Insertion of discrete phonological units: An articulatory and acoustic investigation of aphasic speech

Adam B. Buchwald
Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN and Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA

Brenda Rapp
Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA

Maureen Stone
Department of Biomedical Sciences and Orthodontics, University of Maryland Dental School, Baltimore, MD, USA

The traditional view that sound structure is mentally represented by discrete phonological units has been questioned in recent years. Much of the criticism revolves around the necessity of positing gradient or continuous sound structure representations to account for certain phenomena. This paper

Correspondence should be addressed to Adam B. Buchwald, Department of Psychological and Brain Sciences, Speech Research Laboratory, Indiana University, Bloomington, IN 47405, USA; E-mail: abuchwal@indiana.edu

This research was supported by the IGERT Program in the Cognitive Science of Language at Johns Hopkins University, National Science Foundation Grant 997280. Preparation of the paper was supported by National Institutes of Health Grant awarded to David Pisoni. Portions of these data were presented at the Speech Error Workshop at MIT (June, 2005) and to audiences at Indiana University, Ohio State University, and Johns Hopkins. Portions of this work were submitted as part of the first author’s doctoral dissertation (Johns Hopkins University, 2005).

The authors would like to thank Paul Smolensky for invaluable intellectual contributions to this project. Additionally, this work has benefited from the helpful comments of Melissa Epstein, Marianne Pouplier, Justin Halberda, Gregory Ball, Susannah Levi, David Caplan and two anonymous reviewers. We thank Joanna Kochaniak and Ebonye Gussine for comments and assistance with various portions of this work, and David Pisoni for comments and providing a supportive environment during the preparation of this paper. Most importantly, we thank VBR for her willing and enthusiastic participation in this study.

© 2007 Psychology Press, an imprint of the Taylor & Francis Group, an Informa business
http://www.psypress.com/lcp DOI: 10.1080/01690960701273532
presents evidence in favour of discrete sound structure units in addition to gradient representations. We present a case study of aphasic speaker VBR, whose spoken language production errors include vowel insertions in many word-initial consonant clusters (e.g., \textit{bleed} $\rightarrow$ \textit{[b3lid]}). An acoustic and articulatory study is reported comparing the inserted vowels with lexical vowels in similar phonological contexts (e.g., \textit{believe}). The results indicate that these two vowels come from the same population, suggesting discrete insertion of a unit the same size as those used to represent lexical contrast. The implications of these data for theories of sound structure representation are discussed.

**INTRODUCTION**

Traditional descriptions of sound structure in linguistics posit a level of mental representation in which speakers represent the constituent sounds of a word as a string of abstract and discrete sound units (e.g., Chomsky & Halle, 1968). For example, the representation of the word \textit{bleed} is argued to consist of four segments, /b/, /l/, /i/, and /d/, and the contrast between \textit{bleed} and \textit{breed} (/b/, / /r/, /i/, /d/) is represented by the difference in the second segment of these forms. Thus, segments act as discrete units for representing lexical contrast. This type of representation has been widely adopted by psycholinguistic accounts of spoken language production as well (e.g., Butterworth, 1992; Dell, 1986, 1988; Garrett, 1980; Goldrick & Rapp, 2007; Levelt, Roelofs, & Meyer, 1999; Roelofs, 1997; Shattuck-Hufnagel, 1987; Stemberger, 1985). Over the past two decades, however, experimental language research measuring articulatory or acoustic properties of spoken language has been used to argue against the existence of discrete representations as active units of psycholinguistic processing (Goldinger, 1998; Pierrehumbert, 2001; see Port & Leary, 2005, for a recent formulation of this argument). These studies have appealed to the incompleteness of phenomena traditionally used to support the existence of discrete phonological processes (as in word-final devoicing, Port & O’Dell, 1985; Port & Crawford, 1989), and to findings that speech errors traditionally transcribed as segmental deletions may result from articulatory gestures associated with multiple different segments being produced simultaneously (e.g., Pouplier, 2003; see Pouplier & Hardcastle, 2005, for a recent review). This evidence reveals that an account of phonological processing requires an appeal to detailed information regarding temporal dynamics of articulation, and that speech errors cannot all be accounted for simply by manipulating the presence or identity of discrete units. In our view, however, although these data support the inclusion of gradient or continuous representations that may interact with motor planning and implementation systems involved in speech production, they do not eliminate the possibility of an additional level of discrete or categorical sound structure representation (for a similar argument, see Frisch & Wright, 2002).
The main issue addressed here is whether there is evidence supporting a level of discrete representation active during spoken language production. In this paper, we present laboratory data analysing the spoken production errors of an aphasic speaker, VBR, who typically inserts a vowel in word-initial consonant clusters (e.g., *bleed* → [bʌlid]). We report on acoustic and articulatory analyses indicating that these spoken production errors are best described as insertions of discrete units of phonological contrast (in this case, the vowel segment schwa), and not as the result of mistiming the articulatory gestures associated with the target form. We argue that the existence of discrete insertion errors reveals a level of representation and processing in which sounds are represented as discrete, categorical units which may be individually manipulated.

The research study reported below integrates methodological and theoretical perspectives from a variety of cognitive scientific approaches to language. To ground this work, we motivate the use of data from brain-damaged individuals to address these issues, and then we review relevant findings and claims from previous research on vowel insertion.

**Aphasic speech and spoken language production**

Jakobson (1941/1968) famously argued that the same principles of phonological complexity that constrain the cross-linguistic distribution of sound patterns also constrain the patterns observed in aphasia. For Jakobson, this proposal entailed that patterns of performance from aphasic speakers can provide insight into the nature of phonological knowledge. This notion dovetails with the key assumptions and goals of cognitive neuropsychology; as argued by Caramazza (1986) among others, one main objective of cognitive neuropsychology is to explicitly articulate the nature of mental representations and processes that underlie cognitive abilities such as spoken language production (see Rapp & Goldrick, 2006, for a review of the contribution of cognitive neuropsychology research to our understanding of spoken production). Examining aphasic speech helps address issues of representation in speech production by identifying the level at which errors arise, and then considering the possible type(s) of mental representation or processes that give rise to the patterns of errors that are observed.

Previous investigations of aphasic speech errors have typically relied on transcriptions of the spoken productions. Several of these studies have reported that aphasic speech errors appear to be influenced by the same regularities of linguistic structure and phonological complexity that constrain natural human languages (e.g., Blumstein, 1973; Romani & Calabrese, 1998) and they appear to be produced by the same mechanisms that produce speech errors in unimpaired subjects (e.g., Buckingham, 1980, 1986).
Hardcastle and Edwards (1992) reported results from electropalatography (EPG) studies of brain-damaged speakers which exposed the limitations of transcription on the errors in this population and favours the use of articulatory and acoustic analysis of aphasic speech. In particular, they cautioned that in transcription, errors that arise in articulatory timing of appropriately selected segments may look like errors in selecting appropriate units for production of a word (e.g., segmental substitution errors), and that the former is often mistaken for the latter. For example, there are (at least) two possible explanations for why a /k/ may be perceived as the initial consonant in an erroneous production of the word *tick*. One explanation is that the speaker has erroneously activated /k/ in onset position, either during the process of retrieving the long-term memory representation of the word or in subsequent phonological encoding, and then produced the appropriate motor plans and implementation of the intended form. This is a common interpretation of errors perceived as segmental substitutions. However, another possibility is that the /t/ and /k/ are both correctly activated in phonological encoding, but an error at the level of motor planning or implementation leads to an error in the serial ordering of these gestures and they are produced simultaneously. This simultaneous alveolar/velar closure may be perceived as a /k/ (due to the velar closure providing the first obstruction in the air flow through the vocal tract). In this case, the same perception of a segment-sized substitution arises (substitution of /k/ for /t/ in word-onset position), but there is a clear difference in both the aetiology and the articulatory patterns that generate these errors. Using EPG, Hardcastle and Edwards (1992; also see Wood & Hardcastle, 2000) identified the existence of the latter type of error in the productions of at least certain types of aphasic speakers.

The work of Wood and Hardcastle underscores the importance of directly examining the articulatory movements associated with aphasic speech errors (and perhaps with all speech errors) in identifying the nature of the error, as is done in this paper. In the next section, we focus on possible accounts of vowel insertion that an acoustic and articulatory examination will allow us to test.

**Inserted vowels**

This section discusses four possible accounts of inserted vowels in speech production. One account holds that inserted vowels in speech production arise from epenthesis of a discrete vowel segment; two additional accounts based in the sound structure representations of Articulatory Phonology (Browman & Goldstein, 1986, 1989, 1990) hold that inserted vowels arise from the mistiming of articulatory gestures coupled with vocal tract
dynamics; and a final account holds that the vowels could arise due to impairment at the level of articulatory implementation.

Browman and Goldstein (1990, 1992b) considered the possibility that inter-consonantal schwas in English (i.e., the schwa in the initial syllable of *cologne*) do not require their own gesture or underlying representation, and that the acoustic derivation of schwa can arise from variation in the temporal coordination of the flanking consonants. However, contrary to this strong proposal, Browman and Goldstein (1992b) reported x-ray tracings evidence that there is indeed an articulatory target for schwa in American English which cannot be determined from production of the adjacent gestures alone. These results revealed that inter-consonantal schwa can indeed function as a unit of lexical contrast in American English; that is, the gestural specification of inter-consonantal schwa can be used to distinguish between two forms constituting a minimal pair (e.g., *clone* ~ *cologne*). The present work builds on this finding. In particular, we assume that evidence indicating that VBR’s inserted vowel is comparable to her lexical schwa constitutes evidence that the inserted schwa has an articulatory target, and does not arise from mistiming other articulatory gestures.

