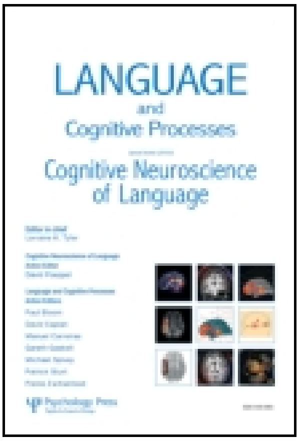

Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp20

A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation

Mark A. Pitt [a] & Christine M. Szostak [a]
[a] Department of Psychology, Ohio State University, Columbus, OH, USA

Version of record first published: 21 Feb 2012

LANGUAGE AND COGNITIVE PROCESSES
2012, 27 (7/8), 1225–1239

Psychology Press
Taylor & Francis Group

A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation

Mark A. Pitt and Christine M. Szostak

Department of Psychology, Ohio State University, Columbus, OH, USA

Words are pronounced in multiple ways in casual speech, which from the perspective of information transmission can be viewed as distortions that the listener must overcome to recognise the word intended by the talker. Two experiments explored the proposal that the recognition of pronunciation variants is facilitated by a lexically biased attentional set, which listeners adopt to compensate for fluctuations in signal reliability. Lexical decision responses were collected to multi-syllabic words in which a phoneme in one of four positions was gradually altered to make it a nonword. In Experiment 1, attention was manipulated through instruction. In Experiment 2, a lexically biased attentional set was induced by altering the design of Experiment 1. Results suggest that attention modulates lexical acceptability, damping lexical influences when attention is focused on perceiving the speech signal veridically (i.e., as pronounced), and maximising lexical biases when attention is focused on ensuring successful message transfer (i.e., perceiving the intended word).

Keywords: Attentional set; Pronunciation variant; Spoken word recognition.

Words spoken in casual speech can undergo considerable pronunciation variation. Their temporal and spectral properties can be altered to such an extent that a word is realised in multiple ways, some of which only partially resemble their citation pronunciation (e.g., *probably* can be pronounced /prali/ and /prai/). A puzzle in the quest to explain how spoken words are recognised is understanding why this wide variation in speech quality rarely causes comprehension to suffer.

Theoretical accounts of how variants are recognised appeal to processing and representational mechanisms that tend to be specific to language, tuned to the variation and the format in which words are represented in memory (Gaskell & Marslen-Wilson, 1998; Gow, 2003; Ranbom & Connine, 2007). For example, Gaskell and Marslen-Wilson (1998) suggest that variants are recognised by recovering the citation form of the word by engaging phonological processes to perform the mapping between how the word was said and how it is stored in memory. The purpose of this study is to offer another perspective on how listeners recognise pronunciation variants. It is not meant to compete with those above, but rather highlight some properties of variant recognition that any theory has to explain.

Correspondence should be addressed to Mark Pitt, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210, USA. E-mail: pitt.2@osu.edu

I thank Lisa Shoaf and Erik Tracy for assistance in many phases of this project. Interpretation of the results also benefitted from input from Dan Mirman.

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2011.619370

Pronunciation variation can be thought of as a perceptual adversity for the listener. Verbal communication requires successful transmission of information from the talker to the listener, with the outcome that the listener comprehends the talker's message. Pronunciation variation adds noise to this communication channel by adding variability to the signal. The scope of the problem is nontrivial because the forms, degrees, and frequency of variation are quite varied (Bell et al., 2003; Patterson & Connine, 2001; Pitt, Johnson, Hume, Kiesling, & Raymond, 2005). The data in Figure 1 provide a global snapshot that conveys its ubiquity in English. Shown are the proportion of occurrences in the Buckeye Corpus of conversational speech (Pitt et al., 2007) in which phonemes in each position in a word were spoken as different phonemes (substitutions) or not spoken at all (deletions). For all word lengths, the frequency of variation increases the further into the word the phoneme occurs, topping out at 35–40% by the fifth phoneme. Talker characteristics add to the variation, such as whether the person is a native speaker of the language.

When viewed as a type of distortion, it can be informative to compare pronunciation variation with other types of distortions that affects speech intelligibility, in particular energetic masking, which refers to the situation in which sounds (e.g., environmental noise) blend with speech acoustics to cause a reduction in signal clarity (Brungart, 2001). Pronunciation variation, however, is an independent source of distortion that precedes energetic masking. It originates from variability in speech production, where articulatory gestures are not completed or do not achieve their target endpoints. Variation can be conceptualised as one speech gesture partially obscuring, even masking, another (Browman & Goldstein, 1990).

