modifications has for the most part focused on relatively abstract properties of the nonnative structures, such as their phonemic content and whether they conform to sonority sequencing principles. The current study finds that fine-grained phonetic details of the stimulus can be equally important for predicting cross-language production patterns. Several acoustic–phonetic properties were manipulated to create stimulus variants that are phonemically identical (i.e., exhibit non-contrastive variation) in the target language (Russian). In a shadowing experiment, English speakers' correct productions and detailed error patterns were significantly modulated by the acoustic manipulations. The results highlight the role of perception in accounting for cross-language production, and establish limits on the perceptual repair of nonnative sound sequences by phonetic decoding.

Introduction

Research on cross-language speech perception and production has shown that nonnative sound patterns can be misperceived and modified in systematic ways. Perhaps best known are cases in which listeners fail to reliably distinguish individual sounds that do not contrast in their native language. For example, Japanese listeners have difficulty discriminating English word-initial /l/ and /ɹ/ (e.g., Guion, Flege, Akahane-Yamada, & Pruitt, 2000), and English listeners cannot reliably categorize the Hindi dental and retroflex stops (Pruitt, Jenkins, & Strange, 2006; Werker & Tees, 1984). There has also been considerable research on the perception of sounds in particular positions and combinations that do not occur natively, especially consonant clusters and word-final consonants. Dupoux, Kakehi, Hirose, Pallier, and Mehler (1999) provide evidence that Japanese listeners often 'perceptually epenthesize' a vowel between word-medial French consonants (e.g., /ebzo/ → [ebuzo]). Related cases of perceptual epenthesis and other types of perceptual 'repair' have been reported for a wide range of nonnative clusters (Berent, Steriade, Lennertz, & Vaknin, 2007; de Jong & Park, 2012; Hallé, Dominguez, Cuetos, & Segui, 2008; Kabak & Idsardi, 2007).

Psycholinguistic theories of cross-language speech perception, like related second language (L2) models (e.g., Best, 1995; Flege, 1995), have focused on the role of *phonetic decoding*. While details vary across accounts, the following description by Peperkamp and Dupoux (2003) is representative of how phonetic decoding is thought to apply to nonnative inputs:

* Corresponding author. Fax: +1 410 516 8020.
E-mail addresses: colin@cogsci.jhu.edu (C. Wilson), lisa.davidson@nyu.edu (L. Davidson).

http://dx.doi.org/10.1016/j.jml.2014.08.001
0749-596X/

> "During phonetic decoding, a given input sound will be mapped onto the closest available phonetic category . . .. With respect to nonnative sounds, this mapping is of course massively unfaithful, since the phonetic categories to which these sounds are mapped in the foreign language can simply be absent from the native one."

Unfaithful decoding also applies to nonnative sequences, as demonstrated by perceptual epenthesis (e.g., Dupoux, Parlato, Frota, Hirose, & Peperkamp, 2011), and to suprasegmental structures such as stress (e.g., Dupoux, Pallier, Sebastián-Gallés, & Mehler, 1997). In all cases, it is plausible that unfaithful decoding maps nonnative inputs to the most phonetically similar native sound structures (e.g., Best, 1995; Escudero, Simon, & Mitterer, 2012; Flege, 1995).

While phonetic decoding has been extensively investigated with perceptual tasks, a number of basic questions about the process and its connection to other components of the language system remain open. Does phonetic decoding consistently map incoming speech signals to phonetic/phonological representations that are legal in the native language, or are illegal representations sometimes formed? If the latter, what factors determine the relative probability with which phonetic decoding 'repairs' or leaves intact a given nonnative structure? Finally, can nonnative structures that are faithfully represented by phonetic decoding be preserved by subsequent task-dependent processes?

We address these questions by investigating how the acoustic–phonetic details of nonnative inputs affect *speech production*. Specifically, we focus on how English speakers with no prior knowledge of Russian produce consonant clusters such as those at the beginning of words like /knʲigə/ 'book' and /zdarov/ 'healthy'. Adopting a method that has been widely used in perception studies, but which has only rarely been applied to production, we systematically manipulate acoustic properties—including the presence of voicing before the beginning of an obstruent, and the amplitude and duration of stop bursts—to create phonetic variants of the clusters. These properties are part of the non-contrastive system of phonetic realization in Russian speech. Our main focus here is the relation between the (manipulated) acoustics of stimulus clusters and the detailed production patterns of English speakers.

Previous production studies have found that nonnative consonant clusters are often modified by epenthesis and a wide range of other 'repairs', including consonant deletion and change of one or more distinctive features (Broselow, 1992; Broselow & Finer, 1991; Davidson, 2006a, 2010; Hancin-Bhatt & Bhatt, 1997). It is also known that English speakers can produce such clusters correctly—matching the phonetic realization of Russian speakers—a certain proportion of the time. However, the *rates* and *types* of modification and correct production vary across clusters in a way that has not been satisfactorily explained. If detailed modification patterns can be demonstrated to be sensitive to fine-grained phonetic details of the stimulus, this will simultaneously shed light on the phonetic decoding process and provide novel insights about a rich body of cross-language production data.

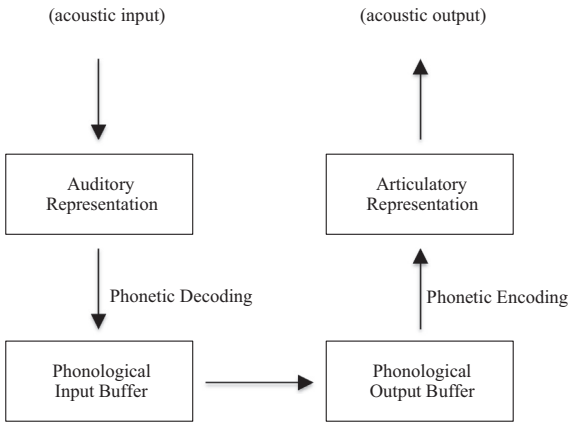

Fig. 1. Processing architecture for nonword repetition (see text for details).

Our discussion is framed within a cognitive processing architecture that has been developed for native word perception and production (Ellis & Young, 1988; Goldrick & Rapp, 2007; Patterson & Shewell, 1987; Ramus et al., 2010). In this architecture, illustrated in Fig. 1, our task (nonword repetition) is decomposed into phonetic decoding, which maps an auditory form to a representation in the phonological input buffer, and a production process, in which phonological encoding creates a representation in the phonological output buffer that is implemented by vocal tract movements. While some modifications of nonnative structures may originate in phonological encoding or articulatory execution (Davidson, 2006a et seq., see General Discussion), our experimental manipulations target phonetic decoding. We begin by considering how patterns of nonnative perception and production bear on the nature of this process, and then turn to the motivation and design of our production experiment.

Phonetic decoding in nonnative perception and production

The input to phonetic decoding is an auditory representation of the incoming acoustic signal. Evidence for language-specific shaping of the auditory system is presently mixed (Breen, Kingston, & Sanders, 2013; Dehaene-Lambertz, Dupoux, & Gout, 2000; Jacquemot, Pallier, LiBihan, Dehaene, & Dupoux, 2003), so we take auditory representations to be largely language-independent (Kingston, 2005). These representations contain measurements of acoustic–phonetic properties—such as formants, durations, and intensities—that are commonly referred to as *cues* in the speech perception literature (e.g., Lisker, 1986; Wright, 2004). Phonetic decoding interprets the cues from the stimulus as phonetic/phonological structures consisting of segments, syllables, etc. Language-specific sound structures begin to influence processing at the level of phonetic decoding, but the nature and extent of the influence there (and at later levels) is not fully understood.

As indicated by the quote from Peperkamp and Dupoux (2003) above, previous perceptual investigations of

phonetic decoding have highlighted *misinterpretation* of nonnative stimuli. Patterns of misperception have been taken as evidence for 'warping' of the perceptual space by the native language (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), for the active role of phonotactic constraints in speech perception (Dupoux et al., 2011), and for universal principles of phonological well-formedness (Berent, Lennertz, Jun, Moreno, & Smolensky, 2008; Berent, Lennertz, Smolensky, & Vaknin-Nusbaum, 2009; Berent et al., 2007). However, as in the case of cross-language production, listeners' identification and discrimination of nonnative stimuli often exceeds chance levels. For example, while Japanese listeners identify a vowel in French productions of clusters (as in /ebzo/) approximately 65% of the time, this is still far lower than the approximately 90% rate at which they perceive full vowels actually produced by French speakers (as in /ebuzo/ or /ebizo/; see Dupoux et al., 2011, Table 1). Similarly, Berent et al. (2007) have shown that English listeners can reliably distinguish Russian CC and CVC sequences (e.g., /bdif/ vs. /bedif/) under task conditions that encourage attention to phonetic detail.

These complex patterns of correct and incorrect performance may appear to resolve one of the questions raised above, namely whether phonetic decoding always maps stimuli to native structures or does so only probabilistically (but see 'Accurate and modified output patterns in nonnative speech production' on alternative interpretations of the perceptual findings). However, since the issue under consideration is the influence of acoustic cues in cross-language speech processing generally, converging evidence should be sought from other types of performance. An important question is whether relatively small, non-contrastive differences in the acoustic structure of the stimuli can lead to qualitatively different outputs in *production*. This type of evidence would indicate that not only is the phonetic decoding stage sensitive to the presence of fine-grained differences in the auditory representation of nonnative inputs, but also that such differences are preserved in some form in downstream systems like those dedicated to phonological encoding and articulation.

With respect to *cross-language* speech production in particular, phonetic decoding has a clear but relatively neglected role to play in accounting for performance. In common laboratory production tasks, the stimulus is a recording of naturally-produced speech from another language, and is therefore rich in auditory cues (e.g., Hawkins, 2003). When the participant is familiar with the lexicon and orthographic conventions of the other language, as in many L2 studies, or with conventionalized loanword adaptation patterns, as in cases of extensive language contact, phonetic decoding may be subordinated to other sources of information about the production target (Hayes-Harb, Nicol, & Barker, 2010; Vendelin & Peperkamp, 2006; Young-Scholten, Akita, & Cross, 1999). But on first exposure to the nonnative language, without orthographic or other supporting information, speakers are likely to rely primarily on phonetic decoding to determine their production targets.

The architecture in Fig. 1 provides many opportunities for knowledge of the native language to shape production patterns: first in phonetic decoding, then in phonological encoding and articulatory execution. Note that the production system, like that of perception, is known to be biased in favor of sounds patterns that occur often in the native language (Dell, Juliano, & Govindjee, 1993; Vitevich & Luce, 2005). The sources and nature of production modifications must therefore be determined by careful experimental manipulation and analysis of the responses. However, *correct* productions are less ambiguous: if a nonnative structure is produced without modification, it is highly likely to have been faithfully represented by phonetic decoding. It is this logic that allows evidence from production to bear on the question of whether phonetic decoding consistently 'repairs' nonnative stimuli, and (if not) what conditions are most favorable to faithful decoding.[1]

There are two potential concerns with this logic that we address here, one methodological and one theoretical. The methodological concern applies to the classification of production responses as correct or modified. While many previous cross-language production studies have relied on perceptual coding (i.e., native speaker transcriptions of nonnative responses), it is now generally recognized that careful phonetic analysis is necessary in order to understand the production differences between native and nonnative speakers (Davidson, 2006a, 2010; de Jong & Park, 2012; Zsiga, 2003). With respect to the study of sound sequences, Davidson (2006a, 2010) has argued that production of nonnative consonant clusters can be properly characterized only through the analysis of acoustic–phonetic properties such as vowel formants and articulatory parameters such as gestural overlap. We follow stringent coding guidelines in analyzing the cluster production data in this study.

The theoretical concern is that the processing architecture in Fig. 1 may not be sufficient for understanding performance in cross-language production tasks. The architecture does not include phonological input and output lexicons, which are clearly relevant for perception and production of known words. However, previous production studies using our nonnative production paradigm have failed to find strong effects of native lexical

Table 1
Proportions of production responses to zN stimulus items in Davidson (2010). Stimulus-specific prothesis rates are highlighted.

Stimulus item	Correct	Epenthesis	Prothesis	C1 deletion
/zmafo/	.75	.13	**.13**	.00
/zmagu/	.36	.00	**.64**	.00
/znafe/	.78	.22	**.00**	.00
/znagi/	.45	.00	**.55**	.00

[1] Orthographic transcription of nonnative auditory stimuli could potentially serve the same purpose (e.g., Berent et al., 2007; Best, McRoberts, & Goodell, 2001; Hallé et al., 2008). However, interpretation of transcription data depends upon a clear understanding of the system that relates perceived to spelled forms, a system that is only quasi-regular in languages like English (Venezky, 1970). Furthermore, orthographic tasks necessarily impose discrete categories on mental representations that may be more gradient or continuous, an issue that is particularly pressing when fine-grained phonetic properties are the focus of investigation.

knowledge (e.g., in the form of neighborhood densities or phoneme-level transitional probabilities, Berent et al., 2007; Davidson, 2006a; Hallé & Best, 2007; Pitt, 1998). While the previous results are consistent with a role for more abstract, feature-based phonotactic restrictions that might be induced from the lexicon (Hayes & Wilson, 2008), such constraints are plausibly embedded within the sublexical processes of phonetic decoding and phonological encoding rather than being due to active lexical processes (e.g., Vitevich & Luce, 1999, 2005).

A potentially more serious omission from the processing model is a direct link from auditory representations to articulation, bypassing phonetic decoding (and phonological encoding). Some alternative models, such as DIVA (Guenther, Ghosh, & Tourville, 2006) and the dual-stream model (Hickok & Poeppel, 2007), do posit a relatively direct mapping from fine-grained perceptual representations to speech production processes. Moreover, a tight connection between speech perception and production is supported by a number of conceptual and empirical arguments, including the need for speakers to learn and maintain auditory-motor mappings that are appropriate for their native language (Guenther & Vladusich, 2012; Hickok & Poeppel, 2007), the ability of speakers to rapidly shadow auditorily presented sounds (Fowler, Brown, Sabadini, & Weihing, 2003; Porter & Lubker, 1980) and to adapt to auditory distortions of their own speech (Purcell & Munhall, 2006; Villacorta, Perkell, & Guenther, 2007), as well as neurophysiological connectivity of the dorsal auditory stream (Hickok & Poeppel, 2007). However, while previous evidence suggests that this detailed perception/production link is not mediated by the lexicon, it may nevertheless engage sublexical processes such as phonetic decoding (Cole & Shattuck-Hufnagel, 2011; Mitterer & Ernestus, 2008). Our main analysis focuses on the role of such processes, but we also investigate whether our results are likely to be due to a more direct connection between auditory encoding and articulation (as has been sometimes suggested for phonetic imitation).