The first account we consider here is *schwa epenthesis*, in which inserted vowels arise due to the insertion (or epenthesis) of a discrete sound structure element – schwa. According to this account, a form with schwa insertion between two consonants in a cluster (e.g., *clone* → *[kɔlɔn]*) differs from its target ([klɔn]) in precisely the same way that a form with that vowel in its lexical specification (e.g., *cologne*) differs from the consonant cluster target. Many previous studies of vowel insertion in speech production have focused on identifying patterns of insertion in second language learners. These studies have reported that schwa epenthesis is a common ‘correction’ in the production of non-native consonant clusters that are phonotactically ill-formed in the speaker’s native language (Broselow & Finer, 1991; Davidson, 2003; Davidson, Jusczyk, & Smolensky, 2003; Eckman & Iverson, 1993; Hancin-Bhatt & Bhatt, 1998). The inserted vowel may be a schwa, as reported by Davidson et al. (2003) for English speakers producing Polish clusters (e.g., *zgomu* → *[zɡɔmu*]; schwa was also reported for Korean speakers producing English clusters, Tarone, 1987), but languages without schwa in the inventory may use a different epenthetic vowel (e.g., *[i]*) for Brazilian Portuguese, Major, 1987). Schwa epenthesis has traditionally been described as a discrete phonological process such that a target sound structure representation that contains an illegal consonant cluster (e.g., *[C₁C₂V…]*) is mapped to a different representation that contains a vowel (e.g., *[C₁ɑC₂V…]*)). The schwa epenthesis account of VBR’s vowel insertion errors is depicted in (1). On this view, vowel insertion is the result of a categorical repair of sound structure – epenthesis of a discrete vowel unit, and the output on the right-hand side of (1) is identical to the target for a
word with lexical schwa. Thus, if VBR’s inserted vowel is the result of schwa epenthesis, we expect it to be indistinguishable from lexical schwa in the same phonological context.

\(1\) Schwa epenthesis

One line of evidence from acoustic work that indirectly questions the notion that vowel insertions arise from epenthesis comes from Price (1980), who noted that lengthening a \(C_2\) liquid in a consonant cluster creates the percept of a schwa (e.g., [pl:] perceived as [psl]). This finding motivated Davidson’s (2003); also see Davidson & Stone, 2004) investigation of schwa insertion in non-native clusters, to see whether this schwa results from phonological epenthesis, or from the mistiming of articulatory gestures (as in Browman & Goldstein, 1986, 1989, 1990) associated with producing the target consonants. Davidson and Stone (2004) investigated the production of forms containing non-native fricative-stop clusters (e.g., \(z\)gomu) by English speakers who appear (acoustically) to insert schwa to break up the illegal cluster (e.g., /zɡ/ \(\rightarrow\) [zɡ]). To assess whether the schwa in the acoustic form (e.g., \(z\)ɡomu) resulted from phonological epenthesis, they used ultrasound imaging to compare tongue movements on insertion tokens with tongue movements involved in producing two similar English words that differ in that one has a cluster (e.g., succumb), and the other has a schwa between the same two consonants (e.g., scum). They argued that if the tongue movements of zgomu (acoustically, [zɡomu]) are more like succumb, then the schwa present in the acoustic wave form is likely the result of phonological epenthesis; however, if the tongue movements are more similar to those from scum, then the acoustic schwa is likely the result of a mistiming of the articulatory gestures associated with the production of the consonants. Davidson and Stone (2004) reported that the tongue movements during production of the inserted schwa in non-native clusters were closer to scum more often than to succumb, and contended that these vowel insertion errors were therefore the result of gestural mistiming, or a ‘pulling apart’ of the articulatory gestures associated with the /\(z/\) and /\(C/\) in the /\(zC/\) sequences.\(^1\)

\(^1\) Davidson (2003) argued that CC mistiming results from a grammatical process; constraints on gestural coordination and alignment generate an articulatory plan in which the degree of overlap between the two consonants leads to voicing between the release of \(C_1\) and the target of \(C_2\) (Davidson, 2003). Thus, the appearance of the inserted vowel results from a systematic ‘repair’, with constraints acting on dynamic gestural representations (following Gafos, 2002) rather than only on discrete segmental representations.
Consonant-consonant (CC) mistiming as proposed in these studies is depicted in (2). CC mistiming predicts clear articulatory differences in the production of lexical schwa and of inserted vowels.

(2)  *Consonant-Consonant (CC) mistiming*

```
[diagram showing CC mistiming]
```

A second account of vowel insertion based in gestural mistiming comes from Hall (2003), who presented ‘intrusive’ vowels as resulting from gestural mistiming. Intrusive vowels appear in consonant clusters containing a sonorant consonant, where the stressed vowel adjacent to a sonorant intrudes between the two consonants. Hall argues that they are ‘copies’ of the vowel adjacent to the sonorant, though they may be transcribed as schwa. Hall’s mistiming proposal – referred to here as Consonant-Consonant-Vowel (CCV) mistiming is depicted in (3).

(3)  *Consonant-Consonant Vowel (CCV) Mistiming*

```
[diagram showing CCV mistiming]
```

Hall proposes several diagnostics for distinguishing vowels resulting from CCV mistiming from vowels that result from phonological epenthesis. For the present purposes, Hall’s most relevant criteria for determining intrusive vowels are: (a) they appear in clusters containing a sonorant; (b) they share acoustic properties of the stressed vowel adjacent to the sonorant; and (c) they are more variable in length than epenthetic vowels and tend to disappear in fast speech. The present experiment focuses on vowel insertion errors in clusters containing a sonorant (a), and we test criteria (b) and (c) in the acoustic portion of the investigation presented in this paper. In addition, the CCV mistiming proposal also predicts articulatory differences between inserted vowels and lexical schwa.

The fourth possibility we discuss which has not been typically raised in the literature on inserted vowels is that VBR’s inserted vowels arises due to ‘articulatory noise’ which affects motor implementation, and is not a repair of sound structure per se. If noise in VBR’s articulations is responsible for generating the inserted vowel, we would expect a great deal of variability (due to the noise) in many facets of her speech. Nevertheless, given the same
phonological context, we should still see differences between inserted vowels and lexical schwa, as noise at the level of articulation is applied to different sound structure representations for each form.

This section reviewed four possible accounts of VBR’s vowel insertion: (1) schwa epenthesis; (2) CC-mistiming (as in Davidson, 2003); (3) CCV-mistiming (as in Hall, 2003); and (4) articulatory noise. The study described below was designed to identify which type of vowel insertion best describes the errors of the aphasic speaker under investigation, and detailed predictions from each of these accounts are presented in the experimental section of this paper. If either of the two mistiming accounts are the best account of the error, it would require that these errors are instituted at a level where the representations include information regarding the temporal dynamics – the coordination of gestures. However, if schwa epenthesis provides the best account of the error, then we may infer that vowels are represented as discrete units at the level at which the repair is instituted. Thus, the latter result providing evidence that supports the discrete process of vowel epenthesis would reveal the existence of discrete sound structure representations. However, while failure to find such evidence would suggest that the errors arise at a level at which sound structure information is represented dynamically, such results would not rule out the existence of a discrete representational level.

The current study

The work presented below investigated the articulation and corresponding acoustics of the speech production errors of aphasic speaker VBR. In particular, the study was designed to gain insight into the nature of VBR’s vowel insertion errors (e.g., bleed → [bəlid]). Uncovering the nature of VBR’s errors (or ‘repairs’) permits us to constrain theories regarding the nature and content of representations in the spoken production system at the level of her deficit. In particular, we addressed whether the repair involves: (1) a categorical change in production (schwa epenthesis), implying that the error arises at a part of the cognitive system where discrete entities may be manipulated; (2) a change along a temporal dimension (CC mistiming or CCV mistiming), implying that discrete sound structure units need not be represented at the level at which these errors arise; or (3) noise in the articulatory system, such that the error arises at a motor implementation level and is not a repair instituted at a level of phonological processing.

---

2 This possibility does not preclude the representation of dynamic gestural information, as the results of Davidson’s (2003) study clearly indicate that this type of information is represented in the speech production system. However, finding that vowel epenthesis is the best account does require that vowels are represented as units. This issue will be addressed further in the General Discussion.
Case report: VBR

VBR is a 58-year-old right-handed woman who suffered a cerebral-vascular accident (CVA) six years prior to the onset of the current investigation (2/2004). Magnetic resonance imaging scans reveal a large left hemisphere fronto-parietal infarct involving posterior frontal lobe, including Broca’s area, pre- and post-central gyri and the supramarginal gyrus (see Figure 1). VBR has a right hemi-paresis as a result of the CVA; she occasionally uses support to walk, and has limited use of her right arm below the elbow. The CVA also induced strabismus, for which she wears corrective lenses. Prior to her CVA, VBR was the president of a small company. VBR’s language production skills are severely impaired as a result of the CVA, particularly her spoken output.