As with energetic masking, listeners might recruit lexical memory to compensate for variation in signal quality. In fact, the presence of both types of distortion suggest that it would be to the listener's advantage for there to be strong and consistent lexical biases to ensure successful recovery of the words intended by the talker. We suggest that this comes about through a lexically biased attentional set, which the listener

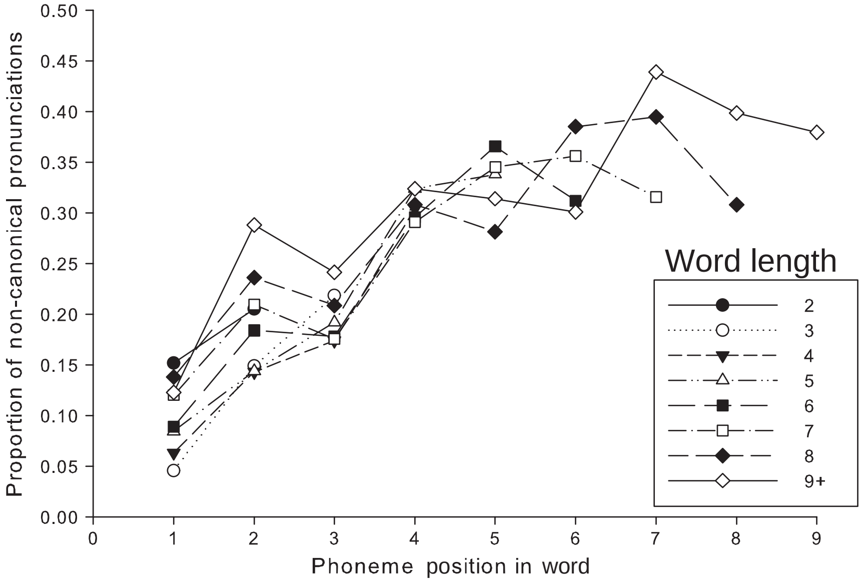

Figure 1. The frequency of noncanonical phone production in the Buckeye Corpus of speech as a function of word length and position of the phone in the word.

develops through experience with the language and the formation of informationally rich lexical representations. Memory is a stable reference from which to interpret distorted and ambiguous speech, and the adoption of a listening mode that takes full advantage of it is likely to ensure comprehension succeeds. Heavy reliance on memory is likely to be a successful listening strategy given the distribution of pronunciation variation found across a word (Figure 1). In addition, word beginnings provide a reliable lexical foothold (Gow & Gordon, 1995), which can then be leveraged to aid recognition of the remainder of the word.

This proposal can be viewed as an extension on the work of Mirman, McClelland, Holt, and Magnuson (2008), who augmented the TRACE model of spoken word recognition with an attention parameter that modulates the degree of lexical activation. As attention is focused more on speech input, global lexical excitation is damped, causing among other things a reduction in lexical influences on phoneme perception. Results from two phoneme detection experiments suggested this conceptualisation, and TRACE simulations with the attention parameter implemented in this way nicely reproduced the behavioral data pattern. Attention has been a central property of adaptive resonance theory as well (Grossberg, 1999), where it has a similar but broader functional role, although it is implemented differently.

The purpose of the current study was to demonstrate that listeners adopt a lexically biased attentional set when processing spoken language, and to test one consequence of doing so, that lexical biases will increase as more of the word is heard. The fact that speech extends in time enables memory to have a prolonged and increasing influence on encoding, the result of which is that listeners should be increasingly insensitive to pronunciation variation the later variation occurs in a word. Note that this prediction matches the qualitative trend in Figure 1 of more variation later in the word, as though a lexically biased attentional set partially inoculates listeners from the variation. Said another way, lexically guided listening is viewed as an adaptation that is designed for recognising time-dependent auditory objects whose clarity varies across the word.

Empirical evidence that supports the prediction of lexical influences increasing across a word can be found in past studies. Using phoneme restoration, Samuel (1987) found greater restoration (larger lexical influences) for phonemes at the ends of three-syllable words than at their beginnings and middles. Cole, Jakimik and Cooper (1978) and Marslen-Wilson and Welsh (1978) used mispronunciation detection, with one phoneme deviating by one feature in various positions in a word. The frequency with which mispronunciations were detected was inversely related to their distance from word onset, with many more mispronunciations detected word initially than word finally. For example, Cole and Jakimik found that mispronunciations were detected 72% of the time when the initial phoneme of a monosyllable was changed, but only 33% of the time when the final phoneme in the monosyllable was changed. When phonemes were changed at the end of trisyllables, Marslen-Wilson and Welsh found detection rates dropped to 24%.

What prior studies have not examined is how these lexical influences across a word are altered by the listener's attentional set. One might expect such influences to be significantly muted when attention is focused solely on the signal, but phoneme restoration requires just this sort of selective focusing and clear lexical influences were found. Only when listeners are presented with a visual prime of the word they are about to hear can lexical biases be significantly inhibited (Samuel & Ressler, 1986). It is also unclear how a change in attention (lexically biased versus focused on the signal) alters speech processing. Does it function only as a simple gain-control on lexical biases, or does processing change in other qualitative ways as listeners attend more or less closely to the talker?

The preceding questions were addressed in two lexical decision experiments in which different attentional sets were induced in participants and the position of a phonetic deviation in a word was manipulated. The lexical decision task was chosen because listeners' sensitivity to variation in word pronunciation is of interest, and like mispronunciation detection, it provides a measure of sensitivity. In Experiment 1, attention was manipulated through instruction. In Experiment 2, a lexically biased attentional set was induced exogenously by altering the design of Experiment 1.

EXPERIMENT 1

The purpose of Experiment 1 was to measure listeners' sensitivity to pronunciation variation across two levels of attention to the spoken words. One group of listeners (focused condition) was informed that the clarity of a phoneme would vary, and that they should pay attention to it to ensure accurate responding. This information was omitted from the instructions to the second group (diffuse condition), with the idea that listeners in the diffuse condition would adopt an attentional set that was closer to what is used in conversation.