Motivation and design of current study

The current study is motivated by close analysis of the cross-language production data reported in Davidson (2010), an analysis that strongly suggests an influence of phonetic decoding. As in the current study, Davidson (2010) examined the ability of English (and Catalan) speakers to produce Russian word-initial consonant clusters. The results showed a gradient pattern of performance, with some clusters eliciting higher rates of correct production (e.g., /zm/) than others (e.g., /bd/). They also revealed a number of modifications in addition to the dominant 'repair' of epenthesis, such as deletion of the first consonant or prothesis of a vowel before the consonant sequence.

While Davidson (2010) used several stimulus items for each cluster type, the original analysis of the results collapsed over individual stimuli. Wilson and Davidson (2013) examined performance on each stimulus item and discovered that English listeners are highly sensitive to phonetic variation across the stimulus items that is non-contrastive for the Russian speaker. This careful inspection of the previous results reveals several findings that would not be expected if nonnative clusters were consistently recoded as native structures, or consistently represented faithfully, by phonetic decoding.

In the most striking instances, illustrated in Table 1, stimuli that begin with exactly the same phonological sequence (e.g., /zm/ or /zn/)—but that differed in their fine phonetic details—were produced with different patterns of modification. In the case of voiced obstruent-initial clusters, the acoustic phonetic property identified as most relevant by Wilson and Davidson (2013) was pre-obstruent voicing (POV): an interval of voicing before the formation of a voiced obstruent constriction with a visibly higher amplitude than the typical voicing during the constriction. Russian speakers tend to produce voiced fricatives with voicing that is continuous throughout, and POV occurs in such sounds when voicing precedes frication (Jones & Ward, 1969:117–118). For voiced stop-initial clusters, Russians regularly produce voicing during the stop closure; POV as we define it occurs when the onset of voicing has a higher amplitude than that during the remainder of the closure. POV was naturally produced in several fricative- and stop-initial tokens by the Russian speaker who recorded the Davidson (2010) items. It was found that stimuli containing this voicing profile (e.g., /zmagu/, /znagi/) were produced with significant rates of prothesis, while stimuli lacking POV (e.g., /zmafo/, /znafe/) were not modified in this way and were most often produced correctly.

Because Davidson (2010) did not deliberately control the phonetic properties of the stimuli, correlations between the acoustic details of individual stimuli and listeners' responses discovered in Wilson and Davidson (2013) are limited to the 'accidental' variation in phonetic implementation across the materials. Thus, the purpose of the current study was to systematically manipulate the low-level acoustic properties that listeners were apparently attending to in the earlier study, and to quantify the influence of such properties on nonnative speech production. On the basis of the previous findings, three relevant acoustic cues were identified as affecting responses to individual stimuli: the presence or absence of pre-voicing (POV), stop burst duration, and burst amplitude.

First, POV was manipulated for both voiced stop-initial and fricative-initial stimuli. For the stimuli of the current study, we manipulated whether or not fricative-initial clusters contained voicing that began before the frication, and whether stop-initial sequences had closure voicing that began at a higher amplitude voicing then tapered off or voicing that retained the same amplitude throughout the closure. It is important to note that the initial voicing of POV tokens lacks visible formant structure, and consequently that items with POV are phonetically distinguishable from stimuli with a vowel before the consonant cluster.

Second, findings from Wilson and Davidson (2013) suggested that longer stop bursts give rise to more epenthesis modifications. Thus, in the current study, we manipulated the duration of the stop bursts along a four-step continuum. We predicted that an increase in duration should correlate with increased epenthesis responses, but the

continuum should allow us to more precisely identify the length at which listeners transition from perceiving only a stop burst to interpreting the longer length as containing a (reduced) vowel.

Third, stop bursts that had higher intensity relative to the following sound also elicited increased rates of epenthesis. To further examine this effect in the current study, higher- and lower-amplitude versions of word-initial stop bursts were created. In the case of stop-stop clusters, the relative amplitude of the burst of the initial stop is high, because the burst is followed by the second stop's closure. Accordingly, a lower-amplitude version of such bursts was created. For stop-nasal clusters, the amplitude of the burst is low relative to the following nasal murmur, thus a higher-amplitude version was created. Two predictions about the amplitude manipulation can be made. First, stop bursts of higher relative amplitude should give rise to higher epenthesis rates. Second, as noted in Wilson and Davidson (2013), stops with lower amplitude bursts should be more susceptible to deletion or misperception.

In addition to performing these individual manipulations, we also combined them to examine how the manipulation of multiple phonetic cues impacts performance. Specifically, we crossed the burst duration manipulation with burst amplitude (for clusters beginning with voiced and voiceless stops) and separately with POV (for clusters beginning with voiced stops only). Expectations about how cues should interact in the two cases are different. Longer burst duration and higher amplitude are both expected to increase epenthesis modifications. In contrast, burst duration and POV support different modifications: longer bursts should result in epenthesis, while presence of pre-voicing should lead to prothesis. Our crossed manipulation allowed us to examine whether one of these cues has greater influence, or if multiple cue manipulations lead speakers to implement more than one modification in a single production response.

The acoustic modifications studied in this paper significantly extend those examined in previous research. Previous studies of nonnative cluster perception have manipulated only the duration of the transition between two consonants (Berent et al., 2007; Dupoux et al., 1999), typically by splicing out increasingly longer portions of a full vowel produced between the consonants in the original recording. This method introduces the confound that coarticulatory traces of the full vowel could still be present on the neighboring sounds (Dupoux et al., 2011). The current study avoids this issue by manipulating the duration of the bursts in recordings of clusters. Moreover, few studies have reported on the phonetic measurements of their stimuli, such as relative amplitude or prevoicing (but see Berent, 2008 for duration measurements of the consonants in a cluster, and Berent et al., 2008 for absolute amplitude measurements). Furthermore, to our knowledge no previous work on nonnative cluster production has deliberately manipulated any phonetic cues. Acoustic–phonetic properties are highly variable across languages (even for the realization of phonologically-identical sequences) and cannot be eliminated from natural speech stimuli. By deliberately modifying these properties in a large-scale (with respect to both the number of stimuli and speakers) study of nonnative consonant cluster production, we aim to examine how modification patterns in cross-language speech production can better inform the nature of phonetic decoding of nonnative sound structures.

Method

Participants

The participants were 24 New York University graduate and undergraduate students. They were all native speakers of American English ranging in age from 19 to 32 who spoke neither Slavic languages nor any other languages with initial obstruent clusters, such as Hebrew. None of the participants reported any speech or hearing impairments. They were compensated $10 for their participation.

Materials

Critical stimuli consisted of nonce words of the form CCáCV ('á' indicates the stressed vowel). The initial consonant clusters were composed of fricative-nasal (FN), fricative-stop (FS), stop-nasal (SN), and stop-stop (SS) sequences; the individual clusters tested are shown in Table 2. Stop-initial clusters contained both voiced and voiceless consonants. For fricative-initial clusters, only voiced fricatives were included to limit the number of stimuli, as previous work has shown that English speakers are quite accurate at producing illegal voiceless fricative-initial clusters (Davidson, 2006a, 2010). Each cluster appeared in four distinct stimulus items (see Appendix A), for a total of 96 CC-initial stimuli. In addition to the critical CC-initial items, there were also fillers of the form CəCáCV (48 items) and əCCáCV (48 items). To create the fillers, two of the four stimulus items for each initial cluster were chosen at random and the –áCV ending from those items was appended 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/). The phonemes of the CV stimulus endings were controlled so that each ending occurred approximately equally often and with a range of initial clusters.

All of the stimuli were produced by a Russian–English bilingual linguist who had no difficulty producing the words with the appropriate stress pattern and with reduced vowels (represented by schwa in the transcription) in the fillers.

The recorded stimuli were modified to conform to the acoustic manipulations discussed previously. The first modification was pre-obstruent voicing (POV), which was

Table 2
Target consonant clusters used in the CCáCV stimuli.

Cluster type	Voiceless C1	Voiced C1
Fricative-Nasal	(not studied)	/vm, vn, zm, zn/
Fricative-Stop	(not studied)	/vd, vg, zb, zg/
Stop-Nasal	/pn, tm, km, kn/	/bn, dm, gm, gn/
Stop-Stop	/pt, tp, kp, kt/	/bd, db, gb, gd/

operationalized as an interval of voicing before the beginning of frication or higher amplitude voicing at the onset of a stop that decreases in intensity throughout the closure. Each item beginning with a voiced obstruent had versions with and without POV. When a recording had naturally-produced POV, this was spliced out to create the non-POV variant. For stimuli that were not originally produced with POV, the initial voiced interval was spliced in from the waveform of a different utterance of the same consonant. All splices were taken at zero-crossings to avoid acoustic artifacts. This manipulation affected voiced FN, FS, SN, and SS, stimuli. Stimuli with POV are illustrated in Fig. 2b (SS) and 2d (FS).

The second manipulated acoustic property was the duration of the burst (initial transient and following frication) of the first consonant in stop-initial stimuli. Four levels of burst duration were generated: 20 ms, 30 ms, 40 ms, and 50 ms. Most of the burst durations as originally produced by the Russian speaker were between 20 and 40 ms, regardless of voicing (SN: mean = 36 ms, *sd* = 17.7 ms; SS: mean = 28 ms, *sd* = 8.7 ms), with several bursts (20%) exceeding 40 ms. Note that the distribution of original productions supports our assumption that burst duration, like POV and burst amplitude, varies subphonemically in Russian clusters approximately within the range tested here. Shorter durations in the stimuli were created by splicing 5–10 ms out of the original burst of the first consonant. Longer durations were created by selecting between 10 and 20 ms of the middle portion of the burst and splicing that material back into the recording. Splices were again taken from or inserted at zero crossings to avoid acoustic discontinuities. The duration manipulation affected voiced and voiceless SN and SS stimuli. Examples of the 20 ms and 50 ms manipulations are shown in Fig. 2a, b (20 ms) and c (50 ms).

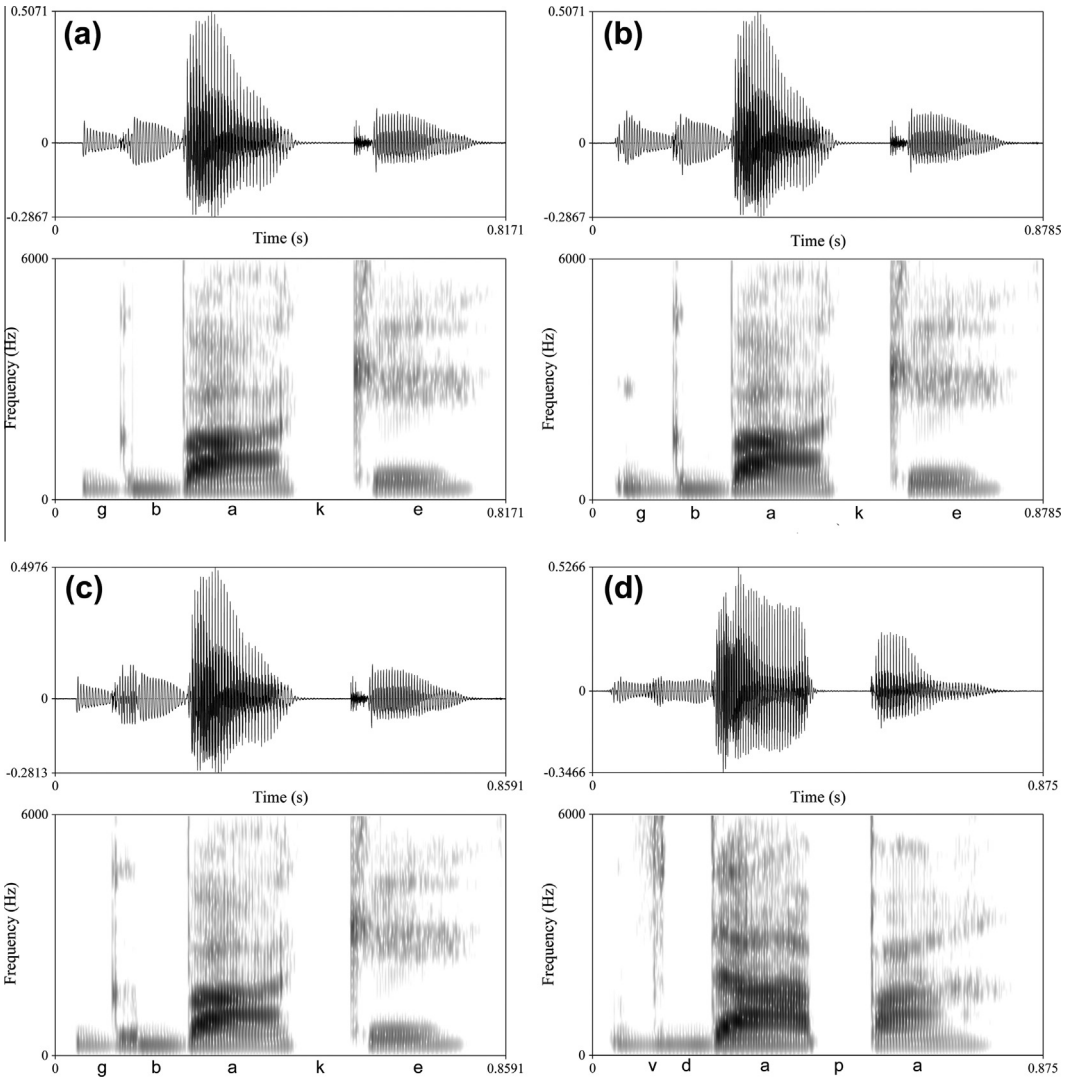

Fig. 2. Waveforms and spectrograms illustrating stimulus variations. (a) Stimulus item /gbake/. Burst duration = 20 ms, low amplitude burst, no pre-obstruent voicing. (b) Stimulus item /gbake/. Burst duration = 20 ms, high amplitude burst, pre-obstruent voicing. (c) Stimulus item /gbake/. Burst duration = 50 ms, high amplitude burst, no pre-obstruent voicing. (d) Stimulus item /vdapa/ with pre-obstruent voicing present.