VBR’s single word comprehension is relatively intact. On the Peabody Picture Vocabulary Test (PPVT-R; Dunn & Dunn, 1981) she scored in the 75th percentile (raw score = 166/175, form M). VBR also correctly matched 14/15 pictures to reversible sentences presented auditorily. VBR’s spelling of single words is moderately impaired; she accurately spelled 71% (39/55) of words from the Length List of the JHU Dysgraphia Battery (Goodman & Caramazza, 1985; see Buchwald & Rapp, 2004 for more information regarding VBR’s spelling deficit).

Figure 1. Left sagittal MRI image of VBR’s lesion.
Localising the deficit in the speech production system

To use VBR’s data to inform theories of language production, it is important to determine the level within the processing system at which her errors arise. This paper is concerned with language production and speech errors that arise during phonological processing. Therefore, it is critical to verify that the errors we are analysing arise at that level. One alternative possibility with respect to VBR’s spoken language production errors is that the errors are reducible to errors in VBR’s auditory perception (e.g., clone is perceived as cologne, and then produced ‘accurately’ as cologne). To address this, we will consider VBR’s performance on tasks that require intact speech perception processing. A second alternative to phonological processing errors is that the errors arise due to impairment in lexical access, or activating the correct target word in production (e.g., when asked to repeat clone, the speaker perceives the sounds correctly but activates the word cologne for production). This issue will be addressed in two ways. First, lexical decision tasks will verify that VBR’s lexical access is intact in spoken word recognition. Second, following Goldrick and Rapp (2007), we assume that accessing the appropriate word for production is required for picture-naming tasks but not for repetition tasks in which the form is presented to the subject (this may be particularly clear when one considers nonword repetition). Therefore, evidence indicating quantitatively and qualitatively similar performance on spoken language production tasks with various types of input – picture naming, repetition and reading – will verify that VBR’s lexical access for spoken language production is not the source of her errors.

VBR was administered two tests to address whether her spoken production errors could arise from incorrect perceptions: the PALPA (Kay, Lesser, & Coltheart, 1992) word same-different discrimination task, and the PALPA nonword same-different discrimination task. In these tasks, the experimenter reads two words (or two nonwords) approximately 1 second apart, and the subject responds whether the two words or nonwords are the same (e.g., word: house-house; nonword: zog-zog) or different (e.g., word: house-mouse; nonword: zog-zog). VBR’s performance was nearly flawless on both the word task (71/72; control subjects = 70.4/72) and the nonword task (71/72; no norms are provided). These results indicate that VBR’s impaired performance on spoken word repetition tasks is unlikely to be due to a problem in parsing auditorily presented linguistic input. In addition to the PALPA test, VBR performed word and nonword minimal pair discrimination for 10 pairs that differed only in the presence or absence of schwa in a consonant cluster (e.g., word: clone-cologne; blisk-belisk) – she performed flawlessly on these pairs.

To address whether her performance could be due to a deficit in lexical access in spoken word recognition, VBR was administered the auditory
lexical decision component of the PALPA. In this task, the experimenter reads a stimulus form (e.g., [tonæko]), and the subject is instructed to identify the stimulus as either a word or a nonword. VBR’s performance on lexical decision was within the normal range for both nonwords (78/80 correct; control subjects = 76) and words (79/80; control subjects = 79.4). This suggests that her ability to correctly identify spoken words and nonwords remains intact and performance problems in repetition tasks are not likely to be due to errors in identifying the target word.

To address VBR’s ability to perform lexical access required for spoken language production, VBR was administered 33 pictures for naming, and the same words were given in both reading and repetition tasks. Her performance reveals quantitatively similar impairment on each task: naming task (64% words correct; 85% phonemes correct); reading (67% words correct; 85% phonemes correct); repetition (67% words correct; 86% phonemes correct). Importantly, errors on these tasks are qualitatively similar as well, consisting of phoneme substitutions (gun → [kan]), deletions (shoulder → [ʃəʊldər]), or some combination of the two (pumpkin → [pʌmpki:n]). VBR’s erroneous output resulted in lexicalisations in 2 of the 22 incorrect pronunciations produced across the three tasks, each of which involved the substitution of a single phoneme (vase → face; kite → cat). In addition to these tasks, VBR was presented with a list of nonwords for repetition. The nonwords were assembled with the same segments (and syllables, as much as possible) as the 33 words in the list discussed above, and VBR correctly repeated 20/33 nonwords (61%). In terms of phonemic accuracy, VBR’s repetition performance with these nonwords is statistically indistinguishable from those reported above (82% phonemes correct, $\chi^2 = 0.69, ns$). These findings demonstrate that VBR’s deficit leads to qualitatively and quantitatively similar performance on tasks that require activation of target words (picture naming and reading) and tasks that do not require the subject to activate the correct target word (word and nonword repetition), indicating that VBR’s spoken production errors are not errors of lexical access.

Taken together, the two lines of evidence presented in this section indicate that VBR’s impairment affects a level of phonological processing involved in spoken language production, and not a level involved in processing input (such as speech perception, reading, etc.).

**Articulatory factors**

VBR’s articulation was assessed by a speech language pathologist as mildly impaired. On a battery of tests designed to assess the strength and mobility of the articulators, the following results were obtained. VBR showed a ‘mild’ asymmetry when asked to close her mouth and pucker her lips (right side), and a ‘mild’ slowness when asked to protrude and retract her
tongue three times in rapid succession. Additionally, tests of tongue strength revealed that her right side was mildly weaker than her left side. No other tests of strength or mobility of the articulators revealed abnormality. In particular, VBR was unimpaired in her ability to move her tongue towards various locations in and outside of her mouth (e.g., upper right of mouth), to move her tongue in specified trajectories (e.g., circle tongue around upper and lower lips), or to hold her tongue still with no movement or tremor. In addition, she was unimpaired in moving her lips in specified ways (e.g., pucker and retract), and was able to form a tight bilabial seal with cheeks inflated and nasal respiration. On diadochokinetic tests involving rapid repetition of /p/, /t/, and /k/ for 10 seconds, VBR produced 48 /p/s, 46 /t/s and 36 /k/s, with the performance on /k/ indicating a very mild slowness. Her performance on a sequence production task (produce /p t k/ for 10 seconds) showed a moderate deficit, as she only produced 3 accurately in the 10 second span.

It is crucial to consider the possible implications of these data for the present investigation. The most problematic possibility for this work is that the errors under investigation may arise at the level of articulation (and that the spoken production impairment is not indicative of errors generated at a level of linguistic processing). This possibility is considered below, as the ultrasound investigation directly addresses the question of whether VBR’s vowel insertion errors are simply the result of ‘noise’ in the articulation. The study investigates the articulation of the vowel she inserts in bleed (i.e., [b3lid]) and the vowel in a word that contains a schwa between the same two consonants (e.g., believe → [b3liv]), a pattern which was not observed in a premorbid speech sample. If her production errors are the result of a motor implementation problem, we would not expect the vowels in these two forms to be articulatorily and acoustically similar across a large number of trials.

**Lexical factors**

Consistent with the findings of Goldrick and Rapp (2007) regarding phonological processing deficits affecting performance on naming and repetition tasks, VBR’s repetition appears to be largely insensitive to lexical factors such as frequency. On a sample of 494 words ranging from four to six phonemes in length, VBR repeated 131 (26.5%) correctly. The frequency of each word was computed using the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), and a Pearson’s correlation was computed to determine whether lexical frequency and percentage of phonemes correct per word were correlated. The results of this analysis indicate that lexical frequency and VBR’s repetition accuracy are not significantly correlated variables ($r = -0.38$, $ns$). A second analysis was performed on a word list ($N = 100$) comparing high- and low-frequency words that were matched for
word stress, number of phonemes, and number (and type) of onset consonant clusters. The list was administered twice, and the performance on the two administrations was statistically similar. In a comparison of word accuracy collapsed across both administrations, VBR performed similarly on each group, correctly producing 43/100 high-frequency words, and 41/100 low-frequency words ($\chi^2 = 0.02, \text{ns}$). There was also no difference between the two groups in segment accuracy, with high-frequency words produced with 84.4% phoneme accuracy (428.5/508) and 83.1% segment accuracy for low-frequency words (422/508; $\chi^2 = 0.22, \text{ns}$).

**Sublexical factors**

VBR’s performance displays a particular sensitivity to the syllabic complexity of the word being produced; on an initial test containing 79 words with word-initial consonant clusters, VBR produced only 22 (27.8%) of the onset clusters appropriately. The majority of the remaining clusters (43/57; 75.4%) were produced with a vowel inserted in the consonant cluster (e.g., *bleed* → *b3lid*). Her performance on singleton onset consonants is significantly better; the onset consonant was correctly produced in 133/150 (88.7%) of words, significantly more accurate than words with cluster consonants ($\chi^2 = 8.85, p < .01$). In addition to these repetition tasks, VBR was presented with 20 pictures to name where the target name contained a consonant cluster (e.g., *broom, glass*). The tendency to insert vowels into consonant clusters was noted in this task as well (14/20 insertions, 70%; also in a reading task, VBR produced 14/20 insertions, 70%). The study reported below explores VBR’s performance on consonant clusters in more detail.

