Article

second language research

Second Language Research
1–32

DOI: 10.1177/0267658316637899
slr.sagepub.com
SAGE

Processing nonnative consonant clusters in the classroom: Perception and production of phonetic detail

Lisa Davidson
New York University, USA

Colin Wilson
Johns Hopkins University, USA

Abstract

Recent research has shown that speakers are sensitive to non-contrastive phonetic detail present in nonnative speech (e.g. Escudero et al. 2012; Wilson et al. 2014). Difficulties in interpreting and implementing unfamiliar phonetic variation can lead nonnative speakers to modify second language forms by vowel epenthesis and other changes. These difficulties may be exacerbated in the classroom, as previous studies have found that classroom acoustics have a detrimental effect on listeners' ability to identify nonnative sounds and words (e.g. Takata and Nábělek, 1990). Here we compare the effects of two acoustic environments – a sound booth and a classroom – on English speakers' ability to process and produce unfamiliar consonant sequences in an immediate shadowing task. A number of acoustic–phonetic properties were manipulated to create variants of word-initial obstruent–obstruent and obstruent–nasal clusters. The acoustic manipulations significantly affected English speakers' correct productions and detailed error patterns in both the sound booth and the classroom, suggesting that the relevant acoustic detail is not substantially degraded by classroom acoustics. However, differences in the response patterns in the two environments indicate that the classroom setting does affect how speakers interpret nonnative phonetic detail for the purpose of determining their production targets.

Keywords

acoustic detail, classroom acoustics, non-native speech perception, non-native speech production, phonotactics

Corresponding author:
Lisa Davidson, New York University, 10 Washington Place, New York, NY 10012, USA.
Email: lisa.davidson@nyu.edu

I Introduction

In order to successfully acquire a new sound system, the second language learner must discover the mapping between acoustic cues in the speech signal and nonnative phonetic and phonological structures. For example, a learner exposed to a range of stop aspiration values must determine whether these values correspond to a single category (voiceless unaspirated stops) or two categories (voiceless unaspirated and voiceless aspirated stops) (Curtin et al., 1998). Many types of knowledge and processing are known to be engaged in mapping acoustic cues to nonnative structures, including low-level and possibly language-independent auditory processes (Breen et al., 2013), filtering by native phonetic and phonological patterns (e.g. Best and Tyler, 2007; Flege, 1995; Kuhl et al., 2008; Strange, 2006), orthographic and other non-acoustic evidence for category distinctions (Showalter and Hayes-Harb, 2013; Hayes-Harb and Masuda, 2008; Swan and Myers, 2013), and developing knowledge of words in the new language (e.g. Trafimovich, 2005).

The present study is part of a series of experiments on English speakers' interpretation of acoustic cues in nonnative initial consonant clusters such as Russian /pt/ (e.g. /ptitsa/ 'bird') and /zn/ (e.g. /znakomɨj/ 'familiar'). Previous results have established that English participants are sensitive to fine-grained and non-contrastive acoustic details of cluster stimuli, such as high-amplitude phonation at the onset of voiced fricatives and the duration and amplitude of stop bursts. This sensitivity modulates the rates of specific errors, such as insertion of intrusive vowels, in participants' cluster productions (Davidson et al., 2015; Wilson et al., 2014). These results converge with those of perception studies showing attention to acoustic cues in challenging nonnative discrimination tasks (e.g. /bu/~/ɓu/ for English speakers, Best et al., 2001; /r/~/l/ for Japanese learners, Iverson et al., 2003; various tone contrasts, So and Best, 2010).

However, several aspects of previous experiments may have particularly encouraged attention to acoustic detail, thus raising questions about the generality and relevance of the findings for second language research. The English participants had no prior experience with Slavic languages (or other languages with similar consonant clusters), and the auditory stimuli were presented to them without accompanying written forms or cues to semantic content. This eliminated the possibility of using two sources of information, orthographic conventions and lexical knowledge, that are often available to second language (L2) learners and that could contribute to the interpretation of spoken words (e.g. Escudero et al., 2008). Furthermore, participants performed an immediate repetition (or shadowing) task, and it is plausible that this fosters low-level phonetic imitation (Dufour and Nguyen, 2013; Flege and Eefting, 1988; Goldinger, 1998; Rojczyk et al., 2013; Zając and Rojczyk, 2014). Most importantly for the present study, the previous experiments were conducted under the ideal listening conditions of a sound-treated room (henceforth, sound booth). While we retain some of the controlled properties of our previous studies in order to make direct comparisons with the current study, the main question we address in this article is whether sensitivity to acoustic detail in nonnative cluster processing can also be found in listening conditions that more closely approximate those encountered by L2 classroom learners.

The classroom environment differs acoustically from the sound booth in many ways (e.g. Crandell and Smaldino, 2004), including ambient speech and non-speech sounds

(e.g. electronic and A/V hums, transient noise from outside the room, etc.), varying distance to the speaker, and the presence of reverberations. Previous research reviewed below indicates that the classroom setting can detrimentally affect listeners' perception of speech sounds, and is challenging for nonnative listeners in particular. The issue investigated here is whether classroom acoustics affect the perceptual mapping and subsequent production of nonnative clusters in ways that differ from the sound-booth environment. Some hypotheses about the possible effects of acoustic environment are formulated at the end of this introduction.

1 Classroom acoustics and the perception of speech sounds

Empirical and modeling studies indicate that speech perception in the classroom is adversely affected by many factors, including ambient speech and non-speech noise, distance to the sound source, and reverberations (Bradley, 1986; Boothroyd, 2004; Crandell and Smaldino, 2004; Plomp et al., 1980). Classroom learners may be faced with noise from fellow classmates, from structural factors such as ventilation systems, and from the outside environment (Hodgson et al., 1999; Knecht et al., 2002; Nelson et al., 2005; Picard and Bradley, 2001). Even if the classroom is very quiet or sparsely populated, the ability to properly hear all of the acoustic information in the speech signal is affected by distance to the speaker. In order to perform a minimal comparison with our previous studies, in this article we chose to study transmission of the speech signal in an open but unoccupied classroom. Effects of distance and intrinsic classroom acoustics are therefore most relevant to the present study.

As a rule of thumb, Boothroyd (2004) approximates that average speech levels drop by 6 dB for every doubling of the distance from the speaker. The transmission from the speaker, which can be referred to as the direct signal, is accompanied by the reverberant signal, which refers to the persistence of sound due to the varied reflections off walls and surfaces in the room. Reverberations are an important consideration in the classroom acoustic landscape, since they are often present even in the absence of additive noise. Short reverberation times can produce echoes that enhance the speech signal, but echoes from long reverberation times are detrimentally shifted relative to the direct signal (Boothroyd, 2004; Bradley, 1986; Klatte et al., 2010; Knecht et al., 2002). Moreover, reverberation has differential effects on speech sounds; typically, consonants are more adversely affected than vowels (Lecumberri et al., 2010; Nábělek, 1988; Nábělek and Donahue, 1984). While some classrooms are designed to meet ANSI standards for optimal acoustic performance (e.g. ANSI/ASA S12.60-2010), up to half of the current classroom stock exceeds the ANSI maximum recommended reverberation time (Knecht et al., 2002).

Studies of the impact of classroom acoustics on speech perception typically examine the effect of both reverberation time and signal-to-noise ratio, with the latter tested by embedding speech either in normal classroom sounds (papers shuffling, chairs scraping, etc.), or in multi-talker babble. For example, Klatte et al. (2010) showed that reverberation in a silent classroom was not by itself sufficient to affect accuracy on word-to-picture matching and instruction-following tasks for either elementary-aged children or adults. However, when reverberation was combined with typical classroom sounds, it

had a significant negative effect compared to the same sounds with no reverberation for both groups (for similar results in a word matching task, see also Nábělek and Pickett, 1974). Other tasks may be more difficult to carry out in the presence of classroom reverberation even without masking sounds. Larsen et al. (2008) found that in the absence of amplification, college-aged participants were much less accurate in writing down CVC monosyllabic words played over speakers located in the front of a classroom that did not meet the ANSI standard for reverberation (44% accuracy) in comparison to one that did (82% accuracy). Moreover, especially in the high-reverberation classroom, word recognition performance degraded with distance from the sound source.

Nonnative listeners show even greater effects of classroom acoustics, relative to native-speaker controls, on speech recognition tasks in their L2. Using a forced-choice word recognition task, Nábělek and Donahue (1984) and Takata and Nábělek (1990) demonstrated that proficient nonnative speakers perform similar to native English speakers when there is little reverberation, but show a consistent disadvantage as reverberation time increases. Shi (2010) manipulated a number of variables, including the English proficiency of the participants, the predictability of the target word in a sentential context, the signal-to-noise ratio of the target words in multi-talker babble, as well as reverberation time. Results showed that reverberation had a significant interaction with the sentential context and the proficiency level of the participant, such that high reverberation levels had more detrimental effects on the late nonnative speakers in comparison to the earlier learners. Though the bilinguals in Shi (2010) performed very similarly to monolinguals on a task in which words were presented in sentences, Rogers et al. (2006) found that Spanish–English bilinguals were less accurate than English monolinguals on repetition of isolated words when they were presented in noise, including when reverberation was also added to the noisy signal.

2 *Effects of fine phonetic detail on nonnative cluster processing*

The findings reviewed above suggest that listeners, and in particular L2 listeners, face difficulties in accurately extracting acoustic phonetic properties from the signal in classroom settings. Indeed, a number of previous studies have provided evidence that a fine-grained level of detail, in addition to contrastive phonological status, affects cross-language speech perception (e.g. Best and Strange, 1992; Escudero et al., 2012; Hallé et al., 1999). In the prior experiment that forms the basis for the present investigation (Wilson et al., 2014), we demonstrated that parallel effects are also found in cross-language shadowing of consonant clusters. American English listeners were presented with obstruent–obstruent and obstruent–nasal clusters (e.g. /vdato/, /zmasa/, /bdafa/, /knapi/) produced by a Russian speaker. The stimuli were digitally manipulated to reflect subphonemic variation that is found in the natural cluster productions produced by our Russian consultants (see Wilson and Davidson, 2013). Voiced fricatives and stops either had a high-intensity onset of voicing (which also began before the onset of aperiodic noise in the fricative), or the voicing was of uniform intensity (and began with the frication noise). We refer to this manipulation as pre-obstruent voicing (POV). Additionally, cluster-initial stops varied in the duration and amplitude of their bursts (for details of these manipulations, see Section II.2).