Table 3
Summary of acoustic manipulations in the stimuli.

Cluster type	Crossed acoustic manipulations
Fricative-initial	POV (present vs. absent)
Voiceless-stop initial	DUR (20, 30, 40, 50 ms) × AMP (high vs. low)
Voiced stop-initial	DUR (20, 30, 40, 50 ms) × AMP (high vs. low)
	DUR (20, 30, 40, 50 ms) × POV (present vs. absent)

The third modification targeted the amplitude of the burst of initial stops. As discussed earlier, SS and SN clusters were subject to different manipulations, consistent with the natural difference in relative amplitude for these two types of sequence. The intensity of the burst in SN clusters was raised, while the intensity of the burst in SS clusters was lowered, relative to average baselines. To determine the appropriate values for the manipulations, we first defined the baseline for each cluster type as the average intensity of the burst relative to the following consonant (stop or nasal) in the original recordings. Bursts in the stimuli were then normalized using Praat (Boersma & Weenink, 2013) so that all tokens within a cluster type had the same baseline relative burst amplitudes—voiceless SN: -18 dB; voiced SN: −7 dB; voiceless SS +23 dB; voiced SS: 0 dB. For the manipulation of the SN clusters, which had a raised amplitude relative to the baseline, the burst of the voiceless tokens were increased to −10 dB and the voiced tokens to 0 dB. For the SS clusters, which were manipulated to have a lower amplitude relative to the baseline, the bursts were lowered to +13 dB for voiceless tokens and −7 dB for voiced tokens. These values were chosen because they ensured that all bursts remained perceptible (especially relevant in the case of the lowered amplitude for SS tokens) and resulted in stimuli that sounded relatively natural. An example of the low amplitude burst for SS sequences is shown in Fig. 2a, and the high amplitude burst is shown in Fig. 2b (along with POV). Raw amplitude values for the stop-initial clusters are provided in the Appendix.[2]

Because we were interested in how cues interact in the determination of production responses, we crossed the acoustic manipulations where possible. A complete summary of the manipulations is provided in Table 3. Together, all of the manipulated stimuli and the fillers came to a total of 800 sound files. To create an experimental procedure that would not be too taxing for the participants, 12 counterbalanced lists were created containing 288 stimuli each. Each list was composed of 32 FN, 32 FS, 64 SN, 64 SS (half voiced, half voiceless), and 48 CəC and əCC fillers each. The manipulations were distributed so that each experimental list contained approximately the same number of each manipulation type (within each cluster type) and each critical version occurred equally often across the lists. Two participants were assigned to each list.

[2] As noted by a reviewer, the amplitude manipulation as implemented in Praat affects all frequencies in the interval equally, whereas natural increases in amplitude may affect higher frequencies more than lower frequencies. We do not consider this to be necessarily a drawback of our method, as it is likely that low-frequency energy contributes most substantially to the perceptual illusion of a vowel, and the effects of amplitude we observe here parallel those found with the naturally-produced stimuli of Wilson and Davidson (2013).

Procedure

Participants were seated in a sound-attenuated room with a computer running ePrime 1.1 (Psychology Software Tools, Pittsburgh, PA). Each stimulus was presented twice before the response; no orthographic or other information accompanied the audio. The two repetitions of a stimulus were separated by 450 ms, and participants were given 1.5 s after the presentation of the second repetition to respond before the program automatically moved on to the next item. The chosen response interval has been used in previous studies (e.g., Davidson, 2010), where it was determined to be sufficiently long to elicit fluent, immediate production responses that do not overlap with the following trial. Participants were not given the opportunity to correct or otherwise evaluate their production responses.

The 288 items were divided into three blocks in order to give the participants a chance to rest. Participants' 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 bit). The experiment began with six practice trials containing clusters different from those used in the study.

Data analysis

The procedure above resulted in more than 5000 spoken production responses. Each response was analyzed by repeated listening to each recording and examination of its waveform and spectrogram in Praat to determine what, if any, modification had been made. Responses were coded by three research assistants and two authors (LD and SM). All of the coding was done blindly; that is, the coders did not know what manipulations corresponded to the utterance. Coding decisions were then discussed by at least two different research assistants and the two authors in regular lab meetings to ensure that all of the coders were in agreement on the labels assigned to each of the responses.

Modifications relative to the native Russian speaker's productions were labeled as shown in Table 4. If multiple errors occurred, each error was labeled, and if none of the errors of Table 4 occurred, the token was labeled as 'correct' (no modification). A token was coded for epenthesis if there was vocalic material containing visible first and second formants following either the frication of a fricative

Table 4
Response codes for CC stimuli.

Response Type	Definition	Example
Epenthesis	Target is produced with vocalic material between the consonants	/pkadi/ → [pᵊkadi]
Prothesis	Target is produced with vocalic material before the cluster	/pkadi/ → [ᵊpkadi]
C1 Deletion	Target is produced with the first consonant deleted	/pkadi/ → [kadi]
C1 Change	Target is produced as a cluster, but with a different first consonant	/pkadi/ → [skadi]

C1, or the burst of a stop C1, that ended with abrupt lowering of intensity at the onset of the second stop, fricative, or nasal. The use of second and higher formant structure as a diagnostic for the presence of a vocalic interval (henceforth, a *vocoid*) is typical of many previous studies of nonnative consonant cluster production, as well as studies of variable vowel deletion and devoicing (see, for example, the studies reviewed in Beckman, 1996). Tokens coded for prothesis had a vocoid with visible first and second formants before the obstruent; voicing during the closure, or voicing which started before the frication for fricative-initial clusters, were not considered as errors because these acoustic implementations are present in some of the stimuli (as in natural Russian productions). More generally, an utterance was coded as correct if the cluster produced by the participant matched the voice, place, and manner specifications of the input, and the consonants were produced in the correct linear order, as determined using the spectrogram. Error coding 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.

A small portion of the data (2.2%) was omitted from all analyses because of disfluency, failure to produce the target, or modifications other than those listed above (e.g., /pkadi/ → Ø, /pkadi/ → [kpadi], /pkadi/ → [spadi]).

Results

The effects of phonetic and other factors on coded responses were analyzed with Bayesian generalized linear mixed-effects models (e.g., Gelman et al., 2013; Gelman & Hill, 2006). Because there are multiple unordered response categories, the appropriate statistical analysis is multinomial (polytomous) logistic regression (Raudenbush & Bryk, 2002). Previous research on multiple-category data in speech perception (e.g., McMurray & Jongman, 2011) and speech production (e.g., Davidson, 2006a, 2010) has used binary logistic regression, which often requires multiple analyses to be performed on overlapping data subsets.[3]

The multinomial dependent variable took the form of several binary columns, one for each of the modification types in Table 4. For each response, a value of 1 in a column indicates that the corresponding modification type was present and 0 indicates its absence. This data format straightforwardly allows for responses with multiple modifications (e.g., epenthesis and C1 change), which constitute a small but non-negligible proportion of the data (7% of total responses). The correct (no-modification) response category served as the baseline against which other response types were compared.

The fixed-effect specifications of the analyses below included acoustic manipulations (as appropriate for each consonant cluster type), as well as factors coding Cluster Profile (e.g., stop-nasal vs. stop-stop) and Cluster Voice (e.g., voiceless vs. voiced). All fixed factors except for burst duration are binary. The binary factors were effect (sum-to-zero) coded and scaled so that each had a mean of 0 and a difference in upper and lower values of 1 (Gelman et al., 2013); this is equivalent to converting such factors to *z*-scores. The burst duration factor, which has 4 levels, was coded with three scaled binary factors, each one comparing a longer duration (30, 40, and 50 ms) to the shortest duration (20 ms). Scaling allows the fixed-effect coefficients, which correspond to log-odds changes in response probabilities relative to no-modification, to be straightforwardly compared with one another. The random-effect specifications included intercepts and slopes for participants and items that were maximal given the experimental design. Only responses to items beginning with clusters were analyzed, as responses to fillers were at ceiling (95% for CVCVCV fillers and 92% for VCCVCV fillers).

Rather than relying on point estimates of the regression coefficients, we performed a Bayesian analysis that sampled coefficients from the posterior probability distribution conditioned on the data and the model's prior. Sampling was performed with the MCMCglmm package (Hadfield, 2010) in R (R Development Core Team., 2012) using prior and other settings that are standard for mixed-effects multinomial models (e.g., Hadfield, 2010). The statistical significance of each coefficient was assessed with 95% highest posterior density (HPD) intervals (Kruschke, 2011) as computed by applying the coda package (Plummer, Best, Cowles, & Vines, 2006) to the output of MCMCglmm. We report results in the form of mean coefficients together with 95% HPD intervals and associated *p*-values (determined by the proportion of posterior samples that lie on the same side of zero as the mean). HPD intervals are Bayesian alternatives to confidence intervals, and have a more straightforward interpretation. If the model is correctly specified, the probability that a coefficient falls within its 95% interval is 0.95; consequently, intervals that do not overlap with zero indicate non-null effects with probability $p < .05$.

[3] In contrast, a single multinomial analysis provides a global view of the data, capturing overall response biases and modulation of response probabilities by experimentally manipulated factors while avoiding the statistical issues raised by non-independent tests. For prior work employing a statistical methodology similar to the one adopted here, though with an empirical focus on neuropsychological data, see Kittredge, Dell, Verkuilen, and Schwartz (2008) and Nozari, Kittredge, Dell, and Schwartz (2010). We did perform binary logistic regressions for particular response types, but only subsequent to finding significant effects in the main multinomial analyses.

Analysis of fricative-initial clusters

For fricative-initial stimuli, the only manipulated phonetic factor was POV. We expected that presence of POV would increase the probability of prothesis modifications. The model also included a fixed effect of Cluster Profile (fricative-nasal vs. fricative-stop), allowing us to assess the influence of the phonemic content of the cluster on response patterns. (Recall that all tested fricative-initial clusters began with voiced /v/ or /z/, so Cluster Voice was not included in this model.)

Fig. 3 shows the observed proportion of each coded response for each level of Cluster Profile and POV. Each bar indicates the mean proportion of trials on which a response type occurs for the relevant stimulus subset across all participants. Error bars indicate 95% BCa bootstrap intervals (Efron, 1987) for the mean proportions, again across participants.

The model has relatively few fixed effects that are estimated to be significantly different from zero (see Supplementary Materials for a complete listing). The intercept (i.e., the grand 'mean' on the log-odds scale) for each modification was significant and negative (*epenthesis* = −2.16, *95% HPD* [−3.03, −1.29], $p < .001$; *prothesis* = −1.29, *95% HPD* [−2.14, −0.44], $p < .01$; *C1-deletion* = −5.01, *95% HPD* [−7.10, −2.71], $p < .001$; *C1-change* = –1.39, *95% HPD* [−2.07, −0.74], $p < .001$), indicating that the highest probability response overall is no-modification. HPD intervals computed on the differences of the sampled coefficients indicate that the overall probabilities of epenthesis, prothesis, and C1-change do not differ significantly from one another and are all significantly higher than the probability of C1-deletion.

Consistent with expectations formed on the basis of our pilot study, POV significantly increased the probability of prothesis relative to trials in which it was absent. This is reflected in the model by a significant interaction between the prothesis response type and POV (*prothesis x POV* = 0.91, *95% HPD* [0.51, 1.32], $p < .001$). This effect can be understood by comparing the predicted rates of prothesis for the two values of the POV factor (present vs. absent). For fricative-nasal clusters, presence of POV

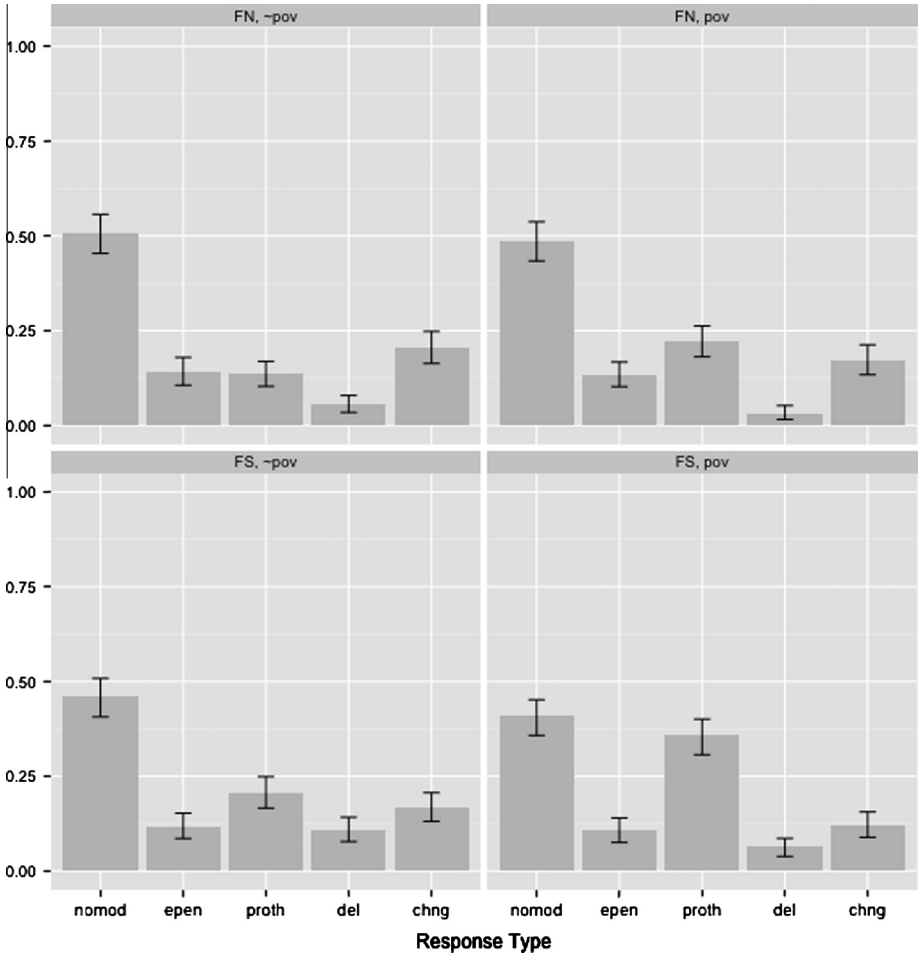

Fig. 3. Proportion of responses for fricative-initial clusters with (pov) and without (~pov) pre-obstruent voicing. Errors bars in this and future figures indicate 95% bootstrap confidence intervals.

more than doubled the fit rate of prothesis (i.e., the coefficients of the multinomial model collectively imply that Pr(proth|POV)/Pr(proth|∼POV) = 0.17/0.07 = 2.43); the effect of POV on prothesis was comparable for fricative-stop clusters (0.33/0.16 = 2.06). Interestingly, presence of POV also significantly lowered the probability of fricative deletion (*C1-deletion x POV* = −0.84, *95% HPD* [−1.52, −0.12], $p < .05$). When it is not interpreted as a prothetic vocoid, an initial modal voicing interval appears to serve as an additional cue to the presence of the fricative.