Two important exceptions to VBR’s pattern with consonant clusters were noted. First, her production of words with /s/-initial consonant sequences. The syllabification of words with /s/-initial clusters has been debated, and the prevailing analysis assumes that /s/ is extrametrical, and not part of the onset in syllabification (see Barlow, 2001 for a discussion). VBR’s performance on these words is difficult to quantify. In words with /s/ followed by one other consonant, she often produces both consonants, but tends to extend the articulation of /s/ for several seconds before producing the remainder of the word (and sometimes produces an extended /S/ instead of the extended /s/). This type of evidence may suggest the veracity of an extrametrical analysis of /s/, but the lack of a consistent pattern coupled with the difficulty in assessing the quality of this error (and her frustration for being asked to produce these sequences) limits the possibility of assessing these productions. Given this limitation, words with /s/-initial clusters were not included in further testing and analysis.

A second deviation from this pattern comes from word-initial sequences with a consonant followed by /j/ followed by a vowel (e.g., *cute* [kjut]). In
these sequences, VBR systematically deleted [j] from the word being produced (e.g., cute → [kut]). Buchwald (2005a) presents an account of this pattern that attributes [j] deletion to a difference between these sequences and the consonant clusters examined in this paper.

VBR was also administered a short list comparing words of high- and low phonotactic probability, which has been shown to influence both spoken word recognition (Vitevitch & Luce, 1998; Vitevitch, Luce, Pisoni, & Auer Jr., 1999) and spoken word production (Vitevitch, Armbrüster, & Chu, 2004). Phonotactic probability is a measure of the frequency with which a segment (or sequence of segments) occurs in the language (Jusczyk, Luce, & Charles-Luce, 1994). She was administered a list of CVC words (N = 28) contrasting high and low phonotactic probability. She performed equally well on both groups of words (12/14 words correct), making a de-voicing error (e.g., bat → [pæt]) and a vowel identity error (e.g., kite → [kæt]) on each list. Thus, given a list of relatively simple (CVC) words, VBR does not show an effect of phonotactic probability on her speech production accuracy. It remains possible that phonotactic probability effects are not seen here because of VBR’s reasonably good performance on monosyllabic CVC words.

ARTICULATORY AND ACOUSTIC INVESTIGATION

As indicated above, VBR’s productions of English words (and nonwords consistent with the phonotactics of English) with word-initial consonant clusters often contain a vowel inserted between the two consonants (e.g., bleed → [bʌlɪd]). The experiment presented in this section included both acoustic and ultrasound imaging components designed to investigate the nature of the repair that leads to VBR’s vowel insertion in consonant clusters. The acoustic component compares lexical schwa (as in believe) with the inserted vowel to determine whether they differ on three key dimensions: degree of coarticulation with the stressed vowel, duration, and overall variability in duration. These parameters were selected as they are direct tests of the CCV mistiming hypothesis (as stated by Hall, 2003). The ultrasound imaging component of the experiment compares the tongue shapes associated with VBR’s production of words with a lexical schwa (e.g., believe) with those of words with the inserted vowel (e.g., bleed → [bʌlɪd]).

Ultrasound imaging has been a useful tool for investigating tongue shapes – both sagittal and coronal slices – in speech production (Stone, 1991, 1995; Stone, Faber, Rafael, & Shawker, 1992; Iskarous, 1998; Davidson, 2003; Davidson & Stone, 2004; Gick & Wilson, 2004; see Davidson, 2005 for a recent review). Ultrasound imaging provides researchers with very good spatial resolution (~1 mm) and good temporal resolution (33Hz), and is
non-invasive and safe for participants (see Epstein, 2005 for a review), particularly when compared to x-ray imaging techniques. Ultrasound images are reconstructions of echo patterns from ultra-high frequency sound that are both emitted and received by piezoelectric crystals contained in a small hand-held transducer. In linguistic research, the transducer is typically placed under the participant’s chin, and the sound reflects off tissue boundaries. The area of interest here is the tissue/air boundary on the upper surface of the tongue, which appears as a bright white line (see Figure 2).

Predictions

The study reported below was designed to determine the nature of VBR’s vowel insertion errors. If the vowel is inserted as part of a schwa epenthesis repair process (depicted in (1)), there should be a clear pattern of results in each portion of the study as these two vowel types come from the same ‘population’. In the acoustic analyses, lexical and inserted vowels should be similar in their degree of coarticulation with the stressed vowel, and in their overall duration and duration variability. In the ultrasound imaging study, we should see that the production of the inserted vowel is similar to lexical schwa, and that differences between the tongue contours associated with the inserted vowel and lexical schwa are not greater than the differences among different tokens of lexical schwa or the differences among different tokens of inserted schwa. Note that these predictions are all supported by the absence of significant differences. To address this issue, we extracted tongue contours representing the $C_1$ and $C_2$ consonants as well as the tongue contours

![Figure 2](image-url)
representing the two vowel types. If VBR’s vowel insertion results from vowel epenthesis, we expect the differences between inserted vowel tongue contours and the consonant tongue contours to be greater than the differences between the inserted vowel and the lexical schwa. Further, we expect similar differences when we compare the inserted vowel to the two consonants and when we compare lexical schwa to the two consonants.

As discussed above, we tested two additional accounts of vowel insertion in which inserted vowels arise from a change in the timing of articulatory gestures. Under one mistiming proposal – CC mistiming (depicted in (2) above) – the coordination of the two consonants is misaligned and the gestures are not fully overlapped, leading to a period during which the vocal tract is open and phonation is occurring, and the schwa that is present in the acoustic record may be a consequence of this vocal tract configuration and timing relationship (Davidson, 2003). If this is the repair strategy used by VBR, there should be clear differences between the two vowels in the ultrasound imaging study. In particular, the differences in articulation of the inserted vowel (e.g., in bleed → [bɔlid]) and lexical schwa (e.g., in believe) should be greater than the differences found within a single category. In addition, because an inserted vowel that arises from CC mistiming does not have its own articulatory target, we would expect this vowel to show more coarticulation with the stressed vowel than a lexical schwa which does have an articulatory target.

Under the CCV mistiming account, the inserted vowel arises from the pulling apart of the two consonantal gestures and the intrusion of the stressed vowel (e.g., the [i] in believe) between the consonants (depicted in (3) above; Hall, 2003). The CCV mistiming account makes additional specific predictions regarding acoustic analysis. If the inserted vowel in VBR’s consonant cluster productions is the result of CCV mistiming, we expect the inserted vowel to be more similar (in F1/C1, F2) to the stressed vowel than lexical schwa is to the stressed vowel in comparable phonological environments. The CCV mistiming repair as discussed by Hall (2003) also entails that the inserted vowel should be more variable in duration than lexical schwa.

A final possible account of VBR’s inserted vowel is that it occurs as a result of noise at the level of articulation, and it is not a repair instituted either by inserting a vowel unit or a particular mistiming of gestures.

---

3 It is difficult to state the precise predictions of the CC-mistiming hypothesis with respect to comparing the inserted vowel contours to the other tongue contours (C₁, and C₂). This difficulty comes from the fact that the snapshot of the inserted vowel tongue contour could correspond to many different points in the transition from C₁ to C₂. Therefore, it is unclear whether the mistiming hypothesis predicts that the inserted vowel tongue contour should be closer to one of these consonants, or to some other tongue configuration.
According to this account, the difference between cluster words (e.g., *bleed*) and lexical schwa words (e.g., *believe*) is maintained at all levels of phonological processing. Thus, if articulatory noise is applied to the target articulation for each of these forms, we still expect differences to emerge in the acoustics. In other words, there is no principled reason why articulatory noise would alter the production of consonant clusters as to make them indistinguishable from structures with lexical schwa; instead, we would expect differences between these two forms to be maintained, with the articulation of each target being somewhat impaired relative to a normal speaker. In particular, we expect differences in the durations of the lexical and inserted schwa reflecting the fact that these come from two different articulatory plans which are produced incorrectly. In addition, it is worth noting that in each of the other acoustic and articulatory analyses, an account of inserted vowels arising as a result of articulatory noise predicts further differences between VBR’s production of lexical schwa and her production of vowels inserted in consonant clusters.

**Participants**

The experimental participant in this study is VBR, an aphasic English speaker who inserts a vowel in legal English obstruent-sonorant consonant clusters. A control subject, GJS (24 M), was recorded to verify that normal speakers show detectable acoustic and articulatory differences between production of words with lexical schwa (e.g., *believe*) and words with consonant clusters (e.g., *bleed*) on the measures used to examine VBR’s productions.

**Materials**

The target stimuli in the study consisted of 22 words with non-coronal obstruent-/l/ consonant clusters /C₁C₂/ in word onset, and 22 control words beginning with /C₁C₂C₂/.\(^4\) The control words were matched to the target words for the vowel following the C₂ as well as for stress. Each experimental word had primary stress on the cluster-initial syllable, whereas each control

\(^4\) The investigation focused on clusters with /l/ as C₂ due to practical considerations. As we will see, it is necessary for the analysis in this section that the tongue movements associated with the C₂ be discernable from the acoustic and articulatory record. This ruled out the use of clusters with /w/, as there is no single tongue shape associated with the production of /w/. The ultrasound experiment was originally carried out using clusters with /r/ as C₂ as well; these were not included in the analysis due to a large number of /r/’s produced as /w/, making it impossible to locate the beginning of the articulation of the /r/.
word had primary stress on the syllable beginning with /l/. Thus, primary stress fell on the vowel following /l/ for each word (e.g., clône ~ colôgne).