In their shadowing responses, English speakers sometimes produced the nonnative clusters accurately (approximately 50% correct overall, with the rate varying by cluster type). They were significantly more likely to apply prothesis (e.g. /zmasa/→ [ᵊzmasa]) when POV was present, and more likely to epenthesize (e.g. /bdafa/→ [bᵊdafa]) when the burst of the initial stop was longer or higher in amplitude.[1] The acoustic manipulations also affected the rates of other modifications, most importantly for present purposes deletion and other changes to the initial consonant (e.g. /kpavo/→ [pavo], /kpavo/ → [tpavo]), both of which were more likely to apply to stops with short or low-amplitude bursts. Taken together, these results suggest that participants did not systematically 'repair' or map the non-native stimuli to native structures, but rather that the probability of a repair type is modulated by acoustic (or auditory) similarity of particular stimulus items to relevant native-language sounds and sequences (for further discussion, see Wilson et al., 2014). Participants appear to have paid careful attention to information in the signal that could be relevant for signaling the intended production targets. This perceptual sensitivity in fact led them to over-interpret the acoustic details: POV and burst duration/amplitude, based on observed non-contrastive phonetic properties of Russian, were intended to be non-contrastive in the experimental stimuli as well, but the participants often took them to signal phonological differences (e.g. between a cluster and a consonant–vocoid–consonant sequence).

For the current study, the null hypothesis is that similar modification patterns will be observed in the classroom environment (as it is implemented here). Findings consistent with the null hypothesis would suggest that our classroom acoustics, while possibly having other effects like those reviewed earlier, does not significantly impact nonnative interpretation of the specific acoustic properties that were modified in our stimuli. In addition to the hypothesis of no difference, we considered two non-null hypotheses about how performance could vary across the sound-booth and classroom contexts.

The first hypothesis, one that seems probable given the perceptual studies reviewed in the previous section, is that the classroom environment will provide listeners with less precise information about the stimulus acoustics, and therefore less detail on which to base their production responses. Several differences between the sound-booth study and the current experiment would follow from this hypothesis, which we refer to as 'Degraded Transmission'. If the classroom environment makes POV more difficult to pick up, and makes the duration and amplitude of a burst more difficult to perceive, the effects of these acoustic manipulations on prothesis and epenthesis repairs should diminish. A further prediction holds for the deletion and feature-change modifications. If participants in the sound-booth experiment deleted or changed initial consonants because they simply failed to detect them or accurately identify all of their features, under Degraded Transmission these modifications should occur at higher rates in the classroom context.

The second hypothesis, referred to as 'Reduced Imitation', focuses not on how much acoustic detail is available to listeners in each environment, but rather on how the classroom differentially affects participants' interpretation of detail in the shadowing task. In the optimal listening environment of the sound booth, listeners may be quite certain that any detail they perceive is attributable to the speech sample and, for that reason, worthy of being produced. Of course, close reproduction of nonnative stimuli will be articulatorily challenging, so attempts at detailed imitation could lead to diverse, stimulus-specific

errors like those observed previously. Contrast this with the classroom context, in which the same acoustic cues could be available but participants may be less certain about their origin: they may opportunistically assume that the speech sample has been somewhat distorted by distance, reverberation, and ambient sound sources. It would be natural for participants to put forth less effort to reproduce details that could, as far as they know, be due to noise in the proximal acoustics rather than being inherent to the speech source.

The predictions of Reduced Imitation overlap with those of Degraded Transmission to some extent, but the hypotheses are nevertheless distinguishable. Both are consistent with weaker effects of the acoustic manipulations in the classroom context. Reduced Imitation would be favored to the extent that weaker effects are not accompanied by other indications of misperception. For example, if participants in the classroom more often resort to a default repair strategy (e.g. epenthesis), rather than producing modifications that impair recoverability of the consonants (e.g. consonant deletion or feature change), this would indicate relatively successful transmission of acoustic detail.

In the next section, we present the details of the current study, aimed at investigating which of these possibilities best accounts for production of nonnative consonant clusters in a more realistic, classroom setting.

II Methodology

1 Participants

The participants were 36 New York University undergraduate and graduate students. All were native speakers of American English ranging in age from 18–35 year. Results from the 24 participants (8 males, 16 females) in the sound-booth condition were previously reported in Wilson et al. (2014), while those of the 12 participants (5 males, 7 females) in the classroom condition are new. Responses to a demographic questionnaire indicated that the participants did not speak Slavic languages nor any other languages with initial obstruent–obstruent or obstruent–nasal clusters (other than /s/-initial clusters), such as Hebrew. No speakers were bilingual in any other language, although they had studied languages such as Spanish, French or Mandarin in high school and college (one reported being proficient in Italian). None of the participants reported any speech or hearing impairments. They were compensated US$10 for their participation. This research was carried out with approval from the Institutional Review Board at New York University.

2 Materials

Critical stimuli consisted of nonce words of the form CCáCV (where á indicates the stressed low vowel /a/). The initial consonant clusters were composed of fricative–nasal (FN), fricative–stop (FS), stop–nasal (SN), and stop–stop (SS) sequences; Table 1 shows the particular clusters that were tested. 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, 2006, 2010). Stop-initial clusters contained both voiceless and voiced consonants. Stop–stop sequences agreed in voicing, but stops of both voicing values appeared before nasals. Each cluster

Table 1. Target consonant clusters used in the CCaCV 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/

appeared in four distinct stimulus items (see Appendix 1), 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.

The stimuli were recorded by a Russian–English bilingual linguist (a PhD candidate in linguistics who came to the USA from Moscow in elementary school and continues to regularly speak Russian with family and friends). Although not all of the consonant clusters used in this study are attested in Russian, the talker was able to produce them without intrusive vowels and in a way that contrasted with the fillers. Previous perception studies have shown that Russian listeners can perceive similar clusters, regardless of whether they are directly attested in Russian or not, with very high accuracy (Berent et al., 2007; Davidson, 2011). Plausibly, many obstruent–nasal and obstruent–obstruent clusters are not observed in the language simply because of restrictions on morphological form and combination, and are hence considered 'accidental gaps' by Russian native speakers. Moreover, since all of the recordings were subsequently acoustically manipulated to create the stimuli for the conditions described below, any potential phonetic differences that might exist between frequent and infrequent consonant sequences should be neutralized in our stimuli.

The recorded stimuli, the same as those in Wilson et al. (2014), were subjected to three acoustic manipulations hypothesized to influence the participants' productions.

a Pre-obstruent voicing. The first modification was pre-obstruent voicing (POV), which is defined as an interval of voicing located before or at the beginning of the formation of an obstruent constriction and with visibly higher amplitude than the voicing that is present during the subsequent constriction. In the case of fricative-initial clusters, POV precedes the onset of frication, while in stop-initial clusters POV occurs at the beginning of closure voicing. In all cases, POV contains low-frequency periodic energy but does not have visible formant structure.

Each item beginning with a voiced obstruent had versions with and without POV. When a recording had naturally-produced POV, this was spliced out using Praat (Boersma, 2001) to create the non-POV variant. For stimuli that were not originally produced with POV, the initial voiced interval was spliced in from a different utterance of

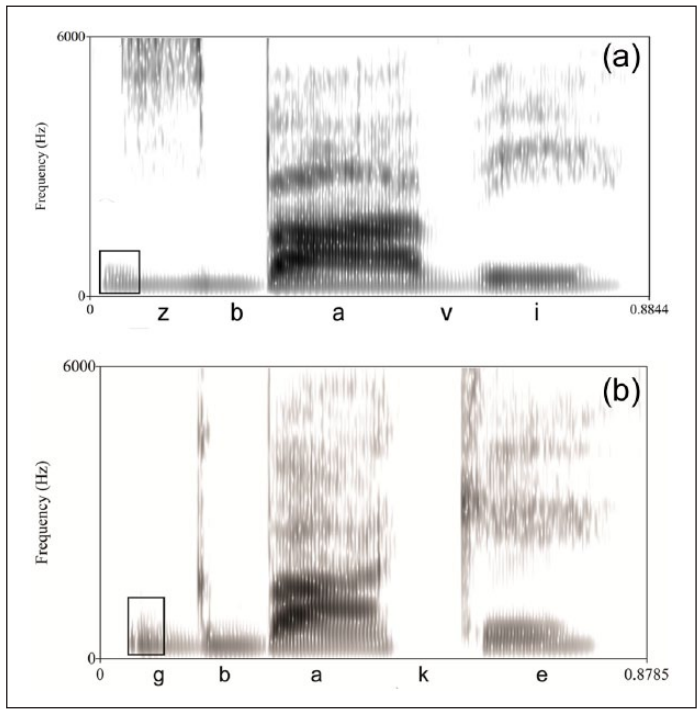

Figure 1a. Illustration of FS-initial cluster with pre-obstruent voicing (POV), indicated by the box.
Figure 1b. Illustration of SS-initial cluster with POV.

the same cluster. All splices were taken at zero-crossings to avoid acoustic artifacts. This manipulation affected all clusters beginning with voiced obstruents. Since the stimuli took advantage of the POV naturally produced by the Russian speaker, there was some variability in its duration. The mean duration of the POV was 53 ms for stop-initial sequences (range: 30 ms–80 ms) and 44 ms for fricative-initial sequences (range: 30 ms–90 ms). Spectrograms illustrating the presence of POV in both fricative-initial and stop-initial clusters are shown in Figures 1a–1b.

b Burst duration. 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 20–40 ms, regardless of voicing (SN: mean = 36 ms, sd = 17.7 ms; SS: mean = 28 ms, sd = 8.7 ms). Shorter durations in the stimuli were created by splicing 5–10 ms out of the middle portion of the original duration of the aspiration of release of the first consonant. Longer durations were created by selecting between 10–20 ms of the middle portion of the aperiodic burst and splicing that material back into the recording. Splices were again taken from and inserted at zero crossings to avoid acoustic discontinuities. The duration manipulation affected voiced and voiceless SN and SS stimuli.

Table 2. 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)

c *Burst amplitude.* The third modification targeted the relative burst amplitude of stimulus-initial stops. Using Praat, we first calculated the amplitudes of the bursts relative to the following oral stop or nasal closure for each recording of the SN and SS stimuli by the Russian talker. Because nasal closures are naturally higher in amplitude than those of oral stops, and hence stop bursts have lower relative amplitude before nasals, the values for this manipulation were determined for each cluster type separately (see further discussion in Wilson et al., 2014). For SN clusters, the low-amplitude versions had values based on the means of the corresponding natural productions (voiceless SN: –18 dB, voiced SN: –7 dB), and high-amplitude versions were raised several decibels above the means (voiceless SN: –10 dB, voiced SN: 0 dB). The direction of manipulation was reversed for SS clusters: the high-amplitude versions mirrored the natural means (voiceless SS: +23 dB, voiced SS: 0 dB), while the low-amplitude versions were reduced in amplitude (voiceless SS: +13 dB, voiced SS: –7 dB). The manipulated values were chosen carefully to ensure that all bursts, and in particular those with lowered amplitude, were audible and sounded intelligible.[2]

These manipulations were crossed as summarized in Table 2. Together, all of the manipulated stimuli and the fillers (which were not modified except by normalization of the amplitude of all of the stimuli to 67dB) constituted 800 sound files. To create an experimental procedure that would not be too taxing for the participants, 12 counterbalanced lists were created with 288 stimuli each. Each list was composed of 32 items each of FN and FS, 64 items each of SN and SS (each with half with voiceless stops, half with voiced), 48 CəC fillers, and 48 əCC fillers. Stimuli were distributed across the experimental lists so that each one contained approximately the same number of each kind of manipulation within each cluster type. In the sound-booth condition, two participants were assigned to each list. As preliminary analysis of this data suggested that 12 participants provide sufficient power to detect the effects of interest, one participant was assigned to each list in the classroom condition.