Cluster Profile had two significant effects on the response patterns. In comparison to fricative-stop clusters, fricative-nasal sequences are less susceptible to both prothesis (*prothesis x Cluster Profile* = −0.86, *95% HPD* [−1.46, −0.25], $p < .001$) and deletion (*C1-deletion x Cluster Profile* = −1.80, *95% HPD* [−3.66, −0.02], $p < .05$). These differences could reflect an asymmetry between the two cluster types in the acoustic–phonetic robustness of the initial consonant, or alternatively indicate that fricative-nasal clusters conform better to the gradient well-formedness pattern of English word-initial clusters (see the General Discussion). The remaining coefficients of the model were non-significant and small in magnitude (<0.3).

Inspection of Fig. 3 suggests an interaction between POV and Cluster Profile, with a larger POV effect for fricative-stop clusters. We think this is likely due to a constellation of small differences between FS and FN clusters, including that (i) FN clusters are numerically more likely to undergo C1-change regardless of whether POV is present and (ii) one FN cluster in particular, /vn/, elicits a high rate of epenthesis (48 responses, or 25% of single modifications for this cluster; cf. no other fricative-initial cluster elicited more than 15 (8%) epenthesis responses). Increases in the rates of C1-change and epenthesis necessarily lower the relative frequency of prothesis. The difference in behavior between vN and zN clusters may be attributable to their relative similarity to legal English onsets. One way of minimally changing zN clusters, namely devoicing the fricative (/zn/ → [sn]), results in sequences that are attested in English, but no single feature repair is available for vN clusters. The high rate of epenthesis for /vn/ could alternatively reflect random variation across the clusters, or perhaps the influence of phonetic detail different from or more fine-grained than investigated here.

To summarize, as anticipated in the design of the current study, English speakers apply prothesis modifications more often when POV is present. In addition to the effect of the acoustic–phonetic POV factor, we found that FS clusters are more likely than FN clusters to undergo certain modification types (prothesis and deletion).

Analysis of stop-initial clusters

Stop-initial clusters were subject to POV, burst amplitude, and burst duration manipulations. In addition to these factors, the model included the fixed factors Cluster Profile (stop-nasal vs. stop-stop) and Cluster Voice (voiceless vs. voiced). Because of the larger number of manipulations and cluster types, statistical analysis of the stop-initial clusters is more involved than for fricative-initial clusters. Therefore, after discussing significant coefficients that apply under all of the phonetic manipulations, we discuss effects of each manipulation separately. We emphasize, however, that a single analysis was performed on all of the data, and the separation is for expository reasons only. All aspects of data analysis were identical to that in the previous section except as indicated below. Results for stop-initial sequences broken down by the amplitude and duration manipulations are shown in Fig. 4, and by the POV manipulation in Fig. 5.

All modification types except epenthesis were estimated to have lower probability than the baseline no-modification response (*prothesis* = −3.50, *95% HPD* [−4.44, −2.45], *C1-deletion* = –3.09, *95% HPD* [−3.77, −2.37], *C1-change* = −1.90, *95% HPD* [−2.44, −1.36], all *ps* < .001). Epenthesis did not differ from no-modification (*epenthesis* = 0.33, *95% HPD* [−0.26, 0.89], n.s.). HPD intervals on differences between the sampled coefficients indicate that epenthesis was more probable than prothesis, C1-change, and C1-deletion; additionally, C1-change was more probable than prothesis and C1-deletion. The finding that epenthesis and correct responses do not differ in their probability is a point of contrast between the stop-initial clusters analyzed here and the fricative-initial clusters considered previously.

In comparison to SS clusters, SN sequences were more likely to be modified by epenthesis (*epenthesis x Cluster Profile* = 0.64, *95% HPD* [0.18, 1.10], $p < .01$) and less likely to undergo deletion (*C1-deletion x Cluster Profile* = −1.88, *95% HPD* [−2.59, −1.13], $p < .001$. (There was also a marginally significant effect of Cluster Profile on C1-change, with SN clusters less likely to undergo this type of modification, *C1-change x Cluster Profile* = −0.64, *95% HPD* [−1.36, 0.10], $p = .08$.) Overall, voiced clusters had higher estimated probabilities of epenthesis and prothesis than voiceless clusters (*epenthesis x Cluster Voice* = 2.36, *95% HPD* [1.86, 2.85], *prothesis x Cluster Voice* = 3.53, *95% HPD* [2.36, 4.52], both *ps* < .001).

POV manipulation

The POV manipulation was applied to clusters beginning with voiced stops. As was the case for fricative-initial clusters, the probability of prothesis increased when POV was present (*prothesis x POV* = 1.60, *95% HPD* [1.01, 2.17], $p < .001$). No other effect involving POV reached statistical significance. Inspection of Fig. 5 suggests that the prothesis-inducing effect of POV on the production of SS clusters weakens as burst duration increases. More specifically, as stimulus burst duration increases the rate of prothesis declines and the rate of epenthesis commensurately increases. This suggests that the burst cue may 'win out' over POV when only one modification is made, a type of cue interaction that we anticipated in the introduction. We discuss possible causes of this interaction in the General Discussion.

Amplitude manipulation

The relative burst amplitude factor was entered into the model as a binary distinction between higher and lower values. This acoustic manipulation had three significant effects. Higher burst amplitude increased the rate of epenthesis (*epenthesis x amp* = 0.44, *95% HPD* [0.18, 0.70],

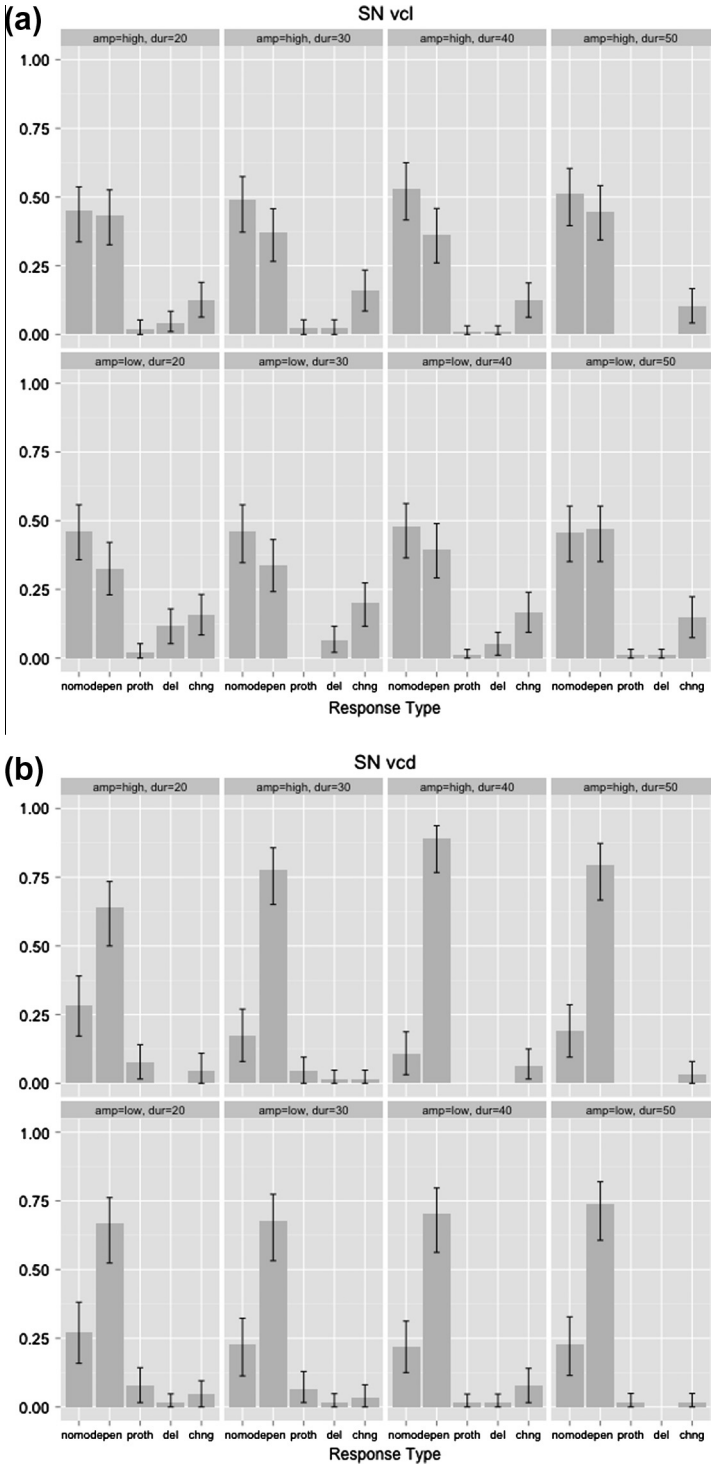

Fig. 4. Proportion of responses for stop-initial clusters with burst duration and amplitude manipulations. (a) Voiceless stop-nasal clusters. (b) Voiced stop-nasal clusters. (c) Voicelesss stop-stop clusters. (d) Voiced stop-stop clusters.

$p < .01$), a finding parallel to that of Wilson and Davidson (2013). The other two effects indicate that, in addition to being interpretable as evidence for a vocoid, higher burst amplitude also serves to protect the initial consonant from modification. Both deletion (*C1-deletion x amp* = −1.93, *95% HPD* [−2.44, −.147], $p < .001$) and feature change

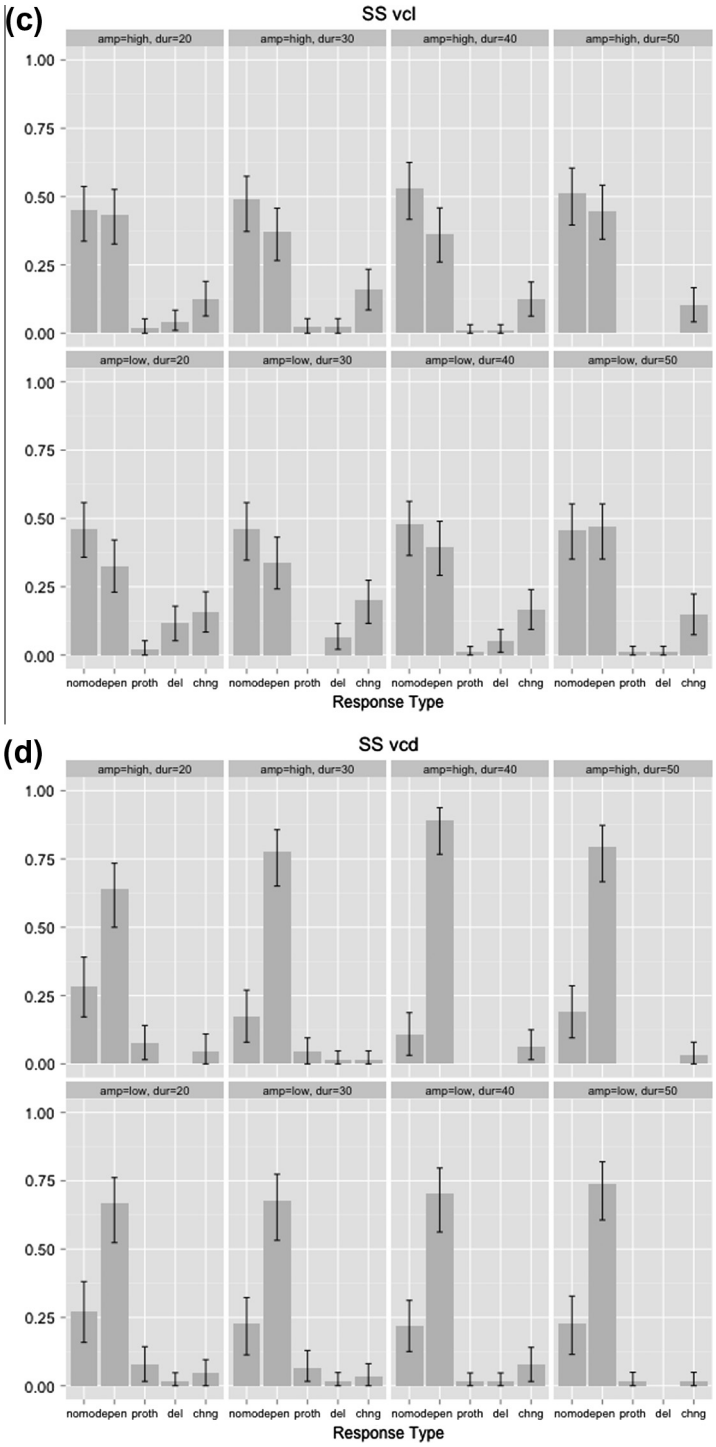

Fig. 4 (*continued*)

(*C1-change x amp* = −0.53, *95% HPD* [−0.93, −0.11], $p < .05$) were less probable for stimuli with higher-amplitude bursts. These effects are illustrated in Fig. 4.