**Ultrasound setup**

Mid-sagittal images of the tongue were collected during speech using a commercially available ultrasound machine (Acoustic Imaging, Inc., Phoenix, AZ, Model AI5200S). Images were collected during the production of the /C1C2/, and /C1əC2/-initial words. A 2.0–4.0 MHz multi-frequency convex-curved linear array transducer that produces wedge-shaped scans with a 90° angle was used. Focal depth was set at 10 cm, producing 30 scans per second.

To ensure that the speaker’s tongue did not change position during data collection, the speaker’s head was stabilised by a specially designed head and transducer support (HATS) system (Stone & Davis, 1995). This was necessary to ensure head stability during running speech, and if the transducer is not immobilised, it is likely to shift by rotation or translation, leading to off-plane images that cannot be compared across tokens. In the HATS system, the speakers’ head is immobilised by padded clamps positioned at the forehead, the base of the skull, and the temples that can be re-sized for different heads. The transducer is held by a motorised arm that can be positioned under the subject’s head and adjusted to optimise the image for a particular speaker. The transducer holder in the HATS system is designed to maintain the transducer in constant alignment with the head and allow for full motion of the jaw.

In ultrasound imaging, piezoelectric crystals in the transducer emit a beam of ultra high-frequency sound that is directed through the lingual soft-tissue. A curvilinear array of 96 crystals in the transducer fire sequentially, and the sound waves travel until they reach the tongue-air boundary on the superior surface of the tongue. The sound waves reflect off the boundary, returning to the same transducer crystals, and are then processed by the computer which reconstructs a 90° wedge-shaped image of the 2-mm thick mid-sagittal slice of the tongue. In the reconstructed image, the tongue slice appears as a bright white line on a grey background. This is shown in Figure 2. Flanking the image of the tongue slice on either side are two shadows; the left shadow is cast by the hyoid bone, and the right is cast by the jaw, since bone refracts the ultrasonic beam.

**Recording procedure**

The subjects were seated in the HATS system, which was adjusted to fit their heads comfortably. The transducer was coated with ultrasound gel and placed in the holder. The position of the transducer was adjusted until the tongue image was visible, and the jaw and hyoid bone were equidistant from the edges of the scan. The target stimuli were read to the subject by an
experimenter who speaks with a neutral American accent. VBR was instructed to repeat each word four times, and then wait for the experimenter to provide the next stimulus. At two points during the recording session, the subjects were asked to swallow a small amount of water (3 cc and 10 cc). The images from the swallows were used to extract renderings of the palate. The recording procedure lasted approximately 30 minutes.

The visual ultrasound image and the synchronised acoustic signal were captured for each token. In addition, the speaker’s head was videotaped throughout the duration of the recording, and a video mixer (Panasonic WJ-MX30) was used to insert both the image of the head and an oscilloscopic image of the acoustic signal. A video timer (FOR-A VTG-33, Natick, MA) was used to superimpose a digital clock in hundredths of a second on each frame. The composite video output, which includes the ultrasound image, the videotaped image of the speaker’s head, the image of the oscilloscope, and the time, was recorded along with the audio digitally on a computer using Final Cut Pro, and simultaneously recorded on a VCR. This can be seen in Figure 2. Each frame during the subject’s verbal productions was exported to jpeg format (using Final Cut Pro) to enable analysis.

DATA ANALYSIS AND RESULTS

This section describes the results of the ultrasound imaging experiment, including the acoustic analyses as well as the analysis of the tongue shapes associated with the articulations of inserted and lexical schwa. The first part of this section discusses the data collected from VBR, followed by a discussion of the control speaker’s data.

For VBR, individual tokens were used for analyses only if each of the target consonants were articulated accurately, although voicing errors were accepted as they are not expected to alter tongue shapes during articulation (Davidson, 2003). Four tokens of each of the 22 stimulus words were recorded. In total, 176 repetition tokens were collected (88 lexical schwa, 88 consonant cluster); 8 tokens (4 with lexical schwa, 3 with consonant cluster targets) were discarded for having one of the consonants produced incorrectly as this type of substitution error affects all neighbouring articulations. One additional token with a consonant cluster target was discarded as there was no schwa present in the acoustic record.

Acoustic analysis

Several crucial comparisons were made between the lexical and inserted vowel types addressing three questions. First, is there a difference in mean duration between the two vowels? An account of insertion errors that arise due to articulatory noise predicts a difference – with lexical vowels longer
than inserted vowels – whereas the epenthesis account does not. (It is not clear whether the mistiming accounts predict a difference in duration.) Second, are the inserted vowels more variable with respect to duration than the lexical vowels? This addresses one of Hall’s criteria regarding CCV mistiming stating that intrusive vowels are more variable in duration than lexical or epenthetic vowels. To address this issue, we compare the standard error of lexical and inserted vowel durations. Third, is there more co-articulation between stressed vowels and the inserted vowels when compared with the amount of coarticulation between stressed vowels and lexical schwa? This addresses another of Hall’s intrusive vowel criteria, which states that intrusive vowels are copies (or the beginning) of the vowel adjacent to the sonorant. This will be addressed by comparing the first two formants of the unstressed vowel (lexical schwa or inserted vowel) with those of the stressed vowel from each token. If the inserted vowel arises due to CCV mistiming, we would expect the distance in acoustic space between the inserted vowel and the stressed vowel to be shorter than the distance between lexical schwa and the stressed vowel. A similar prediction may follow from CC mistiming, as discussed above. The schwa epenthesis account predicts no difference, as these two vowels are from the same population.

**Duration and variability of duration**

The length of each vowel type was computed using the acoustic waveform and the spectrograph image generated by Praat (Boersma & Weenink, 2005). The onset of the vowel was measured from the beginning of vocalic periodic noise following the release of C\(_1\), and the offset was set at the time when there was a change in intensity (from the unstressed vowel to the /l/) and the formant values transition into the sonorant. There was no significant difference in vowel length between the lexical vowels, mean = 125.0 ms, SD = 43.7 ms, and the inserted vowels, mean = 123.8 ms, SD = 45.7 ms; \(t(166) = 0.181, p > .80\). This result suggests that the two vowel types are similar with respect to duration, a finding which is inconsistent with the predictions of the articulatory noise account, and consistent with the schwa epenthesis account. It is worth noting, however, that both inserted vowels and lexical schwa are relatively long. Further, it is clear from the large standard deviations in each group that VBR’s vowel duration was variable for both groups. To determine whether there is greater variability in the duration of the inserted vowel than of lexical schwa, Levene’s test of equality of variances was used. The results indicated that the inserted vowel durations and lexical schwa durations did not differ in their variance (\(F = 0.208, p = .649\)). Thus, it is possible that there is a level of noise in VBR’s articulation, but the fact that there
was no difference between these two vowels suggests that the noise is ‘applied’ to the same intended articulation.

Coarticulation with stressed vowel

The analysis in this section was designed to determine whether the inserted vowel produced in forms like bleed has a greater degree of coarticulation with the stressed vowel (e.g., comparing the [a] the [i] in bleed) than a matched lexical schwa does (comparing the [a] and [i] in believe). According to Hall’s analysis of CCV mistiming, the inserted vowel and the stressed vowel should be closer in articulation (and hence, closer in acoustics) than the lexical vowel and the stressed vowel. Formant values were computed at the midpoint of schwa and after the onset formant transitions of the stressed vowel.5

The results of the analysis clearly show a great deal of coarticulation between both types of reduced (i.e., unstressed) vowel (lexical and inserted) and the stressed cardinal vowels (i.e., /i/, /u/, and /A/). These vowels are plotted according to their first and second formants in Figures 3 and 4. In the plots, $F_2$ is on the x-axis in decreasing units and $F_1$ is on the y-axis increasing from top to bottom. For each plot, the cluster of circles in the upper left hand corner represents the formant plots of VBR’s production of stressed [i] (e.g., believe). The cluster in the upper right-hand corner represents the production of stressed [u] (e.g., clue), and the cluster in the centre of the bottom represents the plots of [A]. Although there is a large degree of variability in these productions, they correspond to the formant frequency range for female English speakers reported in Hillenbrand, Getty, Clark, and Wheeler (1995).

In each figure, the reduced vowels are depicted with transparent shapes matching the solid shape of the stressed vowels in the same word. For example, in Figure 3, the solid squares plot the productions of /i/ (as in bleed) according to $F_1$ and $F_2$, and the transparent squares plot $F_1$ and $F_2$ of the inserted vowel VBR produced in words with /i/ (as in bleed [b3l3d]). It is apparent from Figures 3 and 4 that the $F_1$ and $F_2$ of the reduced

---

5 The diagram of CCV mistiming in (3) suggests that the inserted vowel is the beginning of the stressed vowel. Thus, it may be appropriate to compare formant measurements of inserted and lexical schwa with the formants of the stressed vowel at the vowel onset. However, because the onset of the stressed vowel and the onset of the inserted/lexical schwa vowels have different CV transitions (from [l] and from [p, b, k, g, f] respectively), this is also not an appropriate comparison. It is likely that the appropriate comparison is with stimuli not in the data set, with the stressed vowel in the same phonetic context that the unstressed vowels are in for this study. Despite this limitation, the CCV mistiming account still should predict some difference in the amount of coarticulation with the stressed vowel for the inserted vowel compared to lexical schwa. We express gratitude to an anonymous reviewer for helping us clarify this issue.
Figure 3. Plot of VBR’s stressed cardinal vowels and corresponding *inserted* vowel. Stressed vowels are circled, with /i/ in the upper left, /u/ in the upper right, and /a/ in the lower middle portion of the diagram. Inserted vowels produced in the same utterance are represented in transparent versions of the same shape.