3 Procedure

The sound-booth condition, originally reported in Wilson et al. (2014), provides an ideal-listening baseline against which the classroom condition can be interpreted. Participants (N = 24) were individually seated in a sound-attenuated room with a computer running ePrime 1.1 (Psychology Software Tools, Pittsburgh, PA). In each trial, a single stimulus was presented twice over computer speakers before the response; no orthographic or other information accompanied the audio. The second repetition of the stimulus was

presented 450 ms after the end of the first stimulus. Participants were given 1.5 seconds after the offset of the second repetition to respond before the program automatically advanced to the next item. Participants did not have the opportunity to correct or otherwise evaluate their responses.

The 288 items were divided into three blocks, with short breaks between blocks. The production 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.

The classroom condition was identical in procedure to the earlier experiment, differing only in the listening context. The participants (N = 12) were tested individually in a classroom in the Linguistics department at New York University. The room is 9.7 m × 6.1 m, and can hold approximately 30 people when occupied with tables, as in the experiment. The classroom contains 6 rows of wooden tables and plastic chairs. The maximum occupancy of the room is 60. The front wall of the room is made of drywall, and has a large whiteboard (6.1 m × 2.0 m) covering about 70% of the wall mounted over the drywall. The back wall, also made of drywall, has two windows (each 2.4 m × 1.2 m) that face a quiet air shaft. One side wall of the room is brick, and the other side wall is mostly glass, with a wooden door (2.7 m × 1.9 m) at the entrance to the room. The ceiling is made of acoustic tile, and the floor is covered with high traffic carpeting. During the experiment, the projection system in the room was not on, but an additional quiet hum from the air conditioning in the HVAC system in the room was audible. A Lenovo Ideapad netbook and Harmon Kardon HK206 computer speakers running the experiment were set up in the front of the room, and the participant was seated alone in the middle of the room, about 3.3 m away from the speakers. The computer speakers were the same as those used for stimulus presentation in the sound-booth condition. No changes were made to the volume controls on either the computer or the speakers between participants.

4 Data analysis

Coding of the data followed the same procedure as Wilson et al. (2014) and earlier studies (e.g. Davidson, 2010). Productions were analysed by repeated listening as well as examination of waveforms and spectrograms in Praat. Modifications of a consonant cluster relative to the stimulus were labeled as shown in Table 3. If multiple errors occurred, each error was labeled, and if none of the errors found in Table 3 occurred, the token was labeled as ‘no modification’ (i.e. accurate). A token was coded for epenthesis if it had voiced vocalic material, containing visible first and second formants, that occurred between the two consonants of a target cluster. To be coded for prothesis, a response had to have a voiced vocalic element containing first and second formants before the initial obstruent; voicing during stop closure, or voicing that started before the closure, was not sufficient to qualify as an error because these properties are found in our talker’s natural productions of the target clusters. More generally, to be coded as accurate, participants’ utterances had to match the manner, place, and voice specifications of the input, and the consonants had to be produced in the correct linear order, as

Table 3. 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]

determined using the spectrogram. When coding the errors, small variations from the target stimulus, such as in the duration of a consonant or a burst, did not prevent the token from being classified as a correct production.

The responses were coded by four research assistants and one of the authors (LD). All coding was done blindly, without knowledge of the acoustic acoustic manipulations that had been applied to each stimulus. All responses were then discussed by at least two different research assistants and the first author in regular lab meetings to ensure coding consensus and consistency.

In the sound-booth condition, 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. /kpabi/ → Ø, / kpabi / → [pkabi], / kpabi / → [spabi]). A comparable portion was removed from the classroom condition (3.6%).

III Results

Coded responses for the sound-booth and classroom conditions were combined and submitted to a Bayesian multinomial mixed-effects analysis (as in Wilson et al. 2014; see also Raudenbush and Bryk, 2002; Cunnings, 2012). For each analysis below, the correct or no-modification response category served as the baseline against which the other response types (epenthesis, prothesis, C1-deletion, C1-change) were compared. Fixed effects of condition, cluster type, cluster voice, and binary acoustic manipulations were standardized to have means equal to 0 and standard deviations of 1. Random effects of the response options and all acoustic manipulations were included for participants and items. Analyses were implemented in R (R Development Core Team, 2012) with the MCMCglmm package (Hadfield, 2010), which returns point estimates and 95% highest-posterior density (HPD) intervals for each fixed coefficient. The Bayesian prior on coefficients and other aspects of the model were given default specifications.

In previous studies (e.g. Davidson, 2010; Wilson et al., 2014), we have observed that modification rates vary considerably by cluster type. For example, clusters beginning with voiced stops undergo epenthesis more often than those beginning with voiceless stops. Because our interest lies primarily in the effects of the acoustic details and listening condition, and in light of the fact that the precise manipulated acoustic values differ somewhat across cluster types (see Section II.2), we analyse each type separately below.

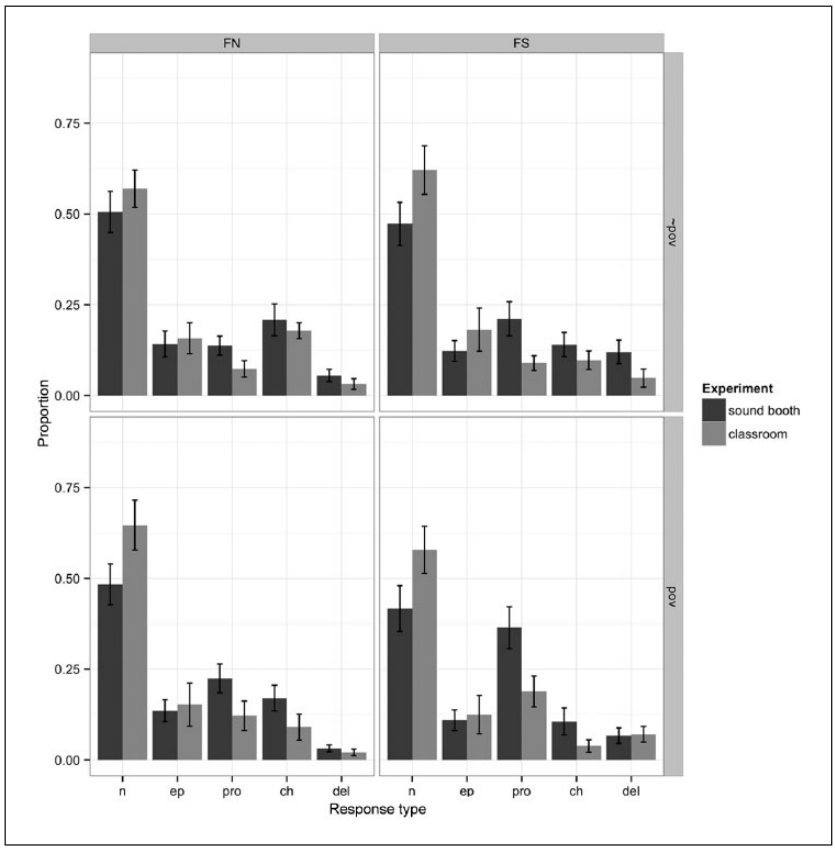

Figure 2. Results of the POV manipulation for fricative-initial (FC) clusters.
Notes. FN = fricative-nasal, FS = fricative stop, pov = POV is present, ~pov = POV is not present. n = no modification, ep = epenthesis, pro = prothesis, ch = C1 change, del = deletion.

1 Fricative-initial clusters

Figure 2 shows the proportion of trials on which each coded modification type (and no-modification) occurred in the sound-booth and classroom conditions. The acoustic manipulation for fricative-initial sequences was POV, and it can be observed that there are differences in the rates of prothesis across values of POV as well as across listening conditions.

a Fricative-nasal clusters (FN_{vcd}). The analysis of the FN clusters included fixed effects of condition (classroom vs. sound booth) and POV (present vs. absent), as well as random intercepts and slopes for participants (POV only) and items (condition and POV). All modification types were less probable than accurate production (epenthesis: −2.17 [−2.82, −1.6]; prothesis: −1.85 [−2.34, −1.36]; C1-change: −1.56 [−2.03, −1.21]; C1-deletion: −4.24 [−5.48, −3.25], all *p*s < .01).[3] The POV manipulation had marginal

effects on productions of this cluster type (POV × prothesis: 0.27 [−0.04, 0.54], $p = .08$; POV × C1-change: −0.24 [−0.49, 0.02], $p = .07$), and there was also a marginal effect of condition on prothesis (condition × prothesis: −0.46 [−0.91, 0.06], $p = .08$), suggesting that prothesis was overall less probable in the classroom condition, though this result should be interpreted with caution. One three-way interaction did reach significance: the effect of the POV manipulation on the probability of C1-change differed across the conditions (condition × POV × C1-change: −0.23 [−0.47, 0], $p < .05$). The presence of POV roughly halved the rate of C1-change in the classroom condition, but had a limited effect on this modification type in the sound booth.

b Fricative-stop clusters (FS_{vcd}). The fixed and random structure for the analysis of FS clusters was identical to that of the preceding analysis. All modifications were less probable than accurate production (epenthesis: −2.07 [−2.75, −1.44]; prothesis: −1.29 [−1.87, −0.78]; C1-change: −2.18 [−2.76, −1.71]; C1-deletion: −3.62 [−4.67, −2.72], all ps < .01). Importantly, the probability of prothesis was increased by the presence of POV (POV × prothesis: 0.52 [0.25, 0.78], $p < .01$) and lower overall in the Classroom condition (condition × prothesis: −0.58 [−1.06, −0.05], $p < .05$). As for fricative-nasal clusters, the probability of C1-change was marginally lower in the Classroom condition (condition × C1-change: −0.49 [−1.02, 0.03], $p = 0.08$). Finally, the effect of POV on C1-deletion was different in the two studies (condition × POV × C1-deletion: 0.37 [0.06, 0.72], $p < .05$), with POV lowering the C1-deletion rate in the sound-booth condition only.