Subsequent investigation of C1-change responses revealed that occurrences of this modification were largely due to the particular cluster /pn/, which was frequently produced as [kn] ($N = 51$). It is likely that the /pn/ → [kn] modification was due to the fact that the higher amplitude bursts in the stimuli, while natural for English /t/ and /k/, are somewhat more intense than is

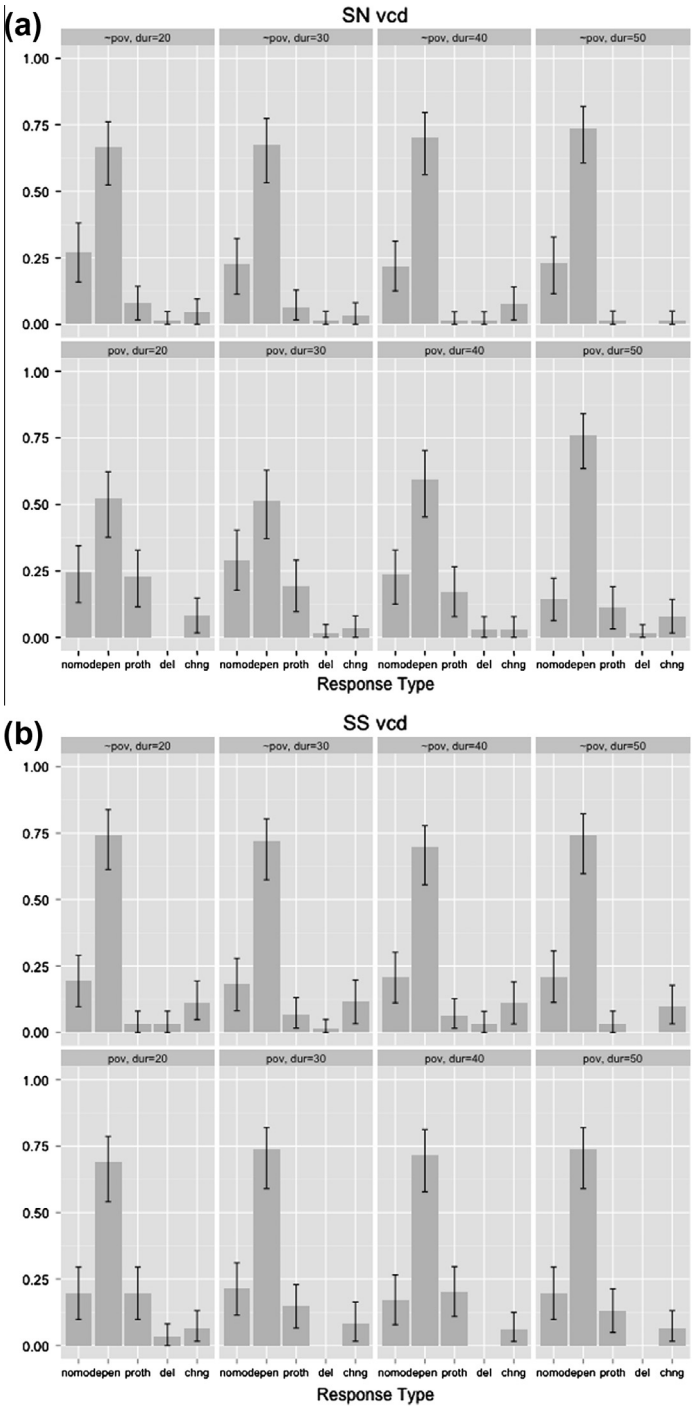

Fig. 5. Proportion of responses for stop-initial clusters with (pov) and without (~pov) pre-obstruent voicing.

typical for English /p/ (Repp, 1984). This modification type is notable both because it provides further evidence that production patterns are heavily influenced by acoustic properties of the stimulus, and because it demonstrates that modifications of nonnative sequences do not uniformly improve phonological well-formedness (as both /pn/ and /kn/ are unattested word-initially in English; see also Davidson & Shaw, 2012 for similar response patterns).

Duration manipulation

We were particularly interested in whether longer burst durations (30–50 ms) would elicit different response patterns from the shortest tested duration (20 ms). The latter

is most similar to that found for word-medial stops of English that appear before another obstruent or nasal consonant (e.g., *actor* or *picnic*; Davidson, 2011). However, release of the first member of a (word medial) stop-initial consonant cluster is relatively rare in English, potentially making even 20 ms bursts susceptible to reinterpretation as containing vocoids. The highest level of burst duration did increase the probability of epenthesis relative to the lowest level (*epenthesis x Dur: 50ms* = 0.43, *95% HPD* [0.02, 0.82], $p < .05$), but other coefficients involving burst duration were non-significant.

The limited effect of the duration manipulation is somewhat surprising in light of the pilot results, but those findings were based on naturally produced bursts in which duration likely covaried with other relevant properties (such as amplitude). To a substantial extent, the influence of bursts on production responses in the present experiment is already reflected in the fact that epenthesis responses to stop-initial stimuli were more prevalent than other modification types and did not differ in their probability from correct responses. Compare this with the earlier findings from fricative-initial clusters, which systematically lack initial-consonant bursts and for which the probability of epenthesis was substantially lower than that of no-modification. The gross acoustic–phonetic disparity between clusters with and without internal releases appears to have largely, though not entirely, overshadowed differences in burst duration.[4]

While the phonetic account of epenthesis in stop-initial clusters is consistent with our general claim that production responses have a perceptual origin, we performed several additional analyses to better understand the role of burst duration specifically. In a first analysis, we crossed Cluster Profile, Cluster Voice, and burst duration in a mixed-effects binomial logistic regression (epenthesis vs. all other response types) with random intercepts and slopes for participants and items. As before, the probability of epenthesis was significantly increased for SN clusters (*Cluster Profile* = 0.95, *95% HPD* [0.31, 1.81], $p < .001$), voiced clusters (*Cluster Voice* = 2.89, *95% HPD* [1.87, 4.79], $p < .001$), and for stimuli with the longest burst duration (*Dur:50ms* = 0.77, *95% HPD* [0.27, 1.47], $p < .05$). Importantly, none of the interactions among burst duration and the other two factors approached significance, indicating that the previous multinomial analysis was not overly simplified. Additional analyses addressing the interaction of burst duration with other acoustic manipulations, and the possibility that what we have coded as vowel epenthesis is in fact phonetic imitation of burst duration, are reported below.

[4] Our previous statistical analyses treated responses to fricative- and stop- initial clusters separately. When the data from all clusters is combined, a logistic regression analysis shows a highly significant effect of the initial consonant type (fricative vs. stop) on the rate of epenthesis (*C1-type* = −3.5, 95% *HPD* [−2.30, −4.57], $p < .001$). This analysis also included fixed factors of cluster voice and second consonant type (nasal vs. stop), as well as random slopes and intercepts for participants and random item intercepts.

Interactions among acoustic manipulations

If phonetic decoding plays an important role in accounting for cross-language production patterns, there should be evidence of well-known perceptual interactions such as phonetic context effects (e.g., Miller & Liberman, 1979) and cue trading or integration (e.g., Repp, 1982). The present experimental design, which crosses the burst duration continuum with the POV and burst amplitude manipulations separately, provides an opportunity to explore such interactions.

Examination of the results for stop-initial items, and in particular SN clusters, indicated that the effect of POV was mitigated by longer burst duration. Specifically, the rate of prothesis declined and (principally) that of epenthesis rose with increasing burst duration. This pattern of cue interaction was especially apparent for voiced SN clusters (as shown by the main analysis above, epenthesis dominates all other response types for voiced SS clusters). The relevant prothesis response proportions are given in Table 5, which also includes values for fricative-initial sequences and SS clusters for purposes of comparison.

To confirm this cue interaction statistically, the rate of prothesis for voiced stop-initial clusters was analyzed with a binary logistic mixed-effects model having Cluster Profile, POV, and burst duration as fixed factors, and random intercepts and slopes for participants and items. Consistent with preceding results, there was a significant bias against prothesis overall (*Intercept* = −3.32, *95% HPD* [−3.99, −2.63], $p < .001$) that was partly counteracted by the presence of POV (*POV* = 1.66, *95% HPD* [1.13, 2.13], $p < .001$). The longest burst duration lowered the probability of prothesis (*Dur:50ms* = −0.79, *95% HPD* [−1.43, −0.09], $p < .01$), approximately halving the effect of POV, an interaction with the duration manipulation that was not made explicit by the main analysis above.

Did burst duration also interact significantly with relative amplitude, the other cue with which it was crossed? Such an interaction could be more difficult to detect, because longer burst duration and higher amplitude have been established to have somewhat parallel effects (both favoring epenthesis) and because independent effects multiply, rather than add, in logistic models. However, while previously we considered duration and amplitude as categorical predictors, a trading or integration relation between them could be better assessed by an analysis in which their raw values are used as predictors.

To this end, we entered burst duration (in ms) and relative amplitude (in dB), together with Cluster Profile and Cluster Voice, into a mixed-effects logistic regression model with epenthesis as the binary dependent variable and random intercepts and slopes for participants and items. The model indicated an overall bias against epenthesis (*Intercept* = −1.06, *95% HPD* [−2.14, −0.07], $p < .05$) and an increase in epenthesis probability for SN clusters (*Cluster Profile* = 1.94, *95% HPD* [0.73, 3.40], $p < .001$) and clusters beginning with a voiced stop (*Cluster Voice* = 3.98, *95% HPD* [2.22, 5.80], $p < .001$). Raw burst duration and relative amplitude both also significantly raised the odds of epenthesis (*Dur* = 0.03, *95% HPD* [0.01, 0.05], $p < .001$; *Amp* = 0.08, *95% HPD* [0.03, 0.14], $p < .001$). The fit

coefficients for the two cues suggest a trading relationship in which an increase of 10 ms in burst duration (the size of the steps in our continua) is equivalent to raising the relative amplitude by approximately 3.75 dB.

Imitation of burst duration

One potential concern with the preceding analyses is that they have been performed entirely on categorically coded responses. In particular, it is conceivable that the responses we have coded as containing epenthesis—and which are more frequent when the stimulus burst duration is longer or burst amplitude is higher—in fact arise from a gradient process of phonetic imitation (Goldinger, 1998; Pardo, 2006) rather than a categorical modification. Perhaps participants were simply intending to mimic the burst durations (or amplitudes) of the stimuli, creating a long transition between consonants that was erroneously coded as containing a vocoid. If it were the case that phonetic imitation operated relatively independently of native phonetic and phonological structure (but cf. Mitterer & Ernestus, 2008; Cole & Shattuck-Hufnagel, 2011), an imitation account of 'epenthesis' responses would be damaging to our main claim that nonnative production data is relevant for understanding phonetic decoding.

We have partly addressed this concern with a coding protocol, similar to those used in several previous studies of native and nonnative speech production, that operationalizes the burst vs. vocoid distinction in terms of acoustic–phonetic properties other than duration. Recall that the central distinction between bursts and vocoids for our purposes is the absence vs. presence of higher formant structure (F2 and above). To verify that higher formants and related properties are systematically absent from the Russian consonant clusters, a balanced subset of the stimuli were coded by a trained research assistant blind to the purpose of the study. The results of this test were definitive, as essentially none of the cluster stimuli were labeled as containing vocoids (less than 1%). It follows that there is at least one qualitative phonetic difference between epenthesis responses and the stimuli that elicited them, and hence that epenthesis as coded here cannot be entirely reduced to phonetic imitation.

An additional analysis was performed to further quantify the phonetic difference between responses with and without epenthesis, and to determine whether any degree of imitation existed in the data alongside categorical modifications. For each stop-initial cluster production that was assigned to either the correct or epenthesis response category, we calculated the total duration of the transition from the initial stop burst to the following consonant (i.e., the release transient and any following aspiration or coded vocoid; see Fig. 6). A mixed-effects linear regression model was fit to the transition durations with coded Response Type (no-modification vs. epenthesis) and stimulus burst duration as crossed fixed factors, and random intercepts and slopes for participants and items. Because phonetic imitation would be most clearly indicated by a trend, with longer stimulus burst durations eliciting long response transitions, Stimulus Burst Duration was entered into this model with orthogonal polynomial coding (linear, quadratic, and cubic terms).

Table 5
Proportion of responses containing prothesis for tokens with and without pre-obstruent voicing (POV).

	Overall proportion of prothesis	Stop burst duration			
		20 ms	30 ms	40 ms	50 ms
FN	0.137	–	–	–	–
FS	0.206	–	–	–	–
SN	0.038	0.079	0.056	0.008	0.008
SS	0.068	0.058	0.058	0.098	0.057
With POV					
FN	0.223	–	–	–	–
FS	0.358	–	–	–	–
SN	0.176	0.230	0.194	0.172	0.111
SS	0.170	0.197	0.145	0.203	0.131

The grand mean response transition duration was estimated to be approximately 50 ms (*Intercept* = 49.74, *95% HPD* [45.51, 53.48], $p < .001$). The largest fixed effect was that of Response Type, with transitions in no-modification responses being about 20 ms shorter, on average, than epenthesis responses (*Response Type* = −20.29, *95% HPD* [−23.54, −17.03], $p < .001$). In addition, there was a linear influence of stimulus burst duration on response transitions in the expected direction (*Dur:Linear* = 6.88, *95% HPD* [5.33, 8.40], $p < .001$). Given the coding of Stimulus Burst Duration, this corresponds to a difference between the effects of the longest and shortest stimulus burst durations (i.e., 50–20 = 30 ms) of less than 10 ms in the elicited response transitions. The quadratic and cubic effects of burst duration were not significant, nor was there any significant interaction between Response Type and Stimulus Burst Duration.