Pronunciation variation was introduced by using stimuli typically created in the Ganong (1980) paradigm, but instead of judging the identity of a phoneme in a word, they judged the lexical acceptability of the word itself. /s/ gradually changed to /ʃ/ (and vice versa), turning a word into a nonword. The question of interest was how the proportion of *word* responses changed across the fricative continuum as a function of instruction and fricative position. Listeners in the focused condition were expected to be less tolerant of variation, but how this would manifest itself in responding, and whether the pattern would be similar across instruction conditions, is less clear. For example, when engaged in a phonetic classification task, lexical status biases responding when the phoneme is ambiguous, but not when it is a clear, unambiguous endpoint. A similar response pattern might be expected when making lexical decision responses in the focused condition because listeners' attention is also focused on the segment of interest.

In the diffuse condition, the only difference might be that lexical biases in the ambiguous (middle) part of the phonetic continuum are stronger. This is what would be expected of a lexically biased attentional set; less attention to the signal leads to larger lexical influences. If such biases are quite strong, one might find that they extend beyond the ambiguous region of the continuum to the nonword endpoint, causing listeners to classify an item with a clear /ʃ/ (e.g., *impreshive*) as a word. Both findings would suggest that a lexically biased attentional set is one means the recognition system uses to overcome the perceptual adversity posed by pronunciation variation.

Method

Participants

Eighteen students from introductory psychology courses served in each of the eight conditions. None reported speech or hearing difficulties.

Stimuli

Three /s/-/ʃ/ continua (initial, medial, and final) were created by first recording clear endpoints of the fricatives in an /ə/ context. After splicing off the vowels, the fricatives in each position were equated for duration and loudness, and then a continuum was constructed by digitally blending the two tokens in proportions ranging from 100% /s/ and 0% /ʃ/ to the reverse, in 5% steps, yielding a total of 21 steps. Fricative durations were 215, 134, and 248 ms for the initial, medial, and final continua, respectively. Results of a pilot experiment showed that endpoints were clearly identifiable.

Attempts to create a single, realistic-sounding fricative continuum that could be used in all word positions were unsuccessful because of the wide variation in duration and amplitude envelope of the fricatives across positions. Although this decision partially complicates comparison of results across some phoneme positions, differences are sufficiently sizeable in many instances to mitigate this concern. Furthermore, this issue does not pertain to the early-medial and late-medial positions, where the fricative continuum was held constant.

The endpoints and nine middle steps on each continuum were then spliced into a position-matched trisyllabic pair of words (e.g., word-initial: *_erenade, _andelier*) for use in a pilot experiment, the purpose of which was to identify three steps from the middle of the continuum that were sufficiently ambiguous to generate strong lexical biases. Three steps that yielded large lexical biases in labeling, and that together covered a sizeable portion of the response range (e.g., 0.15–0.85), were combined with the two endpoint steps to yield a five-step /s/-/ʃ/ continuum. This calibration procedure was repeated for the initial, medial, and final fricative continua.

The steps on each five-step continuum were then spliced into copies of each of the position-matched word-pairs (initial: *serenade-chandelier*; early-medial: *recipe-national*; late-medial: *impressive-condition*; final: *malpractice-establish*). Stimulus duration averaged 903 ms.

There was one stimulus list for each of the four phoneme position conditions. Each contained the 10 target stimuli (2 contexts × 5 steps) plus 20 fillers (10 words and 10 nonwords). Like the targets, the filler nonwords differed from the filler words by only one or two phonetically close phonemes in various positions. Fillers were included to make the targets less predictable by adding stimulus variety to the experiment.

Procedure

Groups of up to four listeners were tested simultaneously in separate sound-insulated rooms. They were instructed to press one key on a button board if the stimulus was a word or an adjacent key if it was a nonword. Fast and accurate responding was stressed. Participants given the focused instructions were informed that the "s" or "sh" letter sound in a particular word position could be ambiguous, and that they should listen closely so as to make the correct response. This instruction was not given to listeners in the diffuse condition.

The stimulus list was presented 10 times for a total of 300 trials. Presentation was blocked so that no stimulus repeated until all stimuli had been presented, and presentation was randomise within block. Participants had two seconds to respond after stimulus offset. There was a 1500 ms pause before the next trial. A rest break was provided halfway through the experiment. Sixteen practice trials preceded the test session.

Results and discussion

Responses to the /s/-biased and /ʃ/-biased stimuli were averaged in the analyses because the data patterns were similar. Mean proportion of *word* responses are shown in Figure 2 as a function of instruction condition (solid symbols and lines) and the continuum step for each of the four phoneme positions. The labels at each continuum endpoint represent superordinate categories that are necessary to use because the fricative at each endpoint is different depending on its lexical context. For example, in the initial position, at the word endpoint the fricative was 100% /s/ when the context was biased toward /s/ (e.g., *serenade*) and 100% /ʃ/ when the context was biased toward /ʃ/ (e.g., *chandelier*). At the nonword endpoint, just the opposite was the case (e.g., 100% /s/ was paired with *andelier*).