Figure 4. Plot of VBR’s stressed cardinal vowels and corresponding *lexical* schwa. Stressed vowels are circled, with /i/ in the upper left, /u/ in the upper right, and /a/ in the lower middle portion of the diagram. Lexical schwas produced in the same utterance are represented in transparent versions of the same shape.
vowels cluster towards the $F_1$ and $F_2$ of the stressed vowel in the same word. This reveals a large amount of coarticulation between each type of reduced vowel and the stressed cardinal vowels (with some reduced vowel tokens appearing to be in the $F_1$–$F_2$ range of the cardinal vowel). Although there is coarticulation for each type of reduced vowel, it is important to consider whether the inserted vowel is more coarticulated with the stressed vowel than is the lexical vowel. To address this issue, $F_1$ and $F_2$ for each vowel token were transformed to Bark-scaled acoustic space (a normalisation procedure performed to account for the finding that the difference between two values in low frequencies is perceptually more salient than the same difference in high frequencies). Once the data were scaled, the Euclidean distance between the stressed vowel and the reduced vowel was computed for each token in the analysis. This Euclidean distance is taken to be the measure of coarticulation, with lower distance values corresponding to a greater degree of coarticulation.

The mean Euclidean difference in Bark-scaled acoustic space between the stressed vowel and the lexical schwa was 2.20 ($SD = 0.65$), and the mean difference between the stressed vowel and the inserted vowel was 2.35 ($SD = 0.67$). A $t$-test revealed no statistical difference between these two sets of Euclidean distances, $t(97) = 1.12$, $ns$. Thus, the degree of coarticulation between the cardinal vowels and the two types of unstressed vowels was not statistically different, confirming the trends evident in Figures 3 and 4.

**Acoustic analyses: Summary**

The analyses provided in this section directly addressed the possibility that the inserted vowels in VBR’s consonant cluster productions are the result of the gestural mistiming based on the CCV mistiming notion of Hall (2003) depicted in (3), and the possibility that the errors arise from noise at the level of articulatory implementation. Two analyses suggested that the CCV mistiming does not provide the best characterisation for VBR’s inserted vowels. First, the variability in the duration of the vowels was not different for lexical and inserted vowels. Second, no difference was found in the degree of coarticulation between the two unstressed vowels and the stressed vowel. In addition, the two vowels were statistically indistinguishable on all acoustic measures, rendering the articulatory noise hypothesis unlikely to be the best account of the inserted vowel. All results were consistent with the predictions of the schwa insertion account. The ultrasound analysis that follows addresses the CC mistiming hypothesis – that the inserted vowels are the result of a ‘pulling apart’ of the consonantal gestures associated with the articulation of the consonants in an onset cluster.
Ultrasound imaging analysis and results

Data processing

A trace representing the palate was created from the images recorded during the swallow by finding the highest point of the tongue from the anterior portion of the hard palate to the posterior portion of the soft palate (following the protocol outlined in Epstein, Stone, Pouplier, & Parthasarathy, 2004), which is the visible area in the swallowing images. This image was superimposed on each of the frames during data analysis, to provide a guideline for assessing the degree of constriction. For each token, the ultrasound frames of interest were chosen by examining the acoustic record to determine the time and duration of each /C1VC2/ sequence (for both lexical and inserted vowels). Each of the four repetition tokens of each stimulus produced by VBR were measured as long as the two consonants were produced correctly. The starting and ending times and the duration of the sequences were ascertained using a combination of Praat and the ultrasound images; this procedure was dependent on the consonants being examined.

For velar C₁ (i.e., /k/ and /g/), the first frame of the /C₁VC₂/ sequence corresponding to the tongue contour representing the velar consonant – was chosen by finding the narrowest degree of velar constriction, and the final frame was chosen by finding the point in the acoustic recording at the release of the sonorant C₂. To locate the ultrasound frame at the release of the sonorant (and onset of the stressed vowel), the acoustic time values corresponding to the transition from /l/ to the stressed vowel were divided by .033 (as each frame is 33 ms apart) yielding an approximate frame number. The ultrasound images were then examined to determine which frame corresponded to the transition from /l/ to the stressed vowel. The frame chosen using the ultrasound images was consistently within one frame (33 ms) of the frame number generated using the acoustic recording. As reported above, VBR’s productions were variable in their duration, and the number of frames analysed with a velar C₁ (i.e., from the frame before the tightest velar constriction to the frame after the first transition into the stressed vowel) varied from 12–22 frames.

The ultrasound images were analysed using EdgeTrak, a semi-automatic system for the extraction and tracking of tongue contours from ultrasound images (Akgul, Kambhamettu, & Stone, 1999; Li, Kambhamettu, & Stone, 2005). The user initiates contour extraction by manually selecting a few points on the tongue image. EdgeTrak uses B-splines to connect the selected points and optimises the edge tracking by determining the steepest black-to-white gradient. The algorithm is then applied to all of the tongue contours in
a sequence, and user correction is also possible. A sample extracted contour is depicted in Figure 5.

Once the contours are tracked over the images in the sequence, specific frames representing C₁ contour, vowel contour, and C₂ contour are separately saved for comparison. These frames were selected based upon specific criteria. For tokens with a velar C₁, these frames include the point of narrowest velar constriction (C₁ contour), the frame before the initial elevation of the tongue tip and tongue body gestures involved in production of /l/ (schwa contour), and the frame before the tongue begins to move to articulate the stressed vowel following the /l/ (C₂ contour). Initial labial consonants do not have a specific target tongue shape; therefore, no C₁ contour was identified for labial-initial utterances. For the purpose of illustration, the frame corresponding to a schwa contour is shown in Figure 6, along with the following frame showing the transition to /l/.

For each individually selected contour, the acoustic record of the production was used to verify that the frame number selected corresponded to an appropriate point in the speech wave. The frames were chosen independently by two members of the research team, and any disputes (which were rare) were resolved by the main experimenter. The analysis proceeded by computing the root mean squared (RMS) deviation value (described below) of each contour representing the inserted vowel with each of the other contours representing: (a) the lexical schwa; (b) C₂ (/l/); and
(c) \( C_1 \) (for velar-initial words). For example, each of the four inserted vowel contours from the four repetitions of clone is compared with each of the four lexical schwa contours (from cologne, yielding 16 RMS values), as well as with each of the four /l/ and /kl/ contours of clone (yielding 16 RMS values per comparison). In addition, the lexical schwa contours for a word were compared with one another (yielding six comparisons), and the inserted vowel contours were compared with one another (yielding six comparisons). Finally, we also compared the lexical schwa contours to the \( C_1 \) and \( C_2 \) contours to determine whether these differences are comparable with those between the inserted vowel and \( C_1 \) and \( C_2 \) contours.

The logic of the comparisons is as follows: the schwa insertion account predicts that the inserted vowel and lexical schwa contours are more similar to one another than the inserted vowel contour is to any of the consonants, and that the differences between the inserted vowel and lexical schwa tongue contours will not be greater than the differences among different repetitions of lexical schwa or the differences among different repetitions of the inserted vowel. In contrast, the CC mistiming account would be supported if the differences among the lexical schwa tongue contours were smaller than the differences between lexical and inserted schwa. Additionally, the gestural mistiming hypothesis does not predict that the inserted vowel and lexical schwa are more similar than the inserted vowel and the consonant gestures (see footnote 3). However, if the tongue contour representing the inserted vowel is more similar to the frame representing one of the consonants, this would suggest that there is a mistiming of articulatory gestures such that

Figure 6. Visual depiction of criteria for selecting schwa frame. In this repetition of the word gloat, the left image is the frame selected as the schwa frame, and the right frame (which is the next frame in the series) shows the transition to /l/, identified as the noticeable elevation of the tongue tip and tongue body. For each schwa frame selected, the time-synchronised acoustic signal was used to verify that the time associated with the frame corresponds to production of schwa. [This figure is available in colour in the online version of the Journal.]
there is still a smooth transition from C_1 to C_2, but the timing leads to the presence of the acoustic schwa.

The RMS deviation between two curves – the dependent variable in the analyses to follow – is computed by translating the curves to a series of 100 discrete points along the x-axis and determining the closest distance between the two curves at each point. An important note here is that the curves may have different minima and maxima along the x-axis, but they need to be the same length for the RMS computation to proceed. Therefore, two possibilities exist for this analysis: the shorter curves may be extended or the longer curves may be truncated. Extending (or kriging) the curves amounts to an extrapolation of the curve, and has been shown to introduce a fair amount of error into the signal (Parsatharathy, Stone, & Prince, 2005); therefore, our analysis proceeded by truncating each curve in a word pair (e.g., each C_1, C_2, and schwa curve from clone and cologne) to the highest minima and the lowest maxima along the x-axis (see Figure 8 for a depiction of tongue contours). Although some of the variation in the minima and maxima comes from noise in the visual signal (and what part of the tongue contour can be accurately extracted from that signal), there is also some systematic variation worth noting. Typically, the tongue contours associated with the production of /l/ extend further (i.e., have higher maxima along the x-axis) given the elevation of the tongue tip towards the alveolar ridge. Therefore, by truncating the curves to the smallest maxima, this portion of the /l/ contour which provides a large part of the contrast between the /l/ and schwa is discarded. In turn, this will favour the similarity of the C_2 and schwa curves.