This analysis established that there is a degree of phonetic imitation in our data (i.e., an effect of burst duration within the coded response categories). However, and most

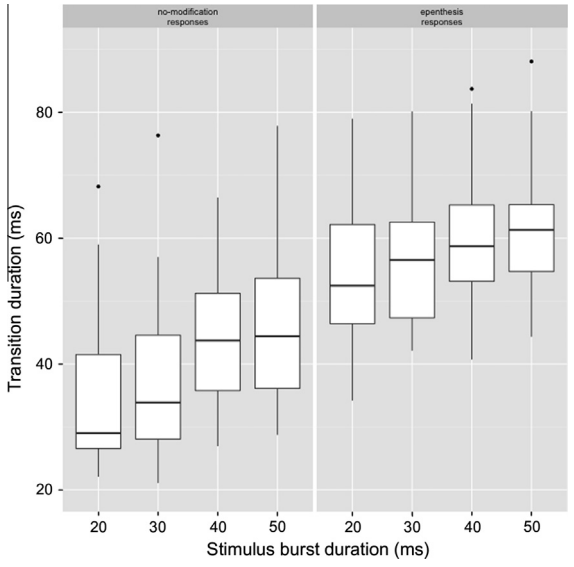

Fig. 6. Duration of the inter-consonant transition for stop-initial tokens with no modification or epenthesis.

importantly for our purposes, the difference between correct and epenthesis responses cannot be explained away by the imitation effect. The fact that imitation was only partial, with responses being constrained by phonetic or phonological categories more than would be expected from pure mimicry, agrees with many previous studies (Nielsen, 2011; Zellou, Scarborough, & Nielsen, 2013). Previous studies of cross-language phonetic imitation have generated mixed results (Flege & Eefting, 1988; Oh & Redford, 2012; Olmstead, Viswanathan, Aivar, & Manuel, 2013; Yeni-Komshian, Caramazza, & Preston, 1977). It is likely that immediate repetition tasks such as the one employed here highlight phonetic properties—and mitigate the influence of dialectal and sociolinguistic affinity (Abrego-Collier, Grove, Sonderegger, & Yu, 2011; Babel, 2012; Kim, Horton, & Bradlow, 2011), shared lexical experience (Johnson, 2006), and other metalinguistic factors (see Chang, 2012 for a review)—relative to tasks involving meaningful communication or language learning, which may elicit weaker imitation effects.

Summary

The preceding analyses have provided extensive evidence for the influence of the acoustic manipulations, indicating an important role for perception in our production task. When a cue such as POV is present in a nonnative cluster, speakers often interpret it as evidence for a reduced vowel and produce a structure that is phonologically permissible in their language. Conversely, participants are more likely to produce nonnative clusters correctly to the extent that such cues are absent or attenuated. We now discuss the implications of these findings for the process of phonetic decoding in particular and (cross-language) speech perception and production more generally, returning to the issues raised in the Introduction.

General discussion

This study has provided evidence about how low-level, subphonemic phonetic detail influences the rates and types of modifications in cross-language speech production. The focus on phonetic detail contrasts with much previous research in nonnative production and perception, which has emphasized phonological motivations for production modifications such as prothesis or epenthesis (Broselow & Finer, 1991; Hancin-Bhatt & Bhatt, 1997) and perceptual confusions such as an inability to distinguish a nonnative sequence from one with a vowel inserted before or between the sequence (Berent, Lennertz, & Balaban, 2012; Berent et al., 2008, 2007; Dupoux et al., 1999, 2011; Hallé et al., 2008; Kabak & Idsardi, 2007). By deliberately manipulating the acoustic properties that were found to be relevant in Wilson and Davidson (2013), a fuller picture emerges of how participants interpret the acoustic stimulus for the purpose of production. In the following sections, we discuss possible phonetic and phonological sources of the speakers' modifications, return to the questions about phonetic decoding that motivated the study, and conclude by discussing broader implications of our findings and ways in which they could be extended in the future.

Effects of acoustic manipulations

The acoustic manipulations studied here had highly specific effects on nonnative cluster production. The first findings that we discussed involve the pre-obstruent voicing (POV) manipulation for stop- and fricative-initial clusters. Recall that the POV manipulation was motivated by the observation that, in the natural tokens produced by a Russian speaker in Davidson (2010), some voiced fricative-initial items contained a short interval of voicing (without visible formants) that preceded the onset of the frication and some stop-initial clusters had a short interval of higher-amplitude voicing at the onset of the stop closure. The occurrence of strong, early voicing within an obstruent following a pause is rare in English fricatives (e.g., Haggard, 1978; Smith, 1997) and stops (e.g., Keating, 1984; Lisker & Abramson, 1964). Because English speakers have little experience with the perception and production of such voicing patterns, and because the voicing cue is consistent with the presence of a short, schwa-like English vowel before the consonant, we expected and found that the rate of prothesis increases when POV was present. The potential of misinterpreting two phonetic components of a single phoneme as evidence for two separate segments is a general prediction of cue parsing models of speech perception (e.g., Gow, 2003), and has previously been observed for other voiced obstruents (e.g., Hallé, Segui, Frauenfelder, & Meunier, 1998; Solé, 2014).

In addition to the increase in prothesis, two related results were obtained. First, C1-deletion was the least likely modification for fricative-initial sequences, which reflects the well-known fact that fricatives are particularly resistant to deletion, plausibly because the frication noise serves as an 'internal' cue to their presence (e.g., Wright, 2004). Second, within the fricative-initial sequences, we found that FS clusters were nevertheless more likely than FN clusters to undergo deletion (as opposed to C1-change). In the analysis of fricative-initial clusters, we suggested that this aspect of the data is possibly a phonological (phonotactic) reflex: clusters that cannot be 'repaired' by a single feature change are somewhat more likely to undergo deletion. Taken together, the results for the fricative-initial sequences support our general point that reference to abstract phonological structure (such as the phonemes in a syllable onset) may be necessary to account for cross-language processing of speech, but is not sufficient. The specificity of the relationship between cue manipulation and nonnative 'repair' demonstrates that subphonemic phonetic detail, which can differ systematically across languages, also plays an important role in explaining production outcomes.

We return to the effect of the duration manipulation below, but at this point it is worth mentioning the finding that for stop-initial stimuli with POV, the proportion of prothesis generally decreased as stop burst duration increased. One interpretation of this result would be to attribute it to cue competition: perhaps the longest burst

duration was more perceptually salient than the POV cue, causing the former to mask the latter. Or perhaps English speakers were reluctant to produce two reduced vocoids at the beginning of a word (e.g., [ᵊdᵊnake]), given that native words never begin with consecutive schwa syllables (Hayes, 1984). When there are two cues that support different loci of insertion, the longer and presumably more robust cue could dominate responses. Alternatively, it may be that participants used the duration of the transition as an estimate of speech rate, in which case the POV may be perceptually shorter—and hence less likely to indicate a reduced vowel—at higher duration values.[5] Adjudication among these possibilities awaits future research on the perception and production of nonnative clusters.

It can be further observed that competition between the POV manipulation and increased burst length was clearest for SN sequences. This is consistent with the general pattern that SN sequences were modified with epenthesis more often than SS sequences even when the stimulus contained only one manipulated cue. Because nasals have voicing and weak formant structure, the overlap between a stop burst and a nasal is generally more acoustically similar to a vowel than the transition in a stop-stop cluster. Since SN clusters are independently more susceptible to epenthesis in the single-cue cases, it is natural that the diminishing effect of POV as burst duration increases would be most evidence for this cluster type.

Returning to the duration manipulation alone, the pilot study of Wilson and Davidson (2013) observed that longer burst durations led speakers to epenthesize more often. This result was broadly confirmed in the current study. Collapsing over all stop-initial sequences that were not manipulated for POV or amplitude (i.e., higher amplitude bursts for SS and lower amplitude bursts for SN), there was significantly less epenthesis at the shortest duration of 20 ms than at the longest duration of 50 ms. The amount of epenthesis for the intermediate duration values was not significantly different from the endpoints, though there is evidence of a linear increase from 20 ms to 50 ms.

A closer look shows that the effect of duration was more pronounced for voiced sequences. Across the board, there was more epenthesis for voiced stop-initial sequences than for voiceless ones; for both SN and SS, epenthesis responses exceeded correct productions for voiced sequences. This finding is expected, since the voiced bursts of these stimuli contain acoustic information that is simultaneously periodic and aperiodic. Since vowels consist of periodic waveforms, listeners are more likely to interpret voiced bursts than voiceless bursts (which have aperiodic energy only) as a vocoid. Moreover, the longer a voiced burst is, the more similar its acoustic profile is to that of a vowel. This raises the question of how even voiceless bursts led to epenthesis responses in 20% to 40% of responses. A possible answer is found in the casual phonetic reduction pattern of English. For example, Davidson (2006b) showed that English speakers can overlap the burst of a stop with a following schwa in words like ‘tomato’ or ‘potato’, such that the schwa portion is realized as voiceless. In the acoustic record, such sequences appear as a silence followed by a lengthened interval of burst plus aspiration. It is plausible that when presented with stimuli containing burst + aspiration in the current study (e.g., [kʰpago]) the participants interpreted this as containing a short devoiced vowel ([kʰə̥pago]), which they then produced with a voiced vocoid ([kʰᵊpago]).

In addition to the burst duration manipulation, stop-initial sequences were also subject to a manipulation of burst amplitude. In natural productions, the burst of a stop relative to a following stop is high in amplitude, whereas it is low in amplitude relative to a following nasal. Preliminary observations of an amplitude effect in Wilson and Davidson (2013) are supported by perceptual results in Davidson and Shaw (2012), who found that discrimination confusions between clusters (especially SN) and either their deleted (e.g., [tmaba∼maba]) or changed (e.g., [tmaba∼kmaba]) counterparts may be due to difficulty in perceiving a low amplitude burst, or in accurately perceiving the information about place in the burst. It is well-known that both stop bursts and formant transitions in the following vowel carry important information regarding the place of articulation of the stop (e.g., Blumstein & Stevens, 1978; Dorman, Studdert-Kennedy, & Raphael, 1977; Ohde & Stevens, 1983; Repp, 1984), so when stops are not followed by vowels that provide formant transitions, the probability of accurately perceiving the stop decreases. To test this hypothesis, we increased the amplitude of bursts in SN sequences; to create a comparably low-quality environment for SS sequences, burst amplitudes were decreased.

The results for the amplitude manipulations showed that in general, there was significantly more deletion and C1-change for lower amplitude bursts and more epenthesis for higher amplitude bursts. Because high amplitude relative to surrounding consonants is characteristic of vowels, it is sensible from an acoustic–phonetic perspective that increased burst amplitude would result in more epenthesis (at the expense of correct responses). The decrease in deletion responses is consistent with the perceptual status of bursts as critical cues for the detection and identification of stops. A word-initial stop with a low amplitude burst is vulnerable to misinterpretation as having a different place of articulation, or to being missed entirely in perception (perhaps being mistaken for noise in the stimulus recording)—and for that reason not realized in production.

While the amplitude manipulation had a similar effect on SS clusters regardless of voicing, SN clusters showed an increase in epenthesis only for voiced clusters. This discrepancy could be due to the fact that voiceless SN sequences are the only ones in which the voicing of C1 does not match C2. If a following nasal naturally causes the amplitude of the burst to increase in this environment (i.e., by anticipatory velum lowering), the experimental increase that we chose may not be sufficiently extreme. Put differently, participants may not have been able to distinguish between an increased burst amplitude due to overlap with a voiced nasal, in comparison to an increase of the burst noise itself. For the voiced cases, it seems that they could apparently tell when an already voiced burst is made louder in the experimentally manipulated tokens.

[5] We thank Lori Repetti (p.c.) for suggesting this possibility to us.

Like the effects of burst duration and pre-obstruent voicing, participants' responses to the amplitude manipulation are easily accounted for by appealing to the interpretation of acoustic–phonetic properties. Like the duration manipulation, increased amplitude between two consonants increases the transition's similarity to a reduced vowel. On the other side of the coin, decreased amplitude may be difficult to hear. The different modifications that these manipulations elicited (epenthesis for increased amplitude and C1 change and deletion for decreased amplitude) are straightforwardly explained by appealing to speech perception.

The results of this study make it evident that manipulations of the acoustic–phonetic cues in the stimuli qualitatively change participants' responses. However, it should also be pointed out that even in the baseline conditions—POV absent, short stop burst, baseline relative burst amplitude (higher for SS, lower for SN)—speakers still made a substantial number of modifications. While the fact that speakers do correctly produce these nonnative sequences some proportion of the time is an important clue in understanding the phonetic decoding process (see further discussion in 'Accurate and modified output patterns in nonnative speech production'), we nevertheless would not expect that speakers could consistently reproduce them faithfully even when subphonemic phonetic cues are least likely to bias the perceptual interpretation toward English-legal structures. In addition to acoustic cues, there are other potential sources of these repairs, such as top-down phonological influences (see 'Relative contributions of phonetics and phonology to nonnative cluster processing') and difficulties with articulatory coordination. The influence of articulatory factors has been discussed in other papers (Davidson, 2005, 2010; Ussishkin & Wedel, 2003; Yanagawa, 2006; Zsiga, 2003), and we would argue that articulation certainly contributed to some of the modifications seen in our productions results. However, our focus in this paper is on speakers' sensitivity to subphonemic phonetic variation above and beyond the contribution of phonological or articulatory factors. The significant differences between the amount and type of modifications in the baseline conditions and the conditions with the duration, amplitude and POV manipulations provide evidence for our argument that sensitivity to sub-phonemic phonetic detail affects how speakers will interpret and produce nonnative sound structures.

Phonetic detail: native language cues vs. language-general speech processing

It has been implicit in the preceding sections that the responses of the English-speaking participants were primarily influenced by their native language-specific interpretation of the acoustic cues in the stimuli. Because the phonology of English does not allow obstruent-obstruent or obstruent-nasal word-initial sequences, English listeners may be predisposed to 'over-interpret' any phonetic cues that are present as evidence for alternative, phonologically legal structures. Thus, for example, the onset of voicing before the start of frication is interpreted as a vocoid (i.e., the prothesis modification), resulting in a word that is at least phonotactically permissible in English (e.g., [ᵊzmabe]).

While it makes sense to analyze speakers' modifications as effects of language-specific cue interpretation, it is also possible that some modifications reflect general properties of speech perception. For example, that decreased burst amplitude led to greater deletion rates and numerically larger C1-change rates likely does not reflect something particular to English, but would be found regardless of the participants' language background. The explanations for the effects of burst duration and increased amplitude follow from the fact that English has reduced vowels and that they can appear pretonically between consonants (e.g., t[ə]mórrow, p[ə]táto), which is the same environment as the stop bursts in our critical stimuli. It would be interesting to examine the responses of speakers of other languages to the manipulated stimuli to help shed light on the relative contribution of language-specific and language-general cue processing. For example, one might contrast the results for English speakers to speakers of a language that has a subset of the obstruent-obstruent or obstruent-nasal sequences permitted in Russian, such as Serbian or Croatian (Morelli, 1999). If language-specific phonetic interpretation is paramount, we might expect that because the Serbian/Croatian speakers have experience with at least some stop-obstruent sequences, they would be less likely to interpret longer or higher amplitude bursts as a vowel than English speakers do.