Looking first at the graph containing the data from the initial position, across continuum steps the two functions overlap at the word endpoint and then diverge toward the nonword endpoint, with *word* responses in the diffuse condition being on average 0.18 more frequent over the last three steps. A two-way ANOVA with instruction and continuum step as factors yielded a reliable interaction, $F(4, 136) = 8.06$, $p <.01$, indicating that listeners in the diffuse condition were more tolerant of fricative variation as the stimulus became more nonword-like. This same data pattern is also visible in the three subsequent phoneme positions, although it failed to reach statistical significance in the final position [early-medial: $F(4, 136) = 5.29$, $p <.01$; late-medial: $F(4, 136) = 16.21$, $p <.01$; final: $F(4, 136) = 1.93$, $p <.11$]. The main effect of instruction was reliable across all phoneme positions, showing that *word* responses were more frequent in the diffuse than focused conditions [initial: $F(1, 34) = 10.09$, $p <.01$; early-medial: $F(1, 34) = 10.58$, $p <.01$; late-medial: $F(1, 36) = 37.59$, $p <.01$; final: $F(1, 34) = 7.43$, $p <.01$].

Comparison of the labeling functions across the first three phoneme positions shows that the focused and diffuse functions spread further apart as the phoneme moved further into the word, suggesting that the effect of instruction was amplified later in the word. Keep in mind that only in the two medial positions could the fricative continuum be held constant, so some differences across position could be due to variation in the fricative continuum. The difference in mean *word* responses (across

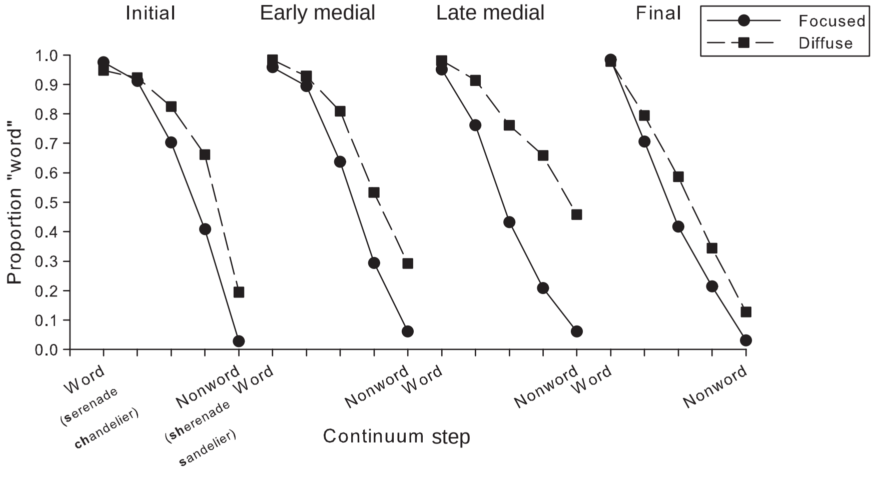

Figure 2. Proportion of *word* responses as a function of continuum step in the four phoneme positions across (Experiment 1).

all five steps) between the focused and diffuse functions was 0.10 in the initial position, 0.14 in the early medial position, and 0.27 in the late medial position. Statistical comparisons between adjacent positions yielded a reliable effect between the two medial positions only, $F(1, 34) = 4.86$, $p <.03$.

The cause of the amplified effect of instruction is two-fold and can be understood most easily by considering the data in the two instruction conditions separately. Listeners in the diffuse condition were increasingly tolerant of fricative variation as the fricative moved further into the word. This is most visible in the tails of the labeling functions at the nonword endpoint, which lift progressively higher from a value of 0.19 in the initial position to 0.46 in the late-medial position. Statistical comparisons of the labeling functions across adjacent positions yielded a reliable position by continuum step interaction for the initial versus early-medial analysis, $F(4, 136) = 2.88$, $p <.03$, and for the early-medial versus late-medial analysis, $F(4, 136) = 3.64$, $p <.01$.

Focused instructions affected responding in a qualitatively different way across the first three positions. Responses at the continuum endpoints changed minimally across positions. At the word endpoint, response proportions never dipped below 0.95, and at the nonword endpoint they never increased above 0.06. What changed were the middle three steps of the labeling functions, which were pushed downward (fewer *word* responses) across phoneme positions, indicating that listeners were increasingly less tolerant of fricative variation. This effect is most visible in the middle step, where *word* responses averaged 0.70 in the initial position and 0.43 in the late-medial position. Statistical analyses were performed over the middle three steps only, and they showed that the mean drop in *word* responses between the initial and early-medial positions was reliable, $F(2, 68) = 3.07$, $p <.05$, as was that between the early-medial and late-medial positions, $F(2, 68) = 4.37$, $p <.02$.

The trend of progressively fewer word responses in the middle of the continuum in the focused instruction condition extends to the final fricative position [late-medial vs. final: $F(2, 68) = 4.76$, $p <.01$]. On the other hand, at this same position listeners in the diffuse condition broke the pattern of a greater lexical bias (i.e., more *word* responses) further into the word and generated a classification function that is much steeper than the one in the late-medial position. Comparison of this function with its counterpart in the late-medial position yielded a reliable interaction, $F(4, 136) = 8.43$, $p <.01$.

One explanation for diffuse-condition listeners' heightened sensitivity to fricative variation in the final position is that they adopted a focused-attending listening strategy during the course of the experiment, possibly because the word-final fricative was especially noticeable coming at the end of the word. This conclusion is supported by a comparison of classification in the first and last halves of the experiment. In the first half, *word* responses at the nonword endpoint averaged 0.25. In contrast, in the last half, this value dropped to 0.06, indicating that over the course of the experiment diffuse participants listened more closely to the word-final phoneme. No such drop across the two halves of the experiment was found in the three other phoneme position conditions. These comparisons will be revisited in the "General discussion".