A strong version of the Degraded Transmission hypothesis predicts that POV should have essentially no effect on shadowing responses in the classroom condition. That is, the low-frequency energy associated with POV – or the difference between POV and the following voiced frication – could be so difficult to perceive in the classroom that it has minimal influence on participants' productions. The statistical analysis above is consistent with this possibility, in spite of the fact that no interaction between condition and POV was found.[4] Therefore, a post-hoc test was performed to assess whether POV had any effect on the rate of prothesis in the classroom condition alone. For this analysis, the dependent variable was binary (prothesis vs. no prothesis) and the only fixed predictor was POV; random intercepts and POV slopes were included for participants and items. The effect of POV was significant (intercept: −2.84 [−3.77, −1.97], $p < .01$; POV: 1.23 [−0.08, 2.20], $p < .05$). This provides evidence against the hypothesis that POV was inaudible to participants in the classroom condition, though it does not adjudicate between a weaker version of Degraded Transmission and the Reduced Imitation hypothesis. It could be that POV was more difficult, though not impossible, to perceive in the classroom, or alternatively that it was perceived at the same rate but had a reduced effect on participants' production targets (for further discussion, see Section IV).

2 *Stop-initial clusters*

Recall that stop-initial clusters in the stimuli varied by the manner of the second consonant (stop-nasal vs. stop-stop) and the voicing of the initial stop (voiceless vs. voiced). Again, because the details of the acoustic manipulations depended on the particular cluster composition, we separately analysed the effects of acoustic manipulations on

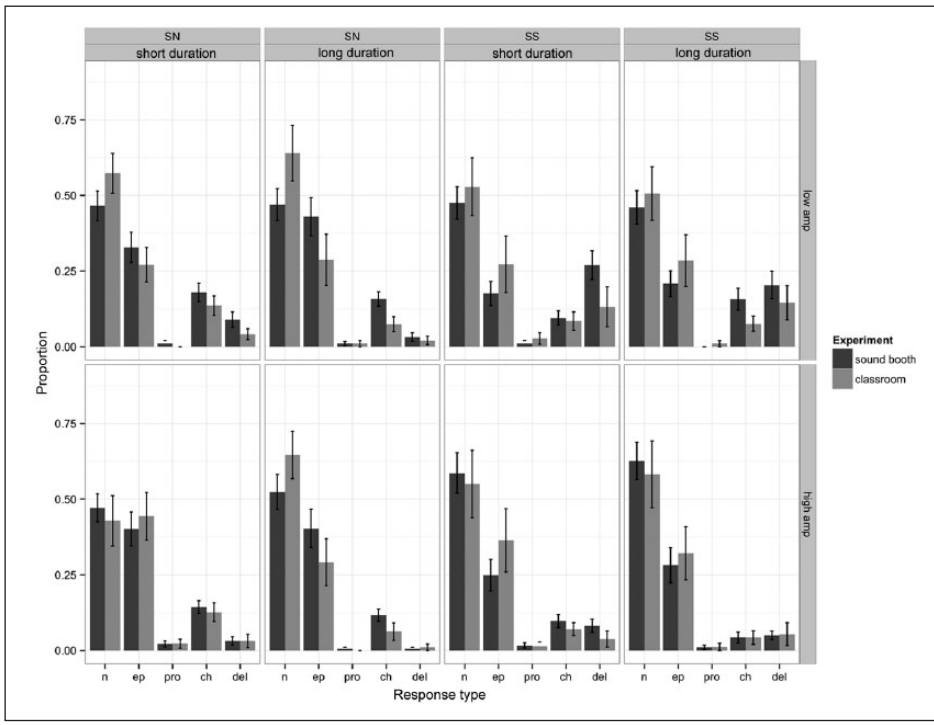

Figure 3. Results of the amplitude and duration manipulations for voiceless stop-initial clusters.
Notes. SN = stop nasal, SS = stop stop. See Figure 2 for response type key.

productions of each of these four types. The manipulations involved burst duration, burst amplitude, and presence of POV. The analyses of Wilson et al. (2014) found no significant differences between the effects of 20 and 30 ms bursts, or between those of 40 and 50 ms bursts; therefore, in the interest of simplifying the present analyses, burst duration was collapsed to a binary distinction (short: 20, 30 ms vs. long: 40, 50 ms). Relative burst amplitude was also coded as a binary factor (high vs. low), though recall that the absolute amplitudes differed by voicing of the first consonant and manner of the second consonant. The POV manipulation (present vs. absent) applied to voiced stop-initial clusters only. Results for the stop-initial sequences are presented in Figures 3–5.

a Voiceless stop-nasal clusters (SN_{vcls}). In this and subsequent analyses, fixed effects of condition (classroom vs. sound booth) and relevant acoustic manipulations were included. Random intercepts and slopes corresponding to the fixed effects were included for participants and items as permitted by the experimental design. Results for voiceless stop-nasal clusters are shown in Figure 3.

All modifications were estimated to be less probable than accurate production for this cluster type (epenthesis: –0.52 [–0.96, –0.16]; prothesis: –4.7 [–5.38, –3.99]; C1-change: –2.37 [–3.01, –1.67]; C1-deletion: –3.57 [–4.22, –3.04], all *p*s < .01). The acoustic

manipulations had a few effects on the production responses. Longer burst duration lowered the probability of modifying the initial stop by feature change or deletion (duration × C1-change: –0.29 [–0.55, –0.01], $p < .05$; duration × C1-deletion: –0.79 [–1.25, –0.35], $p < .01$). Higher burst amplitude had a similar protective effect on the initial stop, reducing the probability deleting it (amplitude × C1-deletion: –0.66 [–1.34, –0.19], $p < .01$). The duration and amplitude manipulations did not have overall effects on epenthesis rate. However, there was one surprising interaction involving burst duration: the effect of duration on epenthesis was reversed in the classroom study (condition × duration × epenthesis: –0.19 [–0.35, –0.01], $p < .05$). This result, which is not expected under any of the hypotheses considered here, is due to parity between accurate and epenthesis responses for the short, high-amplitude bursts in the classroom condition (see the bottom left cell of Figure 3).

In comparison to the sound-booth condition, C1-change was overall less probable in the classroom setting (condition × C1-change: –0.38 [–0.65, –0.08], $p < .01$). This result is surprising from the perspective of the Degraded Transmission hypothesis: if classroom acoustics significantly interfere with the perception of nonnative clusters, how could the rate of feature misperception (and subsequent misproduction) be lower in this environment? The result is, however, consistent with the Reduced Imitation hypothesis. Suppose that participants successfully recovered stop place and other features from low-amplitude bursts in both conditions. Deletion and feature change could have been more frequent in the sound booth because, as discussed in the introduction, participants attempted to mimic the fine-grained details of the stimulus but in doing so failed to produce intelligible stops. Articulatorily, imitating a low-amplitude burst requires partially suppressing or otherwise defusing the intraoral pressure that is characteristic of word-initial voiceless stops in English (e.g. Müller and Brown, 1980). Failure to sustain sufficient pressure, and to properly coordinate pressure build-up and release with constriction gestures, could result in responses that are featurally distinct from the targets. In cases such as this, less phonetic imitation – as we suggest occurs in the classroom – could actually increase the rate of categorically accurate productions.

b Voiceless stop-stop clusters (SS_{vcls}). As for the previous cluster type, all modifications were less probable than accurate productions of voiceless stop-stop items (epenthesis: −1.44 [−2.13, −0.85]; prothesis: −4.84 [−5.38, −4.14]; C1-change: −2.22 [−2.54, −1.92]; C1-deletion: −2.37 −3.05, −1.83], all ps $< .01$). Only the amplitude manipulation influenced modification rates: higher burst amplitude reduced the probability of C1-change and C1-deletion (amplitude × C1-change: −0.41 [−0.7, −0.11]; amplitude × C1-deletion: −0.88 [−1.19, −0.56], both ps $< .01$). There was no effect of burst duration on epenthesis or other categorical response types. These results are shown in Figure 2. However, in Section III.2.e below we show that stimulus burst duration was reflected in the phonetic details of productions of both SN_{vcls} and SS_{vcls} clusters, a result that bears on the perception and degree of imitation of this acoustic manipulation. There were no important effects of condition for this cluster type.[5]

c Voiced stop-nasal clusters (SN_{vcd}). Voiced stop-initial clusters were subject to the most extensive acoustic manipulations, with varying burst duration combined separately with

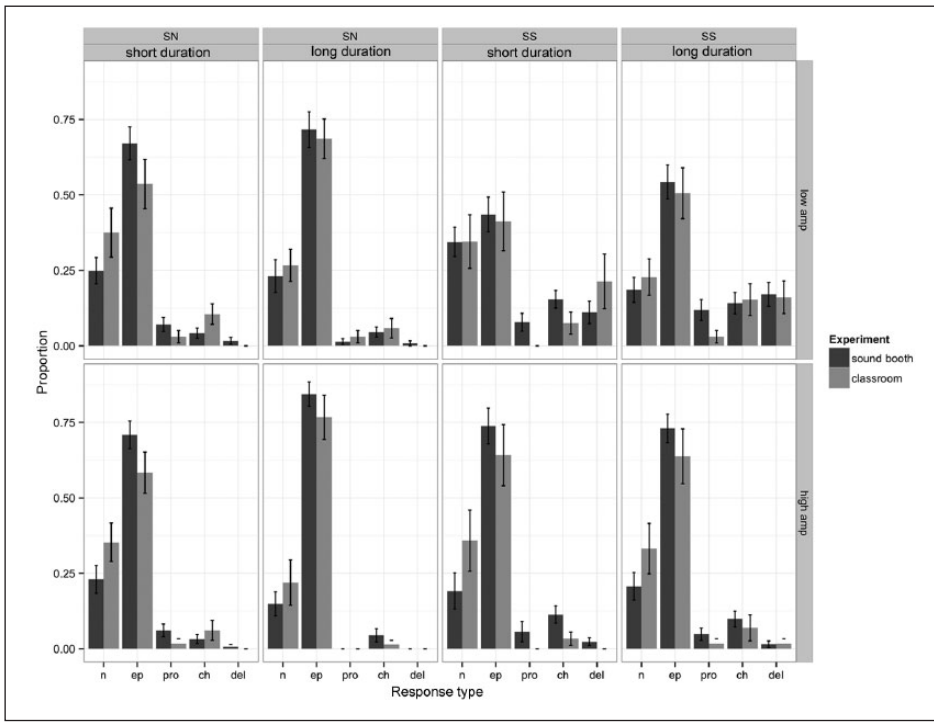

Figure 4. Results of the amplitude and duration manipulations for voiced stop-initial clusters.
Notes. SN = stop nasal, SS = stop stop. See Figure 2 for response type key.

burst amplitude and the POV manipulation. One might therefore expect the strongest evidence for sensitivity to acoustic detail – and differences in sensitivity across listening conditions – to be found for voiced stop-nasal and stop-stop clusters.