**Ultrasound results: VBR**

Root mean squared difference values represent the difference between two contours, such that contours that are more similar have lower RMS values. These were computed using CAVITE (Parsatharathy et al., 2005), a program designed for comparison and averaging of tongue contours. For the first part of the analysis, three sets of RMS difference values were computed. In each case, the contour associated with the production of inserted schwa was compared to the contours associated with: (a) lexical schwa; (b) C_1; and (c) C_2.

The data indicate that the tongue contours associated with the inserted schwa are more similar to the contours associated with the lexical schwa (mean RMS = 2.23, SD = 1.09) than to the contours associated with the production of the neighbouring consonants, C_2 (mean RMS = 3.12, SD = 1.18) or C_1 (mean RMS = 5.22, SD = 1.15). Planned comparisons yield significantly smaller RMS values between inserted vowels and lexical schwa than between: inserted vowels and C_1, t(467) = 27.45, p < .001; and between
inserted vowels and $C_2$, $t(679.9) = 9.78$, $p < .001$. The fraction in the *degrees of freedom* for this latter analysis comes from using the *t*-value without the assumption of equal variances, as Levene’s test for equality of variances yielded a significant difference ($F = 5.173$, $p < .05$). The difference arises from *greater* variance in the comparison of inserted vowels and /l/ than inserted vowels and lexical schwa. Note that this difference in variance may at first appear to support a hypothesis in which the errors result from noise at the level of articulatory implementation.

To address this issue, the RMS differences between inserted vowels and /l/ were compared to the RMS differences between lexical schwas and /l/ (mean = 3.08, $SD = 1.19$). Levene’s test for equality of variances revealed that there was no difference in the variability in this comparison ($F = 0.011$, *ns*). Additionally, a *t*-test revealed that the RMS differences between /l/ and inserted vowels were statistically indistinguishable from the RMS differences between /l/ and lexical schwa, $t(583) = 0.885$, *ns*. A similar comparison was performed comparing the RMS differences between $C_1$ and the inserted vowel, and the RMS differences between $C_1$ and the lexical vowel (mean = 5.26, $SD = 1.20$). These comparisons also revealed that the RMS differences between these contours were statistically indistinguishable, $t(179) = -0.340$, *ns*, and Levene’s test for equality of variances indicated no difference in the variance of these populations ($F = 0.599$, *ns*).

An additional analysis was performed to address the strong prediction of the gestural mistiming account, that the difference among the tongue contours of lexical schwa repetitions should be smaller than the difference between lexical schwa and inserted vowel tongue contours. If the differences between the two schwa types are larger than the differences within each schwa type, this would suggest that the two vowels do not come from the same population. However, if the differences between the two schwas is the same as the variability within each schwa type, this would suggest that the tongue contours associated with each schwa come from the same population, and that the variability is due to other factors. The data indicate that the differences between the lexical schwa and inserted vowel tongue contours (mean RMS = 2.23, $SD = 1.09$) are not greater than the differences among lexical schwa contours (mean RMS = 2.33, $SD = 1.10$) or among inserted vowel contours (mean RMS = 2.09, $SD = 0.99$), $F(2, 519) = 1.12$, *ns*.

This pattern of results indicates that the tongue contours of the inserted vowel and lexical schwa were as similar to one another as the different tokens of the inserted vowel were to one another and the different tokens of the lexical vowel were to one another. However, there was some variability in all three comparisons (inserted-inserted; lexical-lexical; and inserted-lexical). To rule out the possibility that the variability is systematic (e.g., the tongue contours for the two vowel types systematically differ such that one is produced higher, more fronted, etc.), the plots in Figure 7 present tongue
contours associated with lexical and inserted schwa. Given that each type of schwa is coarticulated with the stressed vowel, it is helpful to look at the two schwas in a set of contrast pairs. The figures present the inserted schwa as solid lines and the lexical schwa as dotted lines for the word pairs with labial stop C₁ and /i/ as the stressed vowel (left panel) and for velar stop C₁ pairs with /u/ as the stressed vowel (right panel). It is clear from these images that there are not systematic differences between the two schwa contours.

The data discussed above provide support for the hypothesis that VBR’s inserted and lexical unstressed vowels are of the same type. The contours associated with the inserted vowel are more similar to those associated with the lexical vowel than to any other contour. Further, the variability between the inserted and lexical vowel contours is the same as the variability within each vowel type. Finally, the differences that do exist are not systematic. These data are consistent with the predictions of the hypothesis that the inserted vowels are produced as the result of schwa epenthesis.

To ensure that the results hold for all gestural contexts, the production of tokens with velar C₁ and labial C₁ were analysed separately. The average RMS data are presented in Table 1. These results show that the patterns discussed above hold for sequences with C₁ having either velar or labial place of articulation.

**Ultrasound results: Control subject**

As discussed above, a control subject completed the same experiment to determine whether VBR’s inserted vowel may be an exaggerated version of a normal process. The purpose of this component of the investigation was to
ensure that there is a clear distinction between words with lexical vowels (e.g., *cologne*) and words with consonant clusters (e.g., *clone*) in unimpaired articulation. The data from the control subject clearly indicate that VBR’s productions of target consonant clusters are categorically different from those of an unimpaired speaker. Crucially, none of the comparisons provided in the acoustic and articulatory studies were possible with the control subject, as there was no vowel present in the acoustic record between the consonants in cluster words, and it was impossible to identify the vowel ultrasound frame for the normal speaker on any of the repetitions of consonant cluster words.

The ultrasound images in Figures 8 and 9 illustrate the categorical difference between cluster words (e.g., *clone*) and lexical schwa words (e.g., *cologne*) for the control subject. Figure 8 shows the sequence of frames in the word *cologne*, with the /k/ in the upper left-hand corner and the beginning of the transition to the /l/ in the lower right-hand corner. Following the procedure used to analyse VBR’s ultrasound data, the lexical schwa frame is the image in the lower left, prior to the transition to /l/. In contrast to the articulation of *cologne* in Figure 8, the images in Figure 9 illustrate that the control subject’s articulation of *clone* does not permit us to identify an inserted schwa frame. In these images, the frame immediately before the transition to the /l/ is the frame associated with the velar C1.

Taken together, Figures 8 and 9 reveal that it is not possible to perform the same analyses with the control subject’s data that we performed with VBR. We were able to verify that the difference among different tokens of schwa was smaller than the differences between schwa and C2 for the control subject. Using the same procedure as outlined for VBR above, we found that the RMS differences among lexical schwa tongue contours (mean RMS = 2.23, $SD = 0.97$) were smaller than the differences between lexical schwa and C2 contours, mean RMS = 3.09; $SD = 1.13$; $t(331) = 45.24$, $p < .001$. This result suggests that there is a target for lexical schwa in unimpaired speakers,

### TABLE 1

Root mean squared differences (in mm) for the ultrasound analysis of VBR’s productions. For each column, numbers with different superscripts are significantly different ($\alpha = .05$)

<table>
<thead>
<tr>
<th>RMS comparison</th>
<th>Labial $C_1$</th>
<th>Velar $C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical schwa-Inserted schwa</td>
<td>2.21&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.49&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Inserted schwa-Inserted schwa</td>
<td>2.33&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.19&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lexical schwa-Lexical schwa</td>
<td>2.56&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.47&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>/l/-Inserted schwa</td>
<td>3.21&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.48&lt;sup&gt;y&lt;/sup&gt;</td>
</tr>
<tr>
<td>$C_1$-Inserted schwa</td>
<td></td>
<td>5.21&lt;sup&gt;z&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Figure 8. Sequence of frames in control subject’s production of *cologne*. The frame in the upper left corner corresponds to the production of /k/, and the frame in the lower right portion of the figure shows the beginning of the transition to /l/. The third frame in the sequence (lower left) would be identified as the schwa frame, immediately before the transition to /l/. [This figure is available in colour in the online version of the Journal.]

Figure 9. Sequence of frames in control subject’s production of *clone*. The images in this figure demonstrate that the control subject does not have a ‘schwa’ frame in the production of the cluster-initial word; the schwa frame and the velar C₁ frame would be identified as the same frame. [This figure is available in colour in the online version of the Journal.]
such that there is a tongue contour for lexical schwa that is distinct from the following sonorant. An additional point to note is that the RMS differences among the lexical schwa contours for the normal speaker are of the same magnitude as were comparable differences for VBR.

From Figures 8 and 9 as well as the above result, we conclude that the data from the control subject confirm that normal speakers show a categorical difference in their production of cluster-initial words and words with a lexical schwa between the same consonants. From this finding, it can be inferred that VBR’s data represent a deviation from the normal articulation of cluster-initial words.