The results of Experiment 1 reveal the powerful influence that attention, as manipulated through instruction, can have on word acceptability judgments. In the diffuse condition, listeners were impressively tolerant of phonetic variation, to the point of readily categorising as words stimuli that had an unambiguous fricative and should have been classified as nonwords, which the listeners in the focused condition readily did. Perhaps a lexically based listening mode induces an expansion of the boundaries of internal phonetic categories, such as /s/, into perceptually similar categories (e.g., /ʃ/), thereby increasing tolerance to phonetic variation. If this tolerance increases as lexical activation of the word increases, it would explain why the diffuse functions rise as the fricative moved further into the word.

In contrast, listeners in the focused condition produced labeling functions that resemble what is found when using the Ganong paradigm. Endpoint responses remained fairly stable while performance varied most in the middle steps of the continuum. Because listeners were clued in to the variation in fricative clarity, as the fricative moved further into the word, there was an increasing reluctance to classify utterances with perceptually ambiguous steps as words. It was as though the listening instruction caused the opposite of what may have occurred in the diffuse condition: The boundaries of internal phonetic categories retracted, making listeners hypersensitive to the presence of the lexically inappropriate fricative (e.g., /s/ given *establi*).[1]

The current results lend credence to the idea that lexically biased listening is one means of combating variation in the clarity of the speech signal to ensure successful word recognition. It succeeds because talkers can usually be counted on to speak words the listener knows. As stated above, lexical memory is a stable reference against which to evaluate what can be a highly variable (i.e., noisy) signal.

Although these results demonstrate that listeners in the diffuse condition exhibited strong lexical biases when responding, properties of the experimental design might have caused the degree to which word processing is lexically biased to be underestimated. In particular, repetition of the stimuli (targets and fillers) could have mitigated diffuse attending by prompting participants to listen more closely to the stimuli than they otherwise would, because they undoubtedly noticed stimulus repetition. The purpose of Experiment 2 was to address this issue.

EXPERIMENT 2

Two changes were made to the design of Experiment 1, both of which were aimed at obtaining a more accurate measure of lexically biased attending. The proportion of filler stimuli was increased so that they swamped the targets in the stimulus list (86% vs. 14%, respectively). In addition, phonetic variation of the fricative occurred across 15 target words instead of just one, with never more than two steps from each of the 15 continua presented to the same listener, virtually eliminating target repetition. The purpose of these changes was to approximate a typical lexical decision experiment, with the reasoning that it would induce in listeners in the diffuse condition an attentional set that more closely matched that found when engaged in conversation. If this reasoning is correct, listeners should be at least as tolerant of fricative variation as those in the diffuse condition of Experiment 1, perhaps even more so if the manipulations induced an even greater reliance on lexical memory. Listeners in the focused condition were expected to be less affected by these changes because they were informed of the manipulation of fricative clarity.

[1] Reaction times were also analysed, but for the most part they provided little additional insight into participant performance, so they are not reported. For reference, in all but the late-medial position in Experiment 1, RTs in the focused condition were on average 110 ms faster than in the diffuse condition (1175 vs. 1285, measured from word onset), except in the late-medial position, where there was a large reversal. Mean word RT in Experiment 2 was 1053, differing little across the three positions. In both experiments, analyses of RTs to the filler items yielded a reliable speed up to words over nonwords.

Method

Participants

Four hundred and eighty new participants from the same population as Experiment 1 served as listeners, 120 in each of the three diffuse-instruction conditions and 60 in each focused-instruction condition.[2]

Stimuli

The fricative continua and lexical contexts from Experiment 1 were re-used. An additional 14 word pairs were added to each position condition, for a total of 15. For the word-final trisyllables, it was easy to identify enough words that met the necessary criteria (e.g., the utterance must form a word at one endpoint and a nonword at the other). For the initial and early-medial positions, there were not enough trisyllables in English that satisfied all of the necessary criteria, so bisyllables had to be used. In the initial position, all but one pair was bisyllabic. In the medial position, seven were bisyllabic and eight were trisyllabic.[3]

The to-be-replaced fricative was spliced out of each word and each of the five steps on the position-appropriate fricative continuum (initial, early-medial, and final) was spliced into copies of the word to yield the target stimuli, 150 in each position condition (30 word contexts × 5 fricative steps).

For each of the three phoneme positions, stimuli were distributed across three stimulus lists in a manner that minimised repetition of items from the same continuum. In each position condition (initial, early-medial, and final), each of the five fricative steps occurred 10 times in each list, for a total of 50 target stimuli. What differed across lists was the word in which each step was embedded. Because there were only 30 word contexts, 20 contexts occurred twice in a list and the other 10 occurred once. When a context word occurred only once, the fricative was always the middle (3) step on the phonetic continuum. When it occurred twice, fricative steps 1 and 4 or 2 and 5 were used. This choice of pairing was motivated by the idea that more widely spaced steps provided the best opportunity for reducing the possibility of listeners hearing a stimulus repeat. Across the three lists, each word (i.e., lexical context) occurred once with the middle fricative, once with steps 1 and 4, and once with steps 2 and 5. Stimulus lists were constructed so that word contexts which occurred twice were placed in different halves of the list. Each list was counterbalanced for presentation order. Across all 30 continua, stimuli averaged 787 ms ($SD = 105$) in duration.