It would also be informative to examine speakers of a language that has neither obstruent-obstruent/nasal initial clusters nor any (phonemic) reduced vowels, such as Spanish (see Berent, Lennertz, & Rosselli, 2012 for relevant perceptual findings). For speakers of such a language, longer burst durations and higher burst amplitudes would not provide as good a match to vowels as they do to the short and highly variable reduced vowels of English (e.g., Davidson, 2006b; Flemming & Johnson, 2007; Silverman, 2011). It remains to be seen whether Spanish speakers would nevertheless insert a full vowel, whether they would 'innovate' a reduced vowel to preserve attested phonotactics but minimize the inserted material, or whether they would opt for different modifications altogether (e.g., deletion). Finally, to briefly return to the amplitude manipulation, if the misperception of lower amplitude bursts is a property of language-general perceptual processing, then it would be predicted that speakers of all backgrounds would have increased deletion and maybe C1-change, just as English speakers do.

Relative contributions of phonetics and phonology to nonnative cluster processing

In this paper, we have emphasized—and our manipulations have robustly supported—acoustic–phonetic sources of nonnative production modifications. However, because all of the clusters tested in our experiment are unattested word-initially in English, it is possible that some of our results are due to more abstract phonological principles, such as sonority sequencing or other markedness constraints. Indeed, our statistical analyses included

phonological factors as potential predictors (i.e., cluster type and phonological voicing specification). As such, it is worth considering the extent to which the present results reflect the effect of phonological knowledge. Because phonological and phonetic properties necessarily covary to a certain extent, effects that appear to be due to phonology may in fact be better explained by phonetic detail (or, of course, vice versa). Indeed, we have already offered potential phonological explanations for certain results; for example, our speculation that certain FN sequences may be repaired by C1-change more often than FS sequences because it requires fewer feature changes to obtain an attested English cluster (/zn/ → [sn] only affects one consonant whereas /zd/ → [st] would change two) is a phonologically-oriented explanation.

Another relevant result is that there was more prothesis overall for fricative-initial sequences than for stop-initial sequences. Such a finding is consistent with a well-known cross-linguistic pattern in which prothesis—particularly in comparison to epenthesis—is a common 'repair' for fricative-initial clusters in loanword adaptation (Broselow, 1992; Fleischhacker, 2005; Gouskova, 2004; Zuraw, 2007). On the basis of perceptual studies, Fleischhacker (2005) and Zuraw (2007) argued that this is due to the greater perceptual similarity between FC to °FC in comparison to F°C. The relative perceptual similarity of a cluster and various modifications may have phonological reflexes (e.g., Steriade, 2001/2009), but we think it would be premature to conclude that the present finding was due to a phonological grammar that favors prothesis over epenthesis for fricative-initial clusters specifically. The perceptual similarity factor itself, which is rooted in the lack of a release between fricatives and following consonants, could suffice. Nonnative speakers could be less likely to perceive an epenthetic vowel within a fricative-initial cluster, making them more likely to adopt a different repair, or they could be actively shaping their response distributions to maintain perceptual similarity with the stimulus. In either case, phonetics rather than phonology would account for the prothesis distribution.

In relevant studies on the perception of nonnative consonant clusters, one phonological factor that has been offered as an explanation for the poor discrimination of onset CC sequences is sonority sequencing (Berent et al., 2007, et seq.). However, in the current study, the substantial effect of voicing on accuracy, which has also been found in previous studies of cross-language speech production (e.g., de Jong & Park, 2012), calls into question the extent to which sonority sequencing can account for nonnative consonant cluster processing. In the phonology literature, perhaps the most widely-accepted scale is that of Clements (1990), which treats glides as most sonorous, followed by liquids, nasals, and then obstruents. Clements' scale does not make a distinction between stops and fricatives, nor does it distinguish between voiced and voiceless obstruents. Using this scale, an explanation of experimental performance which takes sonority sequencing to be a major contributor to accuracy predicts that there should be no difference between voiced and voiceless SS or voiced and voiceless SN sequences. Yet, recall that collapsing over the duration manipulation, voiceless sequences for both SN and SS had significantly more accurate productions than voiced sequences did (SN voiceless: 46.3%, SN voiced: 23.3%, SS voiceless: 60.2%, SS voiced: 19.1%).

Alternative scales to the ones proposed by Clements (1990) do sometimes make divisions on the basis of voicing such that voiced obstruents are more sonorous than voiceless ones (Foley, 1972; Selkirk, 1984), which would give voiced SN sequences (e.g., /dmabe/) slightly worse sonority profiles than voiceless SN sequences (e.g., /tmaba/). Even if such a sonority scale were adopted, it would still not account for the voicing-related difference for the SS sequences, which are both sonority plateaus regardless of the voicing value. Instead, both SN and SS sequences behave similarly because listeners interpret the acoustic information in the (lengthened) voiced burst the same way for both types of sequences. This explanation is more parsimonious than a sonority-based explanation, which can only weakly account for SN sequences, and only if a more detailed and possibly language-specific sonority scale is adopted.

Accurate and modified output patterns in nonnative speech production

A central finding of our study is that illegal consonant clusters are not invariably 'repaired' in production, with correct production rates ranging from approximately 25% to over 50% across cluster types (see Figs. 3 and 4). As discussed in the Introduction, correct production of an auditorily-presented stimulus—in the absence of supporting orthographic, lexical, or other information—can provide critical information about the operation of phonetic decoding. The most straightforward interpretation of accurate production is that nonnative clusters can be decoded faithfully (i.e., as in the target language), and that faithful cluster representations can survive intact through the processing levels of Fig. 1. Under this view, phonetic decoding is not restricted to producing representations that are phontactically legal in the native language. Furthermore, accurate perceptual encoding of nonnative inputs is not completely obscured by conversion to phonological and articulatory codes. While it is undeniable that the native sound system significantly shapes both perception and production, these findings establish a limit on the influence of native categories and constraints at all levels of the assumed processing architecture.

Some previous experimental findings could also be interpreted as demonstrating faithful phonetic decoding of nonnative stimuli, but appear more ambiguous on close inspection. In perceptual tasks such as identification and discrimination, above-chance performance on nonnative stimuli could be due to representations at an earlier level of processing (Berent et al., 2007, et seq.; Peperkamp & Dupoux, 2003; Pisoni & Tash, 1974). Such findings would thus support a type of representation that is (relatively) independent of language-specific categories, but would not bear directly on phonetic decoding or its relation to subsequent processes. Alternatively, it could be that correct responses in such tasks are not due to faithful perception at any level, but rather to category goodness

differences among stimuli (Best, 1995; Best et al., 2001; Iverson & Kuhl, 1996; Miller, 1994). For example, if several acoustically-different /ebzo/ stimuli are all perceived as containing a vowel between the two consonants, but some perceived vowels are deemed less typical of the relevant native language category, listeners could respond 'vowel' less often for atypical instances (perhaps on the basis of task expectations that roughly half of the stimuli should contain vowels). Breen et al. (2013) explicitly acknowledge the possible role of category goodness in explaining electrophysiological data that might otherwise indicate faithful phonetic decoding of nonnative clusters. Moreover, the results and interpretation of perception studies are generally dependent on a small set of response alternatives, and are thus only weakly diagnostic of perceptual representations. For example, an English speaker that fails to perceptually represent the initial stop of Russian /dnif/ (e.g., because the burst is short or low-amplitude) could correctly categorize the stimulus as monosyllabic—but clearly this should not be taken as evidence of faithful phonetic decoding of the illegal cluster.

Previous cross-language production studies also point to the possibility of faithful decoding of nonnative structures, but do not show a close connection between stimulus details and production response patterns in the way that we have done. Building on the pilot results of Wilson and Davidson (2013), we found that performance on nonnative consonant clusters is variable but lawful and stimulus-locked. Correct productions were not haphazard, as would be expected if phonetic decoding succeeded in faithfully representing nonnative inputs in a stimulus-independent manner. Instead, productions without modification are more probable when the stimulus lends itself better to perceptual interpretation as a cluster. Furthermore, the production modifications that did occur were highly sensitive to fine-grained details of the auditory stimulus. One clear methodological consequence of our results is that collapsing across items, in the analysis of any task involving auditory stimuli, can obscure predictable variation that is due to phonetic structure below the phonemic level.

Conclusion

This study has demonstrated that low-level, subphonemic phonetic detail influences the production of nonnative sequences, and thereby sheds light on the role of phonetic decoding in cross-language perception and production. In the absence of additional information about target structures, English speakers relied heavily on fine-grained acoustic properties to interpret and reproduce nonnative inputs. The results indicate that the process of phonetic decoding does not invariably map nonnative inputs to forms that are phonotactically legal in the listener's native language, and furthermore that phonotactically illegal representations can be preserved by downstream processes and faithfully articulated. There was a substantial proportion of correct responses, especially at the baseline levels of our acoustic manipulations, suggesting that the signal in such cases did not contain sufficient evidence to consistently support perceptual repairs. The rates at which different nonnative stimuli were modified in production, and the observed types of modification, are also largely explainable in terms of phonetic cue interpretation. An initial interval of modal or higher-amplitude voicing makes prothetic responses more probable; stop bursts that are longer or higher in amplitude are more likely to be interpreted as burst + vocoid sequences, resulting in epenthesis; weak bursts provide less information about the presence and features of stops, leading to deletion and other modifications of cluster-initial consonants.

These results are consistent with previous research in the areas of cross-language speech perception, second language acquisition, and loanword adaptation, all of which have found evidence of variable transfer of native language patterns to nonnative structures. It remains unclear how to apportion the influence of purely phonological effects (such as a preference for native syllable types) and phonetic cue interpretation (which is also 'warped' by the native sound pattern). While native language influences are integral to understanding many of the responses in this study, there was somewhat less evidence for abstract and putatively universal phonological effects, such as sonority sequencing (cf. Berent et al., 2007 *et seq*; see also Daland et al., 2011 for related claims). Importantly, sequences that would be categorized as having the same sonority sequencing profile (e.g., voiced stop-stop and voiceless stop-stop sequence) led to very different patterns of modifications in this study. The present study therefore motivates further research on the contribution of acoustic cues to cross-language speech processing in general, a line of inquiry that may shift the balance of explanation from abstract phonological principles stated at the level of phonemes and syllables to interpretive processes operating on subphonemic cues (see also Henke, Kaisse, & Wright, 2012 for related discussion of typological data).

To better understand the relative contribution of phonological knowledge and cue interpretation, we look forward to computational models of phonetic decoding that combine them (see also Dupoux et al., 2011 for similar ideas). In Bayesian terms familiar from automatic speech recognition systems and previous studies of human speech perception, such a model could integrate knowledge of phonotactics with knowledge of phonetic realization as follows. Phonotactic knowledge (or more generally phonology) could function as a *prior* over phonological representations, while the *likelihood function* would measure the phonetic similarity between nonnative stimuli and expected realizations of phonological structures. In such a model, a sufficiently unambiguous stimulus-based likelihood can overwhelm even a strong phonological prior, resulting in faithful perceptual representation of illegal sounds and sequences. A goal of future research is to understand the contribution of the phonological and phonetic components of such a model to the processing of nonnative structures under a broad range of conditions, especially those that decrease the availability of acoustic–phonetic information (such as real-world loanword adaptation or communication in noisy environments).

Acknowledgments

The authors would like to thank Alice Hall, Francesca Himelman, Johnny Mkitarian, and Elizabeth George for their assistance in coding the data. We would also like to thank the associate editor, three anonymous reviewers, and the members of the NYU Phonetics and Experimental Phonology Lab and the JHU Phonology/Phonetics Lab for comments that significantly improved this paper. This research benefited from discussions with audiences at the University of Pennsylvania, the University of Delaware, Stony Brook University, the JHU IGERT Workshop, and the Laboratory Phonology meeting in Stuttgart, Germany. This research was supported by NSF grants BCS-1052855 to Lisa Davidson and BCS-1052784 to Colin Wilson.

Appendix A. Stimuli

Sequence	Initial cluster	CC item	CVC item	VCC item
Fricative + Nasal	vm	vmado	vemafu	evmado
		vmafu	vemage	evmati
		vmage		
		vmati		
	vn	vnabe	venago	evnabe
		vnadu	venaza	evnadu
		vnago		
		vnaza		
	zm	zmabo	zemagi	ezmabo
		zmagi	zemasa	ezmaku
		zmaku		
		zmasa		
	zn	znade	zenagu	eznade
		znagu	zenapo	eznaka
		znaka		
		znapo		
Fricative + Stop	vd	vdafi	vedafi	evdagu
		vdagu	vedato	evdapa
		vdapa		
		vdato		
	vg	vgabu	vegafi	evgabu
		vgafi	vegase	evgaka
		vgaka		
		vgase		
	zb	zbafo	zebafo	ezbata
		zbase	zebase	ezbavi
		zbata		
		zbavi		
	zg	zgade	zegade	ezgaku
		zgafa	zegafa	ezgapi
		zgaku		
		zgapi		
Stop + Nasal (Voiced)	bn	bnadi	benadi	ebnate
		bnapa	benapa	ebnazo
		bnate		
		bnazo		
	dm	dmabe	demago	edmabe
		dmago	dematu	edmasa
		dmasa		
		dmatu		
	gm	gmafu	gemato	egmafu
		gmape	gemava	egmape
		gmato		
		gmava		
	gn	gnake	genavo	egnake
		gnatu	genazi	egnatu
		gnavo		
		gnazi		
Stop + Nasal (Voiceless)	km	kmabi	kemapo	ekmabi
		kmapo	kemazu	ekmave
		kmave		
		kmazu		
	kn	knadu	kenadu	eknago
		knafe	kenafe	eknapi
		knago		
		knapi		
	pn	pnabu	penabu	epnaso
		pnaso	penata	epnave
		pnata		
		pnave		
	tm	tmaba	temaba	etmado
		tmado	temafe	etmavu
		tmafe		
		tmavu		
Stop + Stop (Voiced)	bd	bdafa	bedafa	ebdate
		bdaki	bedaki	ebdazo
		bdate		
		bdazo		
	db	dbagi	debagi	edbapu
		dbapu	debazo	edbate
		dbate		
		dbazo		
	gb	gbadi	gebake	egbadi
		gbake	gebaso	egbavu
		gbaso		
		gbavu		
	gd	gdape	gedasu	egdape
		gdasu	gedaza	egdavi
		gdavi		
		gdaza		
Stop + Stop (Voiceless)	kp	kpabi	kepabi	ekpaga
		kpaga	kepazu	ekpavo
		kpavo		
		kpazu		
	kt	ktada	ketada	ektapu
		ktapu	ketasi	ektaze
		ktasi		
		ktaze		

(*continued on next page*)

Appendix A. Stimuli (*continued*)

Sequence	Initial cluster	CC item	CVC item	VCC item
	pt	ptage	petage	eptako
		ptako	petava	eptasi
		ptasi		
		ptava		
	tp	tpabe	tepabe	etpada
		tpada	tepaki	etpafo
		tpafo		
		tpaki		

Appendix B. Mean Amplitudes of the bursts of stop-initial stimuli in dB. Standard deviations in parentheses

Cluster type	Relative amplitude	Voiceless C1	Voiced C1
Stop-Nasal	High	66.45 (1.44)	74.73 (1.68)
	Low	61.62 (1.75)	68.43 (1.67)
Stop-Stop	High	58.39 (2.76)	65.10 (1.49)
	Low	48.43 (2.71)	60.04 (1.34)

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2014.08.001.