**GENERAL DISCUSSION**

The ultrasound and acoustic experiments were performed to determine which of three theories of vowel insertion provides the best account of vowel insertion in VBR’s consonant cluster productions: schwa epenthesis, CC mistiming, CCV mistiming, or articulatory noise. The data from the two instruments (ultrasound imaging and acoustic recordings) converged on the claim that the vowel insertion errors produced by VBR were the result of a categorical change \(-/\text{C1\_schwa}\) and they were not the result of mistiming the component gestures in the utterance.

Unlike Hall’s (2003) description of CCV mistiming (depicted in (3)), VBR’s inserted vowel is clearly not due to the stressed vowel ‘intruding’ between the consonantal articulations. Differences between VBR’s data and Hall’s theory are as follows. First, the acoustic results revealed that VBR’s inserted vowel and her lexical schwa are both strongly influenced by the stressed vowel in $C_1(\text{a})C_2V_-$ words, and crucially that the inserted vowel is *not* more coarticulated with the stressed vowel than is the lexical schwa. Second, VBR’s lexical schwa and inserted vowels do not differ in the variance of duration. Each of these results is inconsistent with a CCV mistiming account of VBR’s inserted vowel, and each is consistent with the schwa epenthesis account (1).

The ultrasound imaging study was designed to address whether VBR’s inserted vowel is best characterised by the CC mistiming account (2), or by the schwa epenthesis account (1), and the results are consistent with the predictions of the latter account. Specifically, the evidence presented above showed that the tongue contours associated with the inserted vowel were more similar to lexical schwa than to the contours associated with the flanking consonants, as predicted by the schwa epenthesis account. Moreover, the differences between VBR’s lexical schwa and her inserted vowel were statistically similar to the differences found by comparing her lexical schwas to each other, and her inserted vowels to each other. In particular,
these results support the claim that the tongue contours of the inserted vowel and lexical schwa come from the same population of tongue configurations (as predicted by schwa epenthesis), and is inconsistent with an account in which the inserted vowel is different from the lexical schwa. The consistency of these results suggests that VBR’s vowel insertion errors are the result of schwa epenthesis, and not CC mistiming.

In the case report, we discussed VBR’s mild impairment in her performance on tests involving complex articulatory movements. Thus, it remained possible that her vowel insertion errors were the result of this mild impairment, and that the vowels resulted from ‘articulatory noise’. If the errors arose at an articulatory level, then the inserted vowel would be highly variable and distinct from lexical schwa. Instead, all evidence points to identity between these two vowels, implying that the errors arise at a level of spoken production at which the phonological target is mapped to a discretely different output representation. In particular, the acoustic analysis revealed that the VBR’s inserted vowels and lexical schwa were statistically indistinguishable in duration. In addition, the ultrasound imaging analysis indicated that variation in the production of the inserted vowel was matched by variation in the production of lexical schwa in several respects. First, the difference between VBR’s inserted and lexical vowels was statistically indistinguishable from the difference among tokens of the inserted vowel and among tokens of the lexical vowel. Second, the difference between the tongue contours associated with the inserted vowel and the contours associated with the flanking consonants in words with consonant clusters (e.g., in clone → [kolon]) was statistically indistinguishable from the differences between the lexical schwa contours and the flanking consonants in words with lexical schwa (e.g., in cologne). Third, a comparison of the contours associated with the vowels revealed that there was no latent pattern to the variation. These converging lines of evidence suggest that VBR’s inserted vowel error was not the result of articulatory noise.

It is worth noting that this does not preclude some additional articulatory disturbance, and the results noted both acoustic and articulatory variability in VBR’s productions of both types of words. In fact, there remains some chance that VBR’s deficit at this later level may be indirectly responsible for the schwa epenthesis repair, as the repair could be a type of compensation for a peripheral deficit, permitting more time for articulatory planning and implementation processes. This possible account may be supported by other facets of VBR’s performance (e.g., elongation of [s] in /s/-initial clusters), though it is not necessarily consistent with all facets of her performance (e.g., deletion of [j] in forms like cute would actually decrease the amount of time available for articulatory planning and implementation). However, even in the case that this is the root cause of VBR’s repair, the data presented above
suggest that the repair is instituted at a level that permits insertion of discrete sound structure units.\(^6\)

**Implications for theories of sound structure representation in spoken production**

The conclusion that VBR’s inserted vowel is the result of schwa epenthesis has implications regarding the types of possible operations in spoken production processing. In particular, accounting for this result requires that the representational system active at the level of her deficit allows the insertion of a discrete phonological unit. This result is consistent with several different proposed systems of sound structure representation, but places important restrictions on them. The following discussion considers three representational systems that have been proposed in the psycholinguistic and linguistic literature: ‘symbolic’ representations (e.g., Chomsky and Halle, 1968); gestural representations (e.g., Browman and Goldstein, 1986, 1988, et seq.), and exemplar-based representations (e.g., Pierrehumbert, 2001).

Many theories of spoken language production include symbolic representations of segments, features, and syllables at some stage in the cognitive processes involved in producing speech (e.g., Butterworth, 1992; Dell, 1986, 1988; Garrett, 1980; Goldrick & Rapp, in press; Levet et al., 1999; Roelofs, 1997; Shattuck-Hufnagel, 1987; Stemberger, 1985). These representations encode sound structure as a sequence of discrete units, and do not represent information regarding the temporal dynamics of articulatory movements. Theories that make use of this type of representation can provide a straightforward account of the data presented here. In terms of this representational system, VBR inserts a schwa segment into the form that is being produced, which separates the consonants in the cluster. It is important to note that while these theories lend themselves to an account of discrete speech errors such as the errors reported in this paper, gradient errors – such as consonant gesture mistiming (e.g., Davidson, 2003) – require a somewhat more elaborated representation than is provided by these theories.

The framework of Articulatory Phonology (Browman & Goldstein, 1986, 1988, 1989, 1992a) holds that sound structure representations take the form of dynamic motor units called articulatory gestures. In speech production, words are represented as gestural scores which describe the target gestures to be produced as well as information about the timing and coordination of these gestures. Previous work in this framework examined speech errors from both brain-damaged and neurologically intact populations, and has found

---

\(^6\) We gratefully acknowledge an anonymous reviewer for suggesting this possibility.
errors that occur at the level of gestural timing (Davidson, 2003; Hall, 2003) and gestural overlap (Pouplier, 2003; Wood & Hardcastle, 2000), including the simultaneous production of multiple articulatory gestures. It has thus been argued that the articulatory gesture is the unit of planning and action in spoken language production (Pouplier, 2003).

Saltzman and Munhall (1989) proposed that articulatory gestures may be coordinated with one another into larger units – gestural constellations – roughly corresponding to segments. For a theory of spoken language production grounded in the representations of Articulatory Phonology to account for the insertion pattern described above, it is essential that the set of possible modifications to the gestural score includes the insertion of gestural constellations, as in the case of schwa epenthesis. This possibility is consistent with the view of gestural representations presented in Gafos (2002), in which the phonological grammar may alter the gestural score; however, it is crucial that the set of possible ‘repairs’ during speech production include insertions at the level of gestural constellations and not simply changes to the temporal coordination of the articulatory gestures (for more discussion, see Buchwald, 2005b).

A third proposal of sound structure representation is that speakers store exemplars of the words they have heard and produced along a number of phonetic parameters (Pierrehumbert, 2001; also see Johnson, 2005). In short, exemplar-based representations consist of a map from a ‘category label’ (which may correspond, roughly, to segments, features, etc.) to a set of exemplars in phonetic parameter space (either acoustic space or articulatory space). The strength (or ‘activation’) of the exemplars with respect to the overall representation of the category label is a function of both the frequency and the recency with which the exemplars have been encountered. To the best of our knowledge this proposal has not yet been integrated into a comprehensive theory of the processes involved in spoken language production. Nonetheless, the data presented in this paper place an important restriction on its instantiation. In particular, the process of selecting an exemplar for production must permit the mis-selection of exemplars with different ‘category labels’ (e.g., in this case, selecting ‘[b\l]’ for ‘[b\l]’).

The most important constraint that our work places on theories of speech production is that the set of cognitive operations involved in speech production must include a process in which representations may be repaired or altered – via the insertion of a segment-sized unit – to yield a new sound structure sequence. This type of cognitive process is broadly consistent with work in linguistic theory (e.g., Chomsky & Halle, 1968; Prince & Smolensky, 1993/2004) in which forms are processed by a phonological grammar which can ‘repair’ ill-formed representations that violate well-formedness constraints. In the context of the work presented here, VBR’s performance can be seen as a window into the nature of this spoken language production
system (Buchwald, 2005a, 2005b). Previous articulatory research has underscored the need for representations that include information regarding the temporal dynamics of articulation, and this work suggests that an account of spoken language production additionally operates over discrete, manipulable sound structure representations.

CONCLUSION

The acoustic and articulatory data reported in this paper support the hypothesis that VBR’s vowel insertion errors in word-initial consonant clusters are the result of vowel epenthesis, a discrete ‘repair’ of complex sound structure sequences. The results were inconsistent with two accounts of the vowel insertion repair based on mistiming of articulatory gestures associated with the production of the target words, and were also inconsistent with an account of the error as arising due to articulatory noise. The results suggest that the processes involved in spoken language production involve operations over representational systems that allow the insertion of discrete sound structure units.

REFERENCES