Fillers were 310 words and nonwords. Nonwords were constructed by altering one or more phonemes in a word. Strings with /s/ and /ʃ/ were minimised, although not eliminated, to ensure they were not over-represented in the lists or in a particular phoneme position. Fillers varied in length from one to three syllables.

[2] Participant availability was the reason for the different *N* in each condition. Data in the two instruction conditions were collected at very different times.

[3] Although the use of two-syllable words could reduce lexical biases, data analyses that examined length differences showed that any such effect of word length was negligible. Classification functions for the two word lengths overlapped each other quite closely over the five-step continuum. A two-way ANOVA with continuum step and word length as factors yielded no reliable effect of length. This result gives little reason to think that the use of two-syllable words reduced lexical influences, at least in the early-medial position.

Procedure

The procedure was identical to that of Experiment 1. There was one rest break after half of the 360 trials, and an 18-trial practice session.

Results and discussion

Because participants responded to only 50 of the 150 target stimuli, yielding five observations per step, each to a different stimulus, performance estimates from subject analyses would likely be less accurate than item analyses, so only the latter are reported to conserve space.[4] The proportion of *word* responses at each of the five steps was calculated for each of the 30 target words. These data were then averaged across words in each position condition and the resulting classification functions are plotted in Figure 3.

In the early medial and final positions, the changes to the experimental setup had the intended effect of making listeners more tolerant of fricative variation, responding *word* more often than in Experiment 1. What was unexpected is that listeners in both the focused and diffuse conditions displayed this increased tolerance, with the labeling functions shifted upward relative to the corresponding functions in Figure 2. By increasing the unpredictability of the target stimuli, listeners were biased more by lexical memory, regardless of whether they knew of the fricative manipulation.

Overall, the results replicate what was found in Experiment 1. The main effect of instruction was reliable in all phoneme positions [initial: $F(1, 29) = 209.41$, $p < .01$; early medial: $F(1, 29) = 73.86$, $p < .01$; final: $F(1, 29) = 73.78$, $p < .01$], with the diffuse classification functions pivoting upward at the word endpoint to rise above the focused functions. In the early medial and final positions, this effect is particularly large, and noticeably larger than in Experiment 1, with the difference in responses at

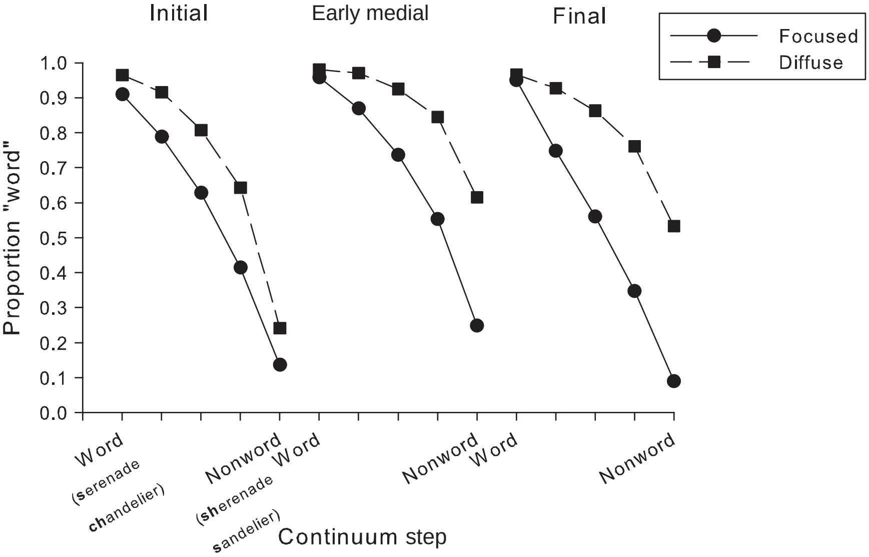

Figure 3. Proportion of *word* responses as a function of continuum step in the four phoneme positions across (Experiment 2).

[4] Subjects analyses yielded the same findings.

the nonword endpoint (diffuse–focused) averaging 0.37 and 0.44, respectively. The size of the latter effect differs from the much reduced effect of instruction found in Experiment 1, and reinforces the suspicion that in the final position listeners in the diffuse condition adopted a focused listening strategy. The vastly different designs across experiments made it difficult to perform meaningful statistical comparisons to evaluate these differences quantitatively.

As in Experiment 1, the effect of instruction increased across phoneme position (initial = 0.14, early medial = 0.19, Final = 0.27); differences between adjacent positions were reliable, [initial vs. early medial: $F(1, 58) = 31.70$, $p <.04$; early medial vs. final: $F(1, 58) = 3.88$, $p <.054$].

By changing the methodology of Experiment 1 so as to virtually eliminate repetition of the target words and to hide the phonetic variation of the fricative, listeners were much less sensitive to fricative variation, but only after word onset. Very strong lexical biases emerged in the diffuse conditions by the onset of the second syllable, and remained strong through word offset. These data are further evidence of how the workings of attention and lexical memory are integrated to aid processing of a highly variable speech signal.