In contrast to voiceless stop-initial clusters, epenthesis was the most probable response type for SN_{vcd} stimuli (epenthesis: 1.26 [0.94, 1.65], $p < .01$); all other modification types remained less probable than accurate production (prothesis: –2.02 [–2.43, –1.69]; C1-change: –1.95 [–2.29, –1.62]; C1-deletion: –4.33 [–5.42, –3.45], all ps $< .01$). Also unlike the clusters beginning with voiceless stops, longer burst duration increased the probability of epenthesis (duration × epenthesis: 0.36 [0.13, 0.54], $p < .01$), but there were no significant effects of burst amplitude. The results for the amplitude and duration manipulations are shown in Figure 4. Presence of POV increased the probability of prothesis (POV × prothesis: 0.63 [0.24, 1.08], $p < .01$) and marginally decreased the probability of C1-deletion (POV × C1-deletion: 0.61 [–0.24, 1.11], $p = .07$). The results for the POV manipulation are illustrated in Figure 4.

The classroom condition elicited an overall lower rate of prothesis (condition × prothesis: –0.66 [–1.08, –0.29], $p < .01$), as found above for fricative-stop clusters. Indeed, for this cluster type the dispreference for prothesis was so strong in the classroom that it effectively nullified the effect of POV, as determined by post-hoc binary logistic

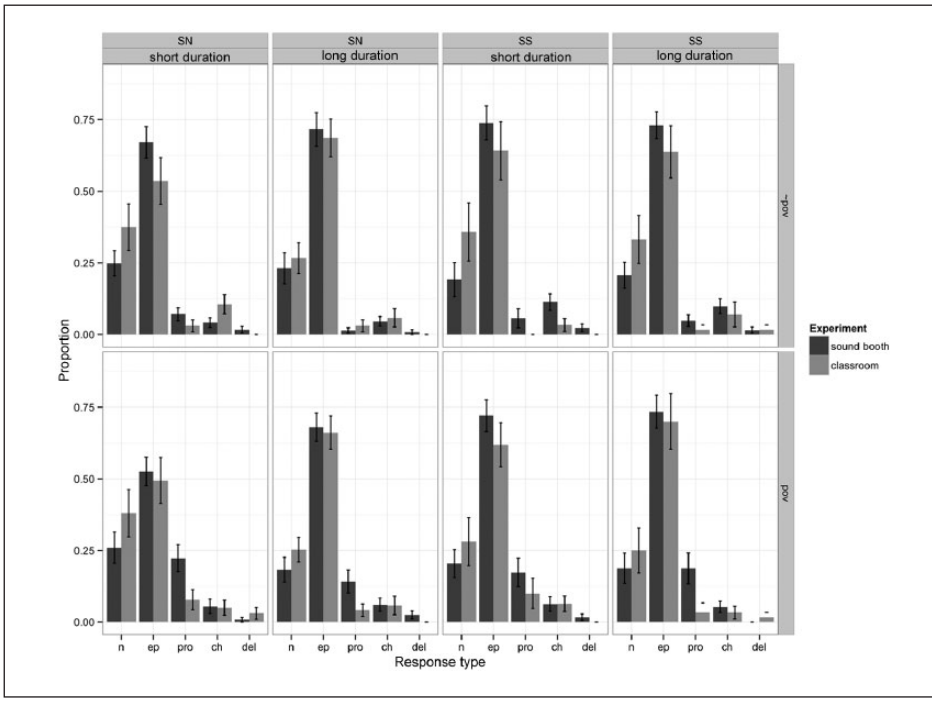

Figure 5. Results of the POV and duration manipulations for stop-initial clusters.
Note. See Figure 2 for response type key.

regression on the classroom data only (intercept: –5.77 [–7.69, –3.82], $p < .01$; POV: 1.28 [–0.71, 3.40], $p = .23$). This difference between the sound-booth and classroom conditions is equally consistent with the Degraded Transmission and Reduced Imitation conditions.[6]

d Voiced stop-stop clusters (SS_{vcd}). As for voiced stop-nasal clusters epenthesis was more probable than no-modification (epenthesis: 1.09 [0.79, 1.45], $p < .01$), which was in turn more probable than all other modification types (prothesis: –1.89 [–2.5, –1.43]; C1-change: –1.29 [–1.56, –0.99]; C1-deletion: –2.54 [–2.9, –2.2], all ps $< .01$). There was a marginal effect of burst duration on epenthesis (duration × epenthesis: 0.18 [0.01, 0.4], $p = .07$). Higher burst amplitude raised the probability of epenthesis and lowered the probability of C1-deletion (amplitude × epenthesis: 0.37 [0.14, 0.61]; amplitude × C1-deletion: –1.1 [–1.49, –0.65], both ps $< .01$), as illustrated in Figure 4. Presence of POV raised the probability of prothesis (POV × prothesis: 0.67 [0.29, 1.15], $p < .01$) and of epenthesis (POV × epenthesis: 0.42 [0.17, 0.63], $p < .01$); it also lowered the probabilities of C1-change and C1-deletion (POV × C1-change: –0.45 [–0.82, –0.07], $p < .05$; POV × C1-deletion: –1.55 [–2.03, –1.04], $p < .01$). These effects are shown in Figure 5.

The probabilities of two modification types, prothesis and C1-change, were lower in the classroom condition (condition × prothesis: –0.82 [–1.46, –0.39], $p < .01$; condition

× C1-change: –0.37 [–0.61, –0.12], $p < .05$). A post-hoc logistic regression of the classroom data confirmed that POV did not significantly influence the rate of prothesis in the present condition (intercept: –6.58 [–8.98, –4.38], $p < .01$; POV: 2.10 [–0.26, 4.84], $p = .11$). This finding does not distinguish between the two hypotheses under consideration. However, the lowering of C1-change rate in the classroom condition converges with the results for voiceless stop-nasal clusters. As discussed earlier for SN_{vcls} clusters, Reduced Imitation but not Degraded Transmission is consistent with a lower rate of this error type under classroom presentation.

e *Imitation of burst duration.* In addition to the coded responses analysed above, there is an additional type of data that bears on the Degraded Transmission and Reduced Imitation hypotheses. Wilson et al., (2014) demonstrated that stop burst durations in no-modification (accurate) productions matched, to a certain degree, the burst durations in the stimuli. This is a type of phonetic imitation effect that has been found in several previous studies (e.g. Fowler et al., 2003; Nielsen, 2011; Shockley et al., 2004). The imitation effect was clearest for clusters beginning with voiceless stops, and we concentrate on those here. Because the imitation effect is likely to be phonetically gradient, in this section we treat burst duration as a numerical factor and distinguish all four manipulated values (20, 30, 40, 50 ms).

The degree of phonetic imitation can be measured by the strength of a linear relationship between stimulus and response burst durations. If the Degraded Transmission hypothesis applies to the type of acoustic detail require to accurately perceive burst duration, we would expect this relationship to be weaker in the classroom setting. Instead of imitation, we could find that participants produce stop bursts of varying but not stimulus-locked durations, or possibly that they resort to a less variable 'default' duration. The Reduced Imitation hypothesis would also be consistent with such outcomes. However, it could be that the duration of burst frication and aspiration is particularly salient to English speakers and therefore imitated in both the sound-booth and classroom conditions.

A linear regression analysis of voiceless stop bursts was performed to assess the evidence for these possible outcomes. The fixed factors were condition (classroom vs. sound booth), cluster type (stop-nasal vs. stop-stop), and stimulus burst duration; these three factors were fully interacted. Random intercepts and slopes were included for participants and items. (Because participants were nested within condition, the random slopes for participant were for cluster type and stimulus duration only. Similarly, items were nested within cluster types therefore only slopes for condition and duration were included.) A handful of responses with outlier burst durations exceeding 120 ms were excluded from the analysis (7 tokens from the sound-booth condition and 1 token from the classroom condition).

The results, displayed in Figure 6, do not support degraded transmission of burst duration in the classroom. Response burst durations increased linearly with stimulus burst duration (7.89 [5.85, 10.37], $p < .01$) and there was no interaction of this factor with condition ($p > .75$). Cluster type also had a significant effect on response bursts, with stop-nasal clusters eliciting longer bursts than stop-stop clusters (cluster type: 7.18 [5.01, 9.07], $p < .01$), but again no interaction with condition (or stimulus duration) was found.[7] While this finding may appear to also be a challenge for the Reduced Imitation

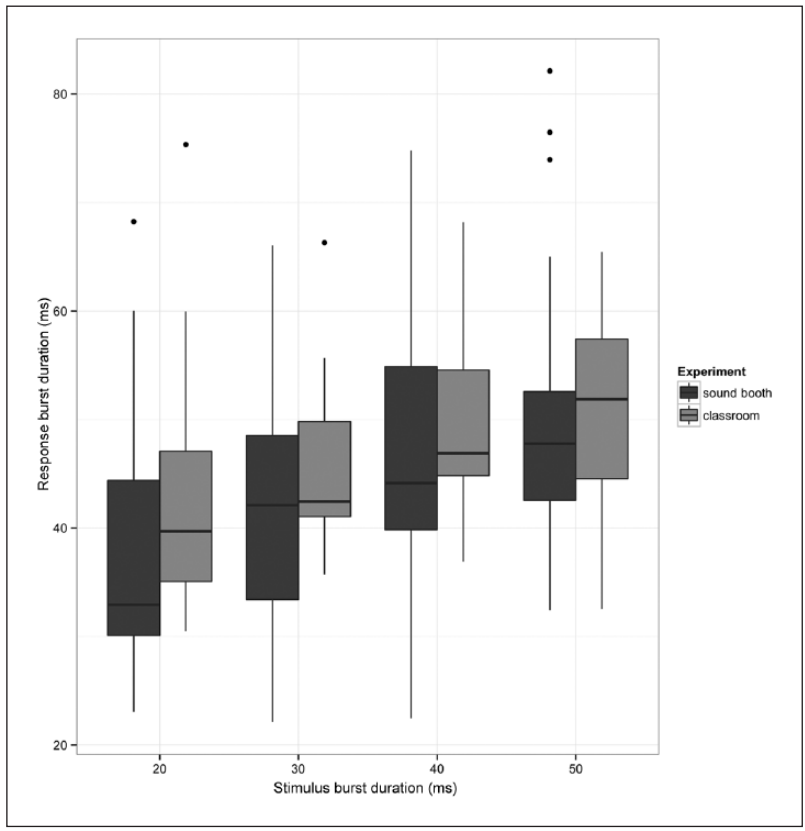

Figure 6. Imitation of burst duration in no-modification (correct) productions of clusters beginning with initial voiceless stops.

hypothesis, in the General Discussion we argue that the distinction between acoustic details that are and are not imitated in the classroom condition can be made on the grounds of native English phonetic patterns. Finally, the present analysis fills a gap in the preceding discussion of voiceless stop-nasal and stop-stop clusters: while no effect of burst duration on categorical epenthesis was found for such clusters (cf. voiced stops), this stimulus property nevertheless did influence responses in the gradient way shown here.

IV General discussion

1 Reduced imitation vs. degraded transmission

The results of this study provide new insight into the challenges of the perceptual interpretation and subsequent production of nonnative consonant clusters. Consistent with previous studies conducted under ideal listening conditions (Davidson et al., 2015;

Table 4. Summary of significant (and marginal) effects of acoustic manipulations on modification types.