References

Abrego-Collier, C., Grove, J., Sonderegger, M., & Yu, A. (2011). Effects of speaker evaluation on phonetic convergence. In *Proceedings of the international congress of phonetic sciences XVII* (pp. 192–195). Hong Kong: International Congress of Phonetic Sciences.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*(1), 177–189.

Beckman, M. (1996). When is a syllable not a syllable? In T. Otake & A. Cutler (Eds.), *Phonological structure and language processing: Cross-linguistics studies* (pp. 95–123). New York: Mouton de Gruyter.

Berent, I. (2008). Are phonological representations of printed and spoken language isomorphic? Evidence from the restrictions on unattested onsets. *Journal of Experimental Psychology: Human Perception & Performance, 34*(5), 1288–1304.

Berent, I., Lennertz, T., & Balaban, E. (2012). Language universals and misidentification: A two-way street. *Language and Speech, 55*(3), 311–330.

Berent, I., Lennertz, T., Jun, J., Moreno, M., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences, 105*(14), 5321–5325.

Berent, I., Lennertz, T., & Rosselli, M. (2012). Universal linguistic pressures and their solutions: Evidence from Spanish. *The Mental Lexicon, 7*(3), 275–305.

Berent, I., Lennertz, T., Smolensky, P., & Vaknin-Nusbaum, V. (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology, 26*(1), 75–108.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition, 104*(3), 591–630.

Best, C. (1995). A direct-realist view of cross-language perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore: York Press.

Best, C., McRoberts, G., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America, 109*(2), 775–794.

Blumstein, S., & Stevens, K. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*(5), 1358–1368.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (Version 5.3.56).

Breen, M., Kingston, J., & Sanders, L. (2013). Perceptual representations of phonotactically illegal syllables. *Attention, Perception and Psychophysics, 75*(1), 101–120.

Broselow, E. (1992). Transfer and universals in second language epenthesis. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (revised ed., pp. 71–86). Amsterdam: John Benjamins.

Broselow, E., & Finer, D. (1991). Parameter setting in second language phonology and syntax. *Second Language Research, 7*(1), 35–59.

Chang, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics, 40*(2), 249–268.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In M. Beckman & J. Kingston (Eds.), *Papers in laboratory phonology 1* (pp. 283–333). Cambridge, MA: Cambridge University Press.

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

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology, 28*(2), 197–234.

Davidson, L. (2005). Addressing phonological questions with ultrasound. *Clinical Linguistics and Phonetics, 19*(6/7), 619–633.

Davidson, L. (2006a). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics, 34*(1), 104–137.

Davidson, L. (2006b). 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. *Journal of Phonetics, 38*(2), 272–288.

Davidson, L. (2011). Characteristics of stop releases in American English spontaneous speech. *Speech Communication, 53*(8), 1042–1058.

Davidson, L., & Shaw, J. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics, 40*(3), 234–248.

de Jong, K., & Park, H. (2012). Vowel epenthesis and segment identity in Korean learners of English. *Studies in Second Language Acquisition, 34*(1), 127–155.

Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: A cross-linguistic study. *Journal of Cognitive Neuroscience, 12*(4), 635–647.

Dell, G., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*(2), 149–195.

Dorman, M., Studdert-Kennedy, M., & Raphael, L. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context dependent cues. *Perception & Psychophysics, 22*(2), 109–122.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1568–1578.

Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J. (1997). A destressing "deafness" in French? *Journal of Memory and Language, 36*(3), 406–421.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language, 64*(3), 199–210.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171–185.

Ellis, A., & Young, A. (1988). *Human cognitive neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Escudero, P., Simon, E., & Mitterer, H. (2012). The perception of English front vowels by North Holland and Flemish listeners: Acoustic similarity predicts and explains cross-linguistic and L2 perception. *Journal of Phonetics, 40*(2), 280–288.
Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.
Flege, J., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *Journal of the Acoustical Society of America, 83*(2), 729–740.
Fleischhacker, H. (2005). *Similarity in phonology: Evidence from reduplication and loan adaptation*. Unpublished Ph.D. Dissertation, UCLA, Los Angeles.
Flemming, E., & Johnson, S. (2007). Rosa's roses: Reduced vowels in American English. *Journal of the International Phonetic Association, 37*(1), 83–96.
Foley, J. (1972). Rule precursors and phonological change by meta-rule. In R. Stockwell & R. Macaulay (Eds.), *Linguistic change and generative theory* (pp. 96–100). Bloomington: Indiana University Press.
Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49*(3), 396–413.
Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.
Gelman, A., & Hill, J. (2006). *Data Analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.
Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition, 102*(2), 219–260.
Gouskova, M. (2004). Relational hierarchies in Optimality Theory: The case of syllable contact. *Phonology, 21*(2), 201–250.
Gow, D. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics, 65*(4), 575–590.
Guenther, F., Ghosh, S., & Tourville, J. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301.
Guenther, F., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics, 25*(5), 408–422.
Guion, S., Flege, J., Akahane-Yamada, R., & Pruitt, J. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America, 107*(5), 2711–2724.
Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package. *Journal of Statistical Software, 33*(2), 1–22.
Haggard, M. (1978). The devoicing of voiced fricatives. *Journal of Phonetics, 6*(2), 95–102.
Hallé, P., & Best, C. (2007). Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/ clusters. *Journal of the Acoustical Society of America, 121*(5), 2899–2914.
Hallé, P., Dominguez, A., Cuetos, F., & Segui, J. (2008). Phonological mediation in visual masked priming: Evidence from phonotactic repair. *Journal of Experimental Psychology: Human Perception & Performance, 34*(1), 177–192.
Hallé, P., Segui, J., Frauenfelder, U., & Meunier, C. (1998). Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 592–608.
Hancin-Bhatt, B., & Bhatt, R. (1997). Optimal L2 syllables: Interactions of transfer and developmental effects. *Studies in Second Language Acquisition, 19*(3), 331–378.
Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*(3), 373–405.
Hayes, B. (1984). The phonology of rhythm in English. *Linguistic Inquiry, 15*(1), 33–74.
Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*(3), 379–440.
Hayes-Harb, R., Nicol, J., & Barker, J. (2010). Learning the phonological forms of new words: Effects of orthographic and auditory input. *Language and Speech, 53*(3), 367–381.
Henke, E., Kaisse, E., & Wright, R. (2012). Is the sonority sequencing principle an epiphenomenon? In S. Parker (Ed.), *The sonority controversy* (pp. 65–100). Berlin: De Gruyter Mouton.
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.
Iverson, P., & Kuhl, P. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America, 99*(2), 1130–1140.
Jacquemot, C., Pallier, C., LiBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *Journal of Neuroscience, 23*(29), 9541–9546.
Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*(4), 485–499.
Jones, D., & Ward, D. (1969). *The phonetics of Russian*. Cambridge: Cambridge University Press.
Kabak, B., & Idsardi, W. (2007). Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech, 50*(1), 23–52.
Keating, P. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60*, 286–319.
Kim, M., Horton, W., & Bradlow, A. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology, 2*(1), 125–156.
Kingston, J. (2005). Ears to categories: New arguments for autonomy. In S. Frota, M. Vigario, & M. J. Freitas (Eds.), *Prosodies: With special reference to Iberian languages* (pp. 177–222). The Hague: Mouton de Gruyter.
Kittredge, A., Dell, G., Verkuilen, J., & Schwartz, M. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology, 25*(4), 463–492.
Kruschke, J. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York: Academic Press.
Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606–608.
Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech, 29*(1), 3–11.
Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*(3), 384–422.
McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*(2), 219–246.
Miller, J. (1994). On the internal structure of phonetic categories: A progress report. *Cognition, 50*(1–3), 271–285.
Miller, J., & Liberman, A. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics, 25*(6), 457–465.
Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173.
Morelli, F. (1999). *The phonotactics and phonology of obstruent clusters in optimality theory*. Unpublished Ph.D. dissertation, University of Maryland, College Park.
Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*(2), 132–142.
Nozari, N., Kittredge, A., Dell, G., & Schwartz, M. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language, 63*(4), 541–559.
Oh, G. E., & Redford, M. (2012). The production and phonetic representation of fake geminates in English. *Journal of Phonetics, 40*(1), 82–91.
Ohde, R. N., & Stevens, K. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America, 74*(3), 706–714.
Olmstead, A., Viswanathan, N., Aivar, M. P., & Manuel, S. (2013). Comparison of native and non-native phone imitation by English and Spanish speakers. *Frontiers in Psychology, 4*(475).
Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*(4), 2382–2393.
Patterson, K., & Shewell, C. (1987). Speak and spell: Dissociations and word-class effects. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The cognitive neuropsychology of language* (pp. 273–294). Hillsdale, NJ: Lawrence Erlbaum Associates.

Peperkamp, S., & Dupoux, E. (2003). Reinterpreting loanword adaptations: The role of perception. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th international congress of phonetic sciences* (pp. 367–370). Barcelona: Universitat Autónoma de Barcelona.

Pisoni, D., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*(2), 285–290.

Pitt, M. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & Psychophysics, 60*(6), 941–951.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnostics and output analysis for MCMC. *R News, 6*(1), 7–11.

Porter, R. J., & Lubker, J. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motor linkage in speech. *Journal of Speech and Hearing Research, 23*(3), 593–602.

Pruitt, J., Jenkins, J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *Journal of the Acoustical Society of America, 119*(3), 1684–1696.

Purcell, D., & Munhall, K. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America, 120*(2), 966–977.

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <http://www.R-project.org4/21/14>.

Ramus, F., Peperkamp, S., Christophe, A., Jacquemot, C., Koiuder, S., & Dupoux, E. (2010). A psycholinguistic perspective on the acquisition of phonology. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10: Variation, phonetic detail and phonological representation* (pp. 311–340). Berlin: Mouton de Gruyter.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92*(1), 81–110.

Repp, B. (1984). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech, 27*(3), 245–254.

Selkirk, E. (1984). On the major class features and syllable theory. In M. Aronoff & R. Oehrle (Eds.), *Language sound structure* (pp. 107–136). Cambridge, MA: MIT Press.

Silverman, D. (2011). Schwa. In M. van Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.), *Companion to phonology* (pp. 628–642). Malden, MA: Wiley-Blackwell.

Smith, C. (1997). The devoicing of /z/ in American English: Effects of local and prosodic context. *Journal of Phonetics, 25*(4), 471–500.

Solé, M. J. (2014). The perception of voice-initiating gestures. *Laboratory Phonology, 5*(1), 37–68.

Steriade, D. (2001). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. In K. Hanson & S. Inkelas (Eds.), *The nature of the word: Studies in honor of Paul Kiparsky* (pp. 151–180). Cambridge: MIT Press.

Ussishkin, A., & Wedel, A. (2003). Gestural motor programs and the nature of phonotactic restrictions: Evidence from loanword phonology. In G. Garding & M. Tsujimura (Eds.), *Proceedings of the West Coast conference on formal linguistics 22, UC San Diego* (pp. 505–518). Somerville, MA: Cascadilla Press.

Vendelin, I., & Peperkamp, S. (2006). The influence of orthography on loanword adaptations. *Lingua, 116*(7), 996–1007.

Venezky, R. (1970). *The structure of English orthography*. The Hague: Mouton.

Villacorta, V., Perkell, J., & Guenther, F. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America, 122*(4), 2306–2319.

Vitevich, M., & Luce, P. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408.

Vitevich, M., & Luce, P. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language, 52*(2), 193–204.

Werker, J., & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75*(6), 1866–1878.

Wilson, C., & Davidson, L. (2013). Bayesian analysis of non-native cluster production. In S. Kan, C. Moore-Cantwell, & R. Staubs (Eds.), *Proceedings of the Northeast linguistics society 40* (pp. 265–278). Amherst, MA: Graduate Linguistic Student Association.

Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 34–57). Cambridge: Cambridge University Press.

Yanagawa, M. (2006). *Articulatory timing in first and second language: A cross-linguistic study*. New Haven: Yale University.

Yeni-Komshian, G., Caramazza, A., & Preston, M. (1977). A study of voicing in Lebanese Arabic. *Journal of Phonetics, 5*(1), 35–48.

Young-Scholten, M., Akita, M., & Cross, N. (1999). Focus on form in phonology: Orthographic exposure as a promoter of epenthesis. In P. Robinson & N. Jongheim (Eds.), *Pragmatics and pedagogy, Vol. 2: The proceedings of the 3rd Pacific second language research forum* (pp. 227–233). Tokyo: Aoyama Gakuin University.

Zellou, G., Scarborough, R., & Nielsen, K. (2013). Imitability of contextual vowel nasalization and interactions with lexical neighborhood density. *Proceedings of Meetings on Acoustics, 19*(060083).

Zsiga, E. (2003). Articulatory timing in a second language: Evidence from Russian and English. *Studies in Second Language Acquisition, 25*(3), 399–432.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog infixation. *Language, 83*(2), 277–316.