GENERAL DISCUSSION

Pronunciation variation adds noise to a verbal communication channel. Explanations of how listeners compensate tend to be confined to a specific type of variation. However, pronunciation variation is an ever-present property of casual speech, being widespread and varying greatly in severity, making a comprehensive explanation elusive. The purpose of the current study was to suggest that a general-purpose mechanism, attention, contributes by ensuring heavy reliance on a reliable information source, lexical memory.

The results of two experiments demonstrate the viability of this proposal. Sensitivity to phonetic variation was influenced by the attentional set adopted in the experiment. When not informed of the variation in the fricative, listeners were very accepting of deviations in pronunciation in the two medial positions, and not just when the fricative was ambiguous (step 3), but all of the way out to the nonword endpoint. When the experimental design was altered to make the manipulation of fricative clarity less obvious, listeners were even more tolerant of phonetic variation, with lexicality judgments at the nonword endpoint in the early-medial and final positions reaching an average of 53% in the diffuse conditions in Experiment 2.

The correspondence between the distribution of pronunciation variation across a word (Figure 1) and listeners' increasing *in*sensitivity to it (Figures 2 and 3) suggests that a lexically biased attentional set is tailored to pronunciation variation in speech. Listeners were most sensitive to phonetic variation where it occurs least often, in word-initial position. By the onset of the second syllable, listeners were highly tolerant of variation, and this tolerance remained relatively constant across the remainder of the word. Listeners recruit lexical memory to its fullest during spoken language comprehension because it is a beneficial listening strategy for coping with the variable clarity of casual speech. More broadly, a lexically biased attentional set is also attractive because of its explanatory scope, being potentially applicable across many forms and degrees of variation. That said, lexical processing alone is probably not sufficient, as recognising /prai/ as *probably* by appealing to lexical activation alone seems a stretch.

The current data are in line with those of Mirman et al. (2008) in suggesting that attention functions as a gain control that modulates the dynamics of word recognition, as exemplified in TRACE, for example. When attention is focused on the signal, lexical influences are reduced, as reflected in consistent classification of the endpoints and a heightened sensitivity to the quality of the fricative when it was ambiguous. When attention is focused away from the signal and more on comprehending the entire word, lexical biases are large, so large that they extend easily to the nonword endpoint in noninitial positions. No evidence was found that word processing changed in other qualitative ways as a function of attentional focus.

Auditory attention is conceptualised here as being relatively constant from one word to the next when listening to a talker, although it can fluctuate across a sentence, with listeners focusing on a novel word or a word that receives accentual prominence. Because attention modulates lexical influences, differences in attention will be found only when lexical influences can be seen. This means that word-initially, when lexical activation is relatively weak, attentional set will have only a minor effect on perception. As lexical activation of a word increases, attentional set can have a greater influence on word perception, which is where the largest effects of instruction were found.

Successful communication also requires that attention be flexible. What a talker will say can be unpredictable (e.g., new words, change of topic) and listening conditions can be poor (environmental noise). Successful communication in these situations may require close attention to the speech so that the listener hears exactly what the talker said. Perception in this case aims to be as veridical as possible, which is achieved by attention minimising lexical biases.

This dual role of attention comports with recent findings by Mattys and colleagues (Mattys, Brooks, & Cooke, 2009; Mattys, Carroll, Li, & Chan, 2010), who examined listeners' reliance on acoustic and lexical information when processing speech under acoustic load (e.g., noise masking) and cognitive load (e.g., secondary task). The results of these studies led to the advancement of the lexical drift hypothesis, whereby listeners rely increasingly on lexical memory as cognitive load increases.

Mattys and Wiget (2011) sought to identify the cause of lexical drift. Their findings are particularly relevant to the current study because they used a pair of Ganong stimuli (/gIft/-/kIft/ and /gIs/-/kIs/ continua) in a variety of behavioral tasks. In their Experiment 1, participants classified the word-initial phoneme as either /g/ or /k/ when a cognitive load (concurrent visual search task) was present or absent. Cognitive load modulated the size of lexical influences on responding. When participants had to divide their attention between the two tasks, biases to respond in a lexically consistent manner (e.g., responding /g/ given /Ift/) increased relative to when the secondary task was absent or reduced in difficulty. The cause of this lexical drift was shown to be due to inefficient or suboptimal stimulus encoding in a subsequent discrimination (same-different) experiment. Adjacent steps on the phonetic continuum were less discriminable when participants had to perform the visual search task than when they did not.

The lexically biased attentional set being advanced in the present study can be considered a form of lexical drift, with pronunciation variation being a signal-based reason for lexical drift. Unclear speech acoustics may be equivalent to, or have processing consequences similar to, poor speech encoding. To maximise accurate perception, the perceptual system compensates by exhibiting lexical drift, increasing reliance on well-formed memory representations.

A lexically biased attentional set is likely to be the signature of an experienced language user, someone who has informationally rich lexical representations and can take full advantage of them. Mattys et al. (2010) showed that, in contrast to the results described above for native listeners, non-native listeners of English do not exhibit a reliance on lexical memory when cognitive load is taxed. Among the reasons they discuss for this pattern of results is inferior lexical knowledge. If non-native speakers rely less on lexical memory, they should display greater sensitivity to pronunciation variation. In the current experimental setup, this should lead to steeper labeling functions in the diffuse conditions, particularly in the noninitial positions. Also, the increase in the strength of lexical biases across a word should be reduced.