	Epenthesis	Prothesis	C1 change	C1 deletion
FN_{vcd}		(Increase for +POV) (*Less probable in classroom*)	Decrease for +POV *in classroom*	
FS_{vcd}		Increase for +POV *Less probable in classroom*	Decrease for +POV *in classroom*	Decrease for +POV *in sound booth*
SN_{vcls}			Decrease for DURlong Decrease for AMPhi *Less probable in classroom*	Decrease for DURlong Decrease for AMPhi
SS_{vcls}			Decrease for AMPhi	Decrease for AMPhi
SN_{vcd}	Increase for DURlong	Increase for +POV *in sound booth*		(Decrease for +POV) (*Less probable in classroom*)
SS_{vcd}	(Increase for DURlong) Increase for AMPhi	Increase for +POV *in sound booth*	Decrease for +POV *Less probable in classroom*	Decrease for AMPhi Decrease for +POV

Notes. +/–POV refers to the presence or absence of pre-obstruent voicing; +/–AMP to high or low amplitude, and +/–DUR to long versus short duration. Blank cells indicate no significant effects. Except where specifically noted that only one condition was affected, the effect of a manipulation was present in both the sound-booth and the classroom conditions.

Wilson et al., 2014), participants in the classroom condition were found to be sensitive to at least some non-contrastive acoustic detail, as summarized in Table 4. For example, presence of POV increased the rate of prothesis for fricative-initial clusters (in particular fricative-stop sequences), higher burst amplitude increased the probability of epenthesis and lowered the probability of C1 deletion for voiced SS sequences, longer burst duration increased the probability of epenthesis for voiced stop-initial sequences, and the duration of the bursts in responses that were not otherwise modified was imitated to some degree. Therefore, the level and type of acoustic detail studied here could in principle be relevant for naturalistic L2 learning of new clusters, at least under quiet classroom conditions.

However, the response patterns found in the classroom condition were not identical to those obtained in the sound booth. For example, the effect of POV on prothesis was reduced in fricative-initial clusters and eliminated for stop-initial clusters. This part of the pattern is consistent with both of the hypotheses raised in the introduction: the POV effect could be reduced because classroom acoustics hinder the perception of this property, as expected under Degraded Transmission, or because participants make less of an effort to reproduce this property, as Reduced Imitation would have it.

Broader examination of the results provides key evidence against across-the-board Degraded Transmission. Perhaps most surprisingly, for several cluster types the rates of

C1-change and C1-deletion were lower in the classroom than in the sound booth. In Wilson et al. (2014), we speculated that such modifications arose when listeners (in the sound booth) misperceived the place or other properties of a stop, or failed to detect the stop altogether. With respect to the classroom experiment, recall that the room used was relatively quiet and unoccupied, but the participant was about 3.3 m away from the computer speakers and a quiet HVAC hum was present during recording sessions. In addition, surfaces such as the glass wall and windows and the whiteboard are not absorptive materials and may lead to an increase in reverberation time and sound reflection (Seep et al., 2000). Accordingly, Degraded Transmission leads to the expectation that these characteristics of the classroom would mask the fragile stop burst cues even more extensively and increase the rates of C1 change (and of the already low-probability C1 deletion modification) in the classroom. The decrease in modifications of initial stops in the classroom indicates, to the contrary, that similar information is available in both environments and is adequate to recover the intended consonants.

The important difference between the two experiments, then, lies not in the availability of acoustic information, but rather in how listeners interpret that information given the embedding environment. In the optimal environment of the sound booth, listeners may be quite confident that the acoustic detail they hear is attributable to the speech sample (i.e. the stimulus they are tasked with shadowing). Difficulties in reproducing the nonnative speech emerge, however, because the consonant clusters are inconsistent with the native language sound system and hence the articulatory patterns necessary to produce them are unfamiliar. The participants' unpracticed production efforts can have negative results: in the case of the low-amplitude releases, for example, an imperfect response lacks an audible release altogether. Thus, in the sound-booth condition, greater certainty about the source of the cues that are present in the phonotactically illegal stimuli plausibly lead to responses that reflect a combination of (1) the difficulty of accurately implementing the articulatory trajectories necessary to match those nonnative acoustic cues, and (2) a tendency to resort to English-possible articulations that match those cues as closely as possible when the correct articulations are especially difficult.

In contrast, in the classroom environment, the same acoustic cues are available, but the participants may be less certain about them, or may assume that they have been distorted or affected by some aspect of the classroom acoustics. It follows that while the manipulations are still reflected in the responses to some extent (i.e. C1change and deletion are still more probable for low amplitude stop-initial stimuli, though implemented at lower rates than in the sound booth), the effect is attenuated. If participants' certainty about the nature or source of the acoustic cues is compromised, then they may make less of an attempt to reproduce them in their own responses. The consequence of this is that their responses generally reflect greater proportions of no modification, since they are now aiming to produce the phonemes without feeling obliged to replicate all of the nonnative acoustic detail, or epenthesis, which we have argued above is the default modification of nonnative sequences for English speakers since it preserves the consonants in the input and is compatible with native articulatory implementations (Abrahamsson, 2003; Davidson et al., 2015; Weinberger, 1994).

Why were some acoustic details, such as burst duration, imitated to similar degrees in the sound-booth and classroom conditions while others, such as POV, had diminished effects in the classroom environment? We believe that such selective imitation can be

traced to the native system of non-contrastive variation. Burst duration of voiceless stops varies considerably by speaker, prosodic context, and other factors within English (e.g., Theodore et al., 2009; Allen et al., 2003; Nielsen, 2011). Furthermore, English speakers show phonetic convergence effects involving voiceless stop burst duration, as discussed in Section III.2.e. It was therefore natural for participants in our condition to modulate the burst durations of voiceless stops, albeit in a novel phonotactic environment, to approximate perceived stimulus values. In contrast, many English speakers may have little experience with voicing starting before the onset of frication (if at all, e.g. Haggard, 1978; Smith, 1997), or with a higher intensity of voicing that decreases throughout the stop closure (again, if voicing occurs at all, e.g. Keating, 1984; Westbury, 1983). Imitation of POV may therefore be particularly challenging. In the sound booth, participants can be confident that POV is part of the target stimulus and therefore attempt to match it; when this is done erroneously, a reduced vocoid can result. In the classroom, participants may at least partly attribute POV to ambient noise, rather than to the speech signal, and for that reason down-weight or disregard it in planning their productions.

2 L2 speech processing in the classroom

In the introduction, we reviewed a number of studies of the effects of classroom acoustics on speech recognition. Most importantly for this study, the research pertaining to second language processing has demonstrated that speech perception and word recognition are hindered under classroom conditions; even reverberation time alone can negatively affect word recognition accuracy for L2 speakers (Nábělek and Donahue, 1984; Takata and Nábělek, 1990), though the effect of reverberation time is exacerbated when it interacts with multi-talker noise (Rogers et al., 2006; Shi, 2010). While it may be tempting to conclude from previous research that classroom acoustics have a detrimental effect on L2 learners' ability to perceive non-native sounds quite generally, the results of the current study are not strictly consistent with this interpretation. Instead, the classroom context has more subtle and complex effects on the perception and interpretation of acoustic detail that cannot be classified simply as degraded perception. One relevant consideration may be that there are substantial differences between the task in this study and much previous research: our task involved producing an unfamiliar word in a language that the participants had not learned before, whereas subjects in previous studies were proficient (or bilingual) speakers who had to match an English stimulus with a word within a set of possible responses. A forced-choice task with real words may be affected by factors other than acoustics, such as familiarity with a word or its frequency (Clopper et al., 2006; Shi, 2010).

Another possible consideration is that the simulated-classroom conditions under which reverberation and multi-talker babble were manipulated have a more detrimental effect than our study. The classroom where this study took place was not acoustically perfect, considering, for example, the glass walls and windows, the white board, and the distance of the participant from the speakers. However, it clearly does not represent the worst listening conditions that L2 learners might face. Future research is necessary to determine what combination of effects, including reverberation time, distance from the talker, occupied vs. unoccupied classrooms, masking noises such as multiple talkers or ventilation, etc. would give rise to measurably degraded perception of acoustic detail.

3 Relevance to naturalistic second language learning

A difficult aspect of L2 speech learning is acquiring the phonetic implementation of phonemes and sequences in the new language (Bradlow, 2008; Flege, 1987; Flege and Dravidian, 1984; Flege and Hillenbrand, 1984; Han et al., 2011; Hazan and Boulakia, 1993; Zampini, 2008). Importantly, successful L2 acquisition requires learners to determine both whether the acoustic detail that they are presented with corresponds to contrastive phonemic categories, and what the proper range of acoustic variability is for the phonological categories in the L2. This study, along with others in the literature, highlights the challenges that may arise in interpreting language-specific phonetic variability in an L2 context. For example, English speakers learning Russian should from a new phonological system that allows two stops to cluster together in an onset, and that requires such clusters to agree in voicing (Burton and Robblee, 1997). They should also learn that all stops must be released regardless of the following phonological context (Davidson and Roon, 2008; Zsiga, 2003), and that voiced stops are produced with phonation during the closure in initial position (Ringen and Kulikov, 2012). At the ends of words, however, they must learn that obstruents are (incompletely) devoiced (Dmitrieva et al., 2010; Kharlamov, 2014). A failure to learn these details of phonetic implementation could lead English speakers to substitute incorrect phonemes (such as voiceless unaspirated stops for prevoiced stops) or insert phonemes that change the contrastive structure of the word (such as inserting a reduced vowel after the burst of a stop-initial cluster). The inability to properly interpret the phonetic implementation and range of variability in a language could potentially have a cascading effect: misinterpretation of the phonetic details could lead to the postulation of inappropriate phonemic categories, ultimately affecting the lexical representations of L2 learners (see Hayes-Harb and Masuda, 2008 on lexical encoding of L2 contrasts).

The results of this study also suggest that the acoustic environment in which an L2 learner is presented with the new language's phonetic information influences how variability is processed. A challenge for second language teaching is to maximize a learner's ability to properly interpret and characterize phonetic detail while minimizing confusing or adverse environmental effects. As teaching in an ideal classroom is usually out of an instructor's hands, other assistance may be available. One such method, already a common practice in many types of L2 learning, is the use of orthography for languages in which this is helpful (Escudero et al., 2008; Rafat, 2015). As Showalter and Hayes-Harb (2015) point out, however, orthography is primarily effective when the orthographic systems of first langauge (L1) and L2 are relatively similar, or when the orthographies are transparent. If there are graphemes that do not correspond to phonemes in the lexical item, or if the orthographies are totally distinct (i.e. English vs. Arabic), then learners can be negatively impacted by orthography (Bassetti, 2006; Hayes-Harb et al., 2010; Showalter and Hayes-Harb, 2015). Thus, under some circumstances, orthography may provide useful information to the learner regarding the phonological categories that phonetic cues must be matched up to, but it is unlikely to be a successful technique in all cases.