In a similar vein, repeated exposure to the ambiguous fricatives might make one wonder whether perceptual learning (Norris, McQueen, & Cutler, 2003) could partially explain why listeners judged steps toward the nonword endpoint as lexically acceptable. If listeners' phonemic categories were broadened by exposing the perceptual system to ambiguous fricatives, one would expect larger lexical biases later in the experiment, after more exposure had occurred. At least for Experiment 1, there is little evidence that this happened. Labeling functions calculated on the data from the first half of the experiment were comparable to those calculated on the data from the second half.

Another reason to doubt that phonemic categories were altered in a meaningful way through perceptual learning has to do with the composition of the stimuli. During the training phase of a perceptual learning experiment, learning is assumed to come about from exposure to ambiguous phonemes. For example, the perceptual system infers that the talker produces /s/ in a very /ʃ/-like manner. Listeners in the current study always heard both endpoints of the fricative continuum, removing any uncertainty about the properties of the talker's /s/ and /ʃ/ categories. In addition, the /s/-biased and /ʃ/-biased contexts were not only intermixed during the experiment but also occurred equally often with all continuum steps, providing no opportunity for the properties of one fricative category (e.g., /s/) to be altered more than the other (e.g., /ʃ/), an outcome that would be necessary to obtain the labeling functions in Figures 2 and 3. One way perceptual learning could have occurred is if the mental representations of /s/ and /ʃ/ expanded independently of one another, a possibility that has not been explored in the perceptual learning literature.

The lexical decision task was used in this study because sensitivity to lexicality is of interest: Listeners are highly tolerant of pronunciation variation, and few tasks can measure this tolerance as unambiguously as responding *word* versus *nonword*. That said, its use can raise concerns about the source of the instructional effects. The different sets of instructions might not have altered just attention, but also induced more stringent criteria for responding *word* in the focused than diffuse conditions. If this occurred, decision biases might be responsible for the general pattern of the diffuse functions being above the focused functions, but a more elaborate account would seem to be required to explain other aspects of the data, in particular why responding changed differentially across positions given the two sets of instructions. Note that, particularly in Figure 2, while the nonword tail lifted progressively higher across positions in the diffuse condition, the middle of the functions bowed downward in the focused condition. It is unlikely that changing only one's response criterion would yield such a pattern.

It is unclear whether other methodologies could avoid similar criticisms when attention is manipulated explicitly through instruction. Regardless of how responses are measured, behavior will be altered by instruction. A change in experimental design is an alternative and minimally obtrusive means of examining attentional influences on word perception, which is why it was employed across experiments. An overarching goal of this study was to estimate listeners' sensitivity to pronunciation variation when engaged in a task that approximates a default mode when listening to speech. Tasks such as eye tracking might do a better job of this than lexical decision because the visual-world paradigm simulates listener engagement with the environment. Nevertheless, a vexing challenge faced by any methodology is manipulating variation across a broad range without listeners consciously noticing the variation, because once they do, listening in a default mode might be abandoned.

Finally, because the stimuli were produced in a laboratory and not tokens of actual variants taken from talkers speaking in a casual style, is generalisation of the findings compromised? That is, are the results ecologically valid? We do not really know enough about variation and spoken word processing to answer this question with confidence. Design considerations affect choice of words and the type of variation examined, and compromises are frequently necessary. For example, on the one hand, talkers vary in the clarity with which they produce /s/, and substituting /ʃ/ for /s/ is an ongoing change in American English in certain phonological contexts (e.g., before /tr/ as in *shtrike* instead of *strike*). Relatedly, few other phonemes are as productive as /s/, occurring in all four word positions across a large set of words. On the other hand, the representativeness of the three middle continuum steps was not assessed, nor was the frequency of /s/ variation in the target words. These oversights might have affected the results, but it is just as easy to speculate that it caused tolerance of /s/ variation to be underestimated as overestimated.

CONCLUSION

A lexically biased attentional set is a powerful means by which to ensure accurate word recognition, and one that is well adapted to the demands of verbal communication. A constant and strong lexical bias is a useful general-purpose means of ensuring successful communication given the unpredictability of environmental noise and speech clarity. At the same time, the flexibility of attention serves a complementary role by enabling listeners to minimise the influences of memory so as to hear what was said veridically when the situation demands it.

REFERENCES

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*, 1001–1024.

Browman, C., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 341–376). Cambridge, UK: Cambridge University Press.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.

Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, *64*, 44–56.

Ganong, W. F. (1980). Phonetic categorization in auditory perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Gaskell, G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 380–396.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*, 575–590.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 344–359.

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. *Consciousness and Cognition*, *8*, 1–44.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*, 203–243.

Mattys, S. L., Carroll, L. M., Li, C. K. W., & Chan, S. L. Y. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech Communication*, *52*, 887–899.

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*, 145–160.

Mirman, D., McClelland, J. L., Holt, L. L., & Magnuson, J. S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science*, *32*, 398–417.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Patterson, D., & Connine, C. M. (2001). Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production. *Phonetica*, *58*, 254–275.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of conversational speech*. Columbus, OH: Department of Psychology, Ohio State University (Distributor). Retrieved from www.buckeyecorpus.osu.edu

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye Corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*, 89–95.

Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, *57*, 273–298.

Samuel, A. (1987). Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language*, *26*, 36–56.

Samuel, A. G., & Ressler, H. R. (1986). Attention within auditory word perception: Insights from the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 70–79.