Another potential avenue for improving L2 learners' chances of discovering the proper range of phonetic variability in their L2 is by presenting learners with input from multiple talkers. Several studies have shown that learners are able to improve their ability to discriminate between contrastive phonemes in a foreign language after being

trained on the contrast in a High Variability Phonetic Training (HVPT) paradigm, which is usually implemented by presenting stimuli recorded by multiple talkers (Bradlow et al., 1997; Iverson et al., 2005; Lively et al., 1993; Wang et al., 1999), though other studies have observed that HVPT may be most successful for learners with high perceptual aptitude (Perrachione et al., 2011; Sadakata and McQueen, 2014). In Davidson et al. (2015), we showed that when the stimuli from the current study were preceded with the same items produced by two other talkers, rates of no modification went up substantially while modifications like prothesis, C1 deletion, and C1 change decreased. Similarly, studies by Barcroft and Sommers showed that English speakers were better and faster at matching newly learned Spanish words to pictures and in L2-to-L1 recall when the words had been taught using multiple talkers as compared to a single talker (Barcroft and Sommers, 2005; Sommers and Barcroft, 2011). Although a few studies provide some evidence that some phonological contrasts may be more successfully facilitated than others using HVPT (Kingston, 2003; Wade et al., 2007), the preponderance of the data suggest that introducing new categories and lexical items by using multiple talkers leads to more accurate characterization of L2 phonetic cues and their variability.

V Conclusion

In line with previous research investigating speakers' sensitivity to phonetic detail in their attempts to interpret and reproduce nonnative speech (Davidson et al., 2015; Iverson et al., 2003; Wilson et al., 2014), the results of this study confirm that such sensitivity is still present to some degree in a classroom setting, which has been discussed in previous literature as a less optimal environment for transmission of speech sounds and second language learning (e.g. Boothroyd, 2004; Lecumberri et al., 2010; Nábělek, 1988; Nábělek & Donahue, 1984; Rogers et al., 2006). Counterintuitively, a quiet classroom may promote an environment that allows learners to react to acoustic detail in a way that is beneficial for second language learning. Whereas the ideal environment of the sound booth may encourage speakers to attempt reproduction of nonnative acoustic detail that is difficult to implement, thereby resulting in production errors, this effect may diminish in the classroom, where learners may be less certain about the source of the fine acoustic detail. Whether or not these results hold up in an occupied classroom with more salient ambient sounds, such as other talkers – perhaps an even more likely scenario for second language learning – is an area for future research. The present findings contribute both to the study of nonnative speech processing in the classroom, showing that this acoustic environment can have perhaps surprising effects on production of foreign sound patterns, and to the study of consonant cluster perception/production, indicating that some production modifications previously interpreted as originating in misperception instead result from failed attempts to match perceived phonetic detail.

Acknowledgment

The authors would like to thank Sean Martin for his longstanding assistance with this project, and Alice Hall, Francesca Himelman, Johnny Mkitarian, Elizabeth George and Steven Foley for their assistance in coding the data. We would also like to thank the members of the NYU Phonetics and Experimental Phonology Lab and the JHU Phonology/Phonetics Lab for their questions and

comments. This research program has benefited from discussions with audiences at various universities and conferences, and from the comments of the anonymous reviewers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by NSF grants BCS-1052855 to Lisa Davidson and BCS-1052784 to Colin Wilson.

Notes

1. Inserted vowels are represented with a superscript schwa /ᵊ/ to acknowledge the finding that they have somewhat different acoustic and articulatory properties from English lexical schwas in matched phonological environments (e.g. Davidson, 2006, 2010).
2. Although intensity perception is likely to be more sensitive to relative values (or changes) than to absolute values, for completeness we report decibel values as measured with Praat for the manipulated stop releases. For each relevant cluster type, we provide the mean and standard deviation in dB for the higher-amplitude manipulation followed by the same statistics for the lower-amplitude manipulation. SN voiceless: 64.1 (1.1), 56.4 (1.0); SN voiced: 73.7 (1.5), 67.3 (1.4); SS voiceless: 58.7 (2.5), 48.7 (2.5); SS voiced: 64.2 (1.8), 57.6 (1.7). The corresponding values for the (unmanipulated) consonant closures following the stop releases are: SN voiceless: 74.0 (1.1), 74.0 (1.0); SN voiced: 74.0 (1.4), 73.9 (1.4); SS voiceless: 36.2 (2.9), 35.9 (3.0); SS voiced: 63.8 (1.9), 63.8 (1.8). The difference between the measured relative amplitude of a stop release and the value specified by the stimulus manipulation was generally less than 1 dB (mean = 0.1, sd = 1.2).
3. Coefficients are reported as mean estimates and 95% HPD interval (in square brackets). According to the model fit, the posterior probability that the population coefficient value lies within the HPD interval is 0.95. P-values are calculated by estimating the posterior probability of coefficient values lying on the opposite side of zero from the mean.
4. In logistic regression models, independent effects multiply (rather than adding as in a linear regression) and predicted probabilities saturate near 0 and 1. To illustrate these properties, suppose that a particular response type has essentially zero probability of occurrence unless POV or some -other acoustic property is present, and then only in the sound-booth condition. A logistic model could match this pattern with a strong overall bias against this response type, together with weak independent effects of acoustics and condition that 'gang up' to raise the probability of the response above a near-zero value only when it is observed.
5. For the prothesis modification, there was a marginal effect of condition (condition × prothesis: –0.38 [–0.99, 0.01], $p = 0.09$) and a marginal interaction of condition and burst amplitude (condition × amplitude × prothesis: –0.38 [–0.99, 0.01], $p = 0.09$). However, the total number of prothesis responses for this cluster type was very low (7 in the sound booth, 5 in the classroom). In the text we do not discuss effects, like these, that pertain to less than 2% of the data.
6. Epenthesis was numerically more probable in the classroom study (condition × epenthesis: –0.26 [–0.51, 0.06], $p = .10$.). As discussed further in Section IV, this is part of a larger pattern of inflated epenthesis rates in the classroom experiment, perhaps because it is a default repair for nonnative clusters quite generally. The classroom condition also elicited an overall lower rate of initial stop deletion for the SN_{vcd} cluster type (condition × C1-deletion: –0.94 [–2.33,

–0.30], $p < .01$). This plausibly has the same explanation as the reduction in C1-change for SN_{vcls} clusters, with reduced imitation averting production failures. However, the rate of C1-deletion was quite low in both experiments; we should therefore be cautious when interpreting this difference.

7. It would be possible to evaluate phonetic imitation of the other acoustic manipulation that affected stop bursts, namely amplitude. Unlike the case of burst duration, however, we are not aware of previous work demonstrating that burst amplitude can be imitated under any conditions (e.g. with native words). Therefore, we leave this possible avenue open to future research.

References

Abrahamsson N (2003) Development and recoverability of L2 codas. *Studies in Second Language Acquisition* 25: 313–49.

Allen JS, Miller J, and DeSteno D (2003) Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America* 113: 544–52.

Barcroft J and Sommers M (2005) Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition* 27: 387–414.

Bassetti B (2006) Orthographic input and phonological representations in learners of Chinese as a foreign language. *Written Language and Literacy* 9: 95–114.

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

Best C, McRoberts G, and 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: 775–94.

Best C and Strange W (1992) Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics* 20: 305–30.

Best C and Tyler M (2007) Nonnative and second-language speech perception: Commonalities and complementarities. In: Munro M and Bohn O-S (eds) *Second language speech learning: The role of language experience in speech perception and production*. Amsterdam: John Benjamins, 13–34.

Boersma P (2001) Praat, a system for doing phonetics by computer. *Glot International* 5: 341–345.

Boothroyd A (2004) Room acoustics and speech perception. *Seminars in Hearing* 25: 155–66.

Bradley JS (1986) Predictors of speech intelligibility in rooms. *Journal of the Acoustical Society of America* 80: 837–45.

Bradlow A (2008) Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/–/l/ contrast. In: Hansen Edwards J and Zampini ML (eds) *Phonology and second language acquisition*. Philadelphia, PA: John Benjamins, 287–308.

Bradlow A, Pisoni D, Akahane-Yamada R, and Tohkura Y (1997) Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America* 101: 2299–2310.

Breen M, Kingston J, and Sanders L (2013) Perceptual representations of phonotactically illegal syllables. *Attention, Perception and Psychophysics* 75: 101–20.

Burton M and Robblee K (1997) A phonetic analysis of voicing assimilation in Russian. *Journal of Phonetics* 25: 97–114.

Clopper CG, Pisoni D, and Tierney A (2006) Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology* 17: 331–49.

Crandell C and Smaldino J (2004) Classroom Acoustics. In: Kent R (ed.) *The MIT encyclopedial of communication disorders*. Cambridge, MA: MIT Press, 442–43.

Cunnings I (2012) An overview of mixed-effects statistical models for second language researchers. *Second Language Research* 28: 369–82.

Curtin S, Goad H, and Pater J (1998) Phonological transfer and levels of representation: The perceptual acquisition of Thai voice and aspiration by English and French speakers. *Second Language Research* 14: 389–405.

Davidson L (2006) Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics* 34: 104–37.

Davidson L (2010) Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics* 38: 272–88.

Davidson L (2011) Phonetic, phonemic, and phonological factors in cross-language discrimination of phonotactic contrasts. *Journal of Experimental Psychology: Human Perception and Performance* 37: 270–82.

Davidson L and Roon K (2008) Durational correlates for differentiating consonant sequences in Russian. *Journal of the International Phonetic Association* 38: 137–65.

Davidson L, Martin S, and Wilson C (2015) Stabilizing the production of nonnative consonant clusters with acoustic variability. *Journal of the Acoustical Society of America* 137: 856–72.

Dmitrieva O, Jongman A, and Sereno JA (2010) Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *Journal of Phonetics* 38: 483–92.

Dufour S and Nguyen N (2013) How much imitation is there in a shadowing task? *Frontiers in Psychology* 4.

Escudero P, Hayes-Harb R, and Mitterer H (2008) Novel second-language words and asymmetric lexical access. *Journal of Phonetics* 36: 345–60.

Escudero P, Simon E, and 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: 280–88.

Flege J (1987) The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics* 15: 47–65.

Flege J (1995) Second-language speech learning: Theory, findings, and problems. In: Strange W (ed.) *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press, 229–73.

Flege J and Dravidian R (1984) Transfer and developmental processes in adult foreign language speech production. *Applied Psycholinguistics* 5: 323–47.

Flege J and 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: 729–40.

Flege J and Hillenbrand J (1984) Limits on phonetic accuracy in foreign language speech production. *Journal of the Acoustical Society of America* 78: 708–21.

Fowler CA, Brown JM, Sabadini L, and Weihing J (2003) Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language* 49: 396–413.

Goldinger S (1998) Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105: 251–79.

Hadfield J (2010) MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package. *Journal of Statistical Software* 33: 1–22.

Haggard M (1978) The devoicing of voiced fricatives. *Journal of Phonetics* 6: 95–102.

Hallé P, Best C, and Levitt A (1999) Phonetic vs. phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics* 27: 281–306.

Han J-I, Hwang J-B, and Choi T-H (2011) The acquisition of phonetic details: Evidence from the production of English reduced vowels by Korean learners. *Second Language Research* 27: 535–57.

Hayes-Harb R and Masuda K (2008) Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research* 24: 5–33.

Hayes-Harb R, Nicol J, and Barker J (2010) Learning the phonological forms of new words: Effects of orthographic and auditory input. *Language and Speech* 53: 367–81.

Hazan V and Boulakia G (1993) Perception and production of a voicing contrast by French–English bilinguals. *Language and Speech* 36: 17–38.

Hodgson M, Rempel R, and Kennedy S (1999) Measurement and prediction of typical speech and background-noise levels in university classrooms during lectures. *Journal of the Acoustical Society of America* 105: 226–33.

Iverson P, Hazan V, and Bannister K (2005) Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults. *Journal of the Acoustical Society of America* 118: 3267–78.

Iverson P, Kuhl PK, Akahane-Yamada R, et al. (2003) A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87: B47–57.

Keating P (1984) Phonetic and phonological representation of stop consonant voicing. *Language* 60: 286–319.

Kharlamov V (2014) Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: Phonological, lexical, and methodological influences. *Journal of Phonetics* 43: 47–53.

Kingston J (2003) Learning foreign vowels. *Language and Speech* 46: 295–349.

Klatte M, Lachmann T, and Meis M (2010) Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. *Noise and Health* 12: 270–82.

Knecht HA, Nelson PB, Whitelaw GM, and Feth LL (2002) Background noise levels and reverberation times in unoccupied classrooms: Predictions and measurements. *American Journal of Audiology* 11: 65–71.

Kuhl P, Conboy B, Coffey-Corina S, Padden D, Rivera-Gaxiola M, and Nelson T (2008) Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B* 363: 979–1000.

Larsen J, Vega A, and Ribera J (2008) The effect of room acoustics and sound-field amplification on word recognition performance in young adult listeners in suboptimal listening conditions. *American Journal of Audiology* 17: 50–59.

Lecumberri MLG, Cooke M, and Cutler A (2010) Non-native speech perception in adverse conditions: A review. *Speech Communication* 52: 864–86.

Lively S, Logan JS, and Pisoni D (1993) Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America* 94: 1242–55.

Müller EM and Brown WS (1980) Variations in the supraglottal air pressure waveform and their articulatory interpretation. In: Lass N (ed.) *Speech and language: Advances in basic research and practice: Volume 4*. Madison, WI: Academic Press, 318–89.

Nábělek A (1988) Identification of vowels in quiet, noise, and reverberation: Relationships with age and hearing loss. *Journal of the Acoustical Society of America* 84: 476–84.

Nábělek A and Donahue A (1984) Perception of consonants in reverberation by native and non-native listeners. *Journal of the Acoustical Society of America* 75: 632–34.

Nábělek A and Pickett JM (1974) Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation, and hearing aids. *Journal of the Acoustical Society of America* 56: 628–39.

Nelson P, Soli S, and Seltz A (2005) *Classroom Acoustics II: Acoustical Barriers to Learning*. , Melville, NY: Acoustical Society of America.

Nielsen K (2011) Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39: 132–42.

Perrachione TK, Lee J, Ha LYY, and Wong PCM (2011) Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America* 130: 461–72.

Picard M and Bradley J (2001) Revisiting speech interference in classrooms. *Audiology* 40: 221–44.

Plomp R, Steeneken HJM, and Houtgast T (1980) Predicting speech intelligibility in rooms from the modulation transfer function. II. Mirror image computer model applied to rectangular rooms. *Acta Acustica united with Acustica* 46: 73–81.

R Development Core Team (2012) *R: A language and environment for statistical computing*. R , Vienna: Foundation for Statistical Computing.

Rafat Y (2015) The interaction of acoustic and orthographic input in the acquisition of Spanish assibilated/fricative rhotics. *Applied Psycholinguistics* 36: 43–66.

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

Ringen C and Kulikov V (2012) Voicing in Russian stops: Cross-linguistic implications. *Journal of Slavic Linguistics* 20: 269–86.

Rogers C, Lister J, Febo D, Besing J, and Abrams H (2006) Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics* 27: 465–85.

Rojczyk A, Porzuczek A, and Bergier M (2013) Immediate and distracted imitation in second-language speech: Unreleased plosives in English. *Research in Language* 11: 3–18.

Sadakata M and McQueen J (2014) Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology* 5.

Seep B, Glosemeyer R, Hulce E, Linn M, and Aytar P (2000) *Classroom acoustics: A resource for creating learning environments with desirable listening conditions*. Melville, NY: Acoustical Society of America.

Shi L-F (2010) Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition. *Journal of Speech, Language and Hearing Research* 53: 821–35.

Shockley K, Sabadini L, and Fowler C (2004) Imitation in shadowing words. *Perception and Psychophysics* 66: 422–29.

Showalter CE and Hayes-Harb R (2013) Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research* 29: 185–200.

Showalter CE and Hayes-Harb R (2015) Native English speakers learning Arabic: The influence of novel orthographic information on second language phonological acquisition. *Applied Linguistics* 36: 23–42.

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

So CK and Best CT (2010) Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech* 53: 273–93.

Sommers M and Barcroft J (2011) Indexical information, encoding difficulty, and second language vocabulary learning. *Applied Psycholinguistics* 32: 417–34.

Strange W (2006) Second-language speech perception: The modification of automatic selective perceptual routines. *Journal of the Acoustical Society of America* 120: 3137.

Swan K and Myers E (2013) Category labels induce boundary-dependent perceptual warping in learned speech categories. *Second Language Research* 29: 391–411.

Takata Y and Nábělek AK (1990) English consonant recognition in noise and in reverberation by Japanese and American listeners. *The Journal of the Acoustical Society of America* 88: 663–66.

Theodore R, Miller J, and DeSteno D (2009) Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America* 125: 3974–82.

Trafimovich P (2005) Spoken-word processing in native and second languages: An investigation of auditory word priming. *Applied Psycholinguistics* 26: 479–504.
Wade T, Jongman A, and Sereno JA (2007) Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica* 64: 122–44.
Wang Y, Spence MM, Jongman A, and Sereno JA (1999) Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America* 106: 3649–58.
Weinberger S (1994) Functional and phonetic constraints on second language phonology. In: Yavas M (ed.) *First and second language phonology*. San Diego,CA: Singular Publishing Group, 283–302.
Westbury J (1983) Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *Journal of the Acoustical Society of America* 91: 2903–10.
Wilson C and Davidson L (2013) *Bayesian analysis of non-native cluster production*. In: Kan S, Moore-Cantwell C, and Staubs R (eds) *Proceedings of the Northeast Linguistics Society* 40. Amherst, MA: Graduate Linguistic Student Association, 265–78.
Wilson C, Davidson L, and Martin S (2014) Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language* 77: 1–24.
Zając M and Rojczyk A (2014) Imitation of English vowel duration upon exposure to native and non-native speech. *Poznan Studies in Contemporary Linguistics* 50: 495–514.
Zampini ML (2008) L2 speech production research: Findings, issues and advances. In: Hansen Edwards J and Zampini ML (eds) *Phonology and second language acquisition*. Philadelphia, PA: John Benjamins, 219–50.
Zsiga E (2003) Articulatory timing in a second language: Evidence from Russian and English. *Studies in Second Language Acquisition* 25: 399–432.

Appendix 1. Stimuli.

Sequence	Initial cluster	CC item	CVC item	VCC item
Fricative + Nasal	vm	vmafu	vemafu	
		vmage	vemage	
		vmado		evmado
		vmati		evmati
	vn	vnago	venago	
		vnaza	venaza	
		vnabe		evnabe
		vnadu		evnadu
	zm	zmagi	zemagi	
		zmasa	zemasa	
		zmabo		ezmabo
		zmaku		ezmaku
	zn	znagu	zenagu	
		znapo	zenapo	
		znade		eznade
		znaka		eznaka
Fricative + Stop	vd	vdafi	vedafi	
		vdato	vedato	
		vdagu		evdagu

Appendix 1. (Continued)

Sequence	Initial cluster	CC item	CVC item	VCC item
		vdapa		evdapa
	vg	vgafi	vegafi	
		vgase	vegase	
		vgabu		evgabu
		vgaka		evgaka
	zb	zbafo	zebafo	
		zbase	zebase	
		zbata		ezbata
		zbavi		ezbavi
	zg	zgade	zegade	
		zgafa	zegafa	
		zgaku		ezgaku
		zgapi		ezgapi
Stop + Nasal (Voiced)	bn	bnadi	benadi	
		bnapa	benapa	
		bnate		ebnate
		bnazo		ebnazo
	dm	dmago	demago	
		dmatu	dematu	
		dmabe		edmabe
		dmasa		edmasa
	gm	gmato	gemato	
		gmava	gemava	
		gmafu		egmafu
		gmape		egmape
	gn	gnavo	genavo	
		gnazi	genazi	
		gnake		egnake
		gnatu		egnatu
Stop + Nasal (Voiceless)	km	kmapo	kemapo	
		kmazu	kemazu	
		kmabi		ekmabi
		kmave		ekmave
	kn	knadu	kenadu	
		knafe	kenafe	
		knago		eknago
		knapi		eknapi
	pn	pnabu	penabu	
		pnata	penata	
		pnaso		epnaso
		pnave		epnave
	tm	tmaba	temaba	

(Continued)

Appendix 1. (Continued)

Sequence	Initial cluster	CC item	CVC item	VCC item
		tmafe	temafe	
		tmado		etmado
		tmavu		etmavu
Stop + Stop (Voiced)	bd	bdafa	bedafa	
		bdaki	bedaki	
		bdate		ebdate
		bdazo		ebdazo
	db	dbagi	debagi	
		dbazo	debazo	
		dbapu		edbapu
		dbate		edbate
	gb	gbake	gebake	
		gbaso	gebaso	
		gbadi		egbadi
		gbavu		egbavu
	gd	gdasu	gedasu	
		gdaza	gedaza	
		gdape		egdape
		gdavi		egdavi
Stop + Stop (Voiceless)	kp	kpabi	kepabi	
		kpazu	kepazu	
		kpaga		ekpaga
		kpavo		ekpavo
	kt	ktada	ketada	
		ktasi	ketasi	
		ktapu		ektapu
		ktaze		ektaze
	pt	ptage	petage	
		ptava	petava	
		ptako		eptako
		ptasi		eptasi
	tp	tpabe	tepabe	
		tpaki	tepaki	
		tpada		etpada
		tpafo		etpafo