A phonetic explanation of pronunciation variant effects

Meghan Sumner

Citation: The Journal of the Acoustical Society of America **134**, EL26 (2013); doi: 10.1121/1.4807432
View online: https://doi.org/10.1121/1.4807432
View Table of Contents: https://asa.scitation.org/toc/jas/134/1
Published by the Acoustical Society of America

Meghan Sumner: JASA Express Letters [http://dx.doi.org/10.1121/1.4807432] Published Online 5 June 2013

A phonetic explanation of pronunciation variant effects

Meghan Sumner
Department of Linguistics, Stanford University, Margaret Jacks Hall, Building 460, Stanford, California 94305-2150
sumner@stanford.edu

Abstract: Effects of word-level phonetic variation on the recognition of words with different pronunciation variants (e.g., *center* produced with/(out) [t]) are investigated via the semantic- and pseudoword-priming paradigms. A bias favoring clearly articulated words with canonical variants ([nt]) is found. By reducing the bias, words with different variants show robust and equivalent lexical activation. The equivalence of different word forms highlights a snag for frequency-based theories of lexical access: How are words and word productions with vastly different frequencies recognized equally well by listeners? A process-based account is proposed, suggesting that careful speech induces bottom-up processing and casual speech induces top-down processing.

PACS numbers: 43.71.Es, 43.71.Sy [DDO]
Date Received: March 9, 2013 **Date Accepted:** May 7, 2013

1. Introduction

Numerous studies have found that carefully-articulated words produced with a canonical variant fare better in the word recognition process than similar productions of the same word with a non-canonical variant (Andruski *et al.*, 1994; Pitt, 2009; Tucker and Warner, 2011). For example, a word like *center* is recognized more quickly and accurately when produced with a canonical [nt] variant (e.g., *se**n-t**er*) than with its more frequent, reduced counterpart (e.g., [n_], *se**n-n**er*). We also know that listeners are sensitive to and use phonetic detail in the processes and representations that underlie spoken word recognition (Goldinger, 1998; McMurray *et al.*, 2002; Nygaard *et al.*, 1994). And, studies have shown that multiple pronunciations of a word with different variants (e.g., [nt] vs [n_]) fare equally well, with no observable difference between the recognition of words with canonical versus non-canonical variants in immediate processing tasks (Gow, 2001; Sumner and Samuel, 2005). The response benefit for canonical variants (*canonical bias*) is typically attributed to either the prominence of an idealized lexical representation that best matches an infrequent, but canonical word form *or* the dependence on *following phonological context* that licenses a particular variant [as in nasal assimilation (Gaskell and Marslen-Wilson, 1996) or consonant cluster reduction (Mitterer and McQueen, 2009)]. In contrast, work that has reported variant equivalence appeals to co-present phonetic and/or allophonic properties present in natural speech. Comparing a canonical production of *green* to a naturally assimilated version, with inherent residual phonetic cues, Gow (2001) found no cost for non-canonical variants—even absent the assimilatory context (Gow, 2003). Sumner and Samuel (2005) examined the recognition of t-final words (e.g., flute) with the final sound produced as a fully released stop, a glottalized unreleased stop, or a glottal stop and found that all variants result in equivalent and robust lexical activation (without a following segmental trigger).

Synthesizing these two seemingly contradictory patterns will help explain how listeners recognize spoken words despite massive variability in speech. Perhaps the types of variation examined by Gow and Sumner and Samuel involve graded within-

sound variation and those that result in a canonical bias involve different sounds with their own distributional properties or the presence/absence of a sound ([t] vs [ɾ], Ranbom and Connine, 2007; Tucker, 2011; [nt] vs [n_], Pitt, 2009; [st] vs [s_], Mitterer, and McQueen, 2009). However, residual cues are likely to be present even in cases of *phonological* deletion (nasal stop vs nasal tap). One possibility is that a particular variant is tightly linked with how a word is uttered, and the canonical bias reflects a cost associated with an incongruent phonetic context and non-canonical variant. For example, post-nasal [t] deletion, while rampant (see Pitt, 2009) is increasingly less likely as the prominence of the following syllable increases (de Jong, 1998), which is likely not independent of the articulation of the first syllable. Similarly, Andruski *et al.* (1994) found that *king* facilitates recognition to a semantically related word when fully aspirated, but not when produced with reduced aspiration. The reduced-aspirate version was created by splicing out the mid-portion of the fully aspirated form, pairing a reduced VOT with a canonical phonetic frame; resulting in an unviable pairing in American English (predicting a *voiced* percept). In studies of spoken words in isolation (where our stimuli are oftentimes naturally, but carefully articulated), a carefully articulated word with a non-canonical variant may have a disadvantage compared to a similarly articulated word with a canonical variant *a priori* (reminiscent of perceptual costs of mismatching formant and release place cues, Marslen-Wilson and Warren, 1994). I suggest that much of the discussed disparity can be attributed to the phonetic composition of the word that *houses* each variant (referred to hereafter as *word-level phonetic variation*). Once we dampen the phonetic cues that favor canonical variants, we find equivalence across variants that differ greatly in terms of frequency. This is an issue for theories that depend on measures of quantitative frequency that mediate lexical access (Goldinger, 1998; Johnson, 2006).

2. Experiment 1: Semantic priming across pronunciation variants and speech styles

This semantic priming paradigm is used to investigate lexical activation of medial-nt words (e.g., *splinter*) produced with either [nt] or [n_]. Effects of word-level phonetic variation are addressed by comparing the variants in careful and casual speech frames. In this paradigm, hearing a word (e.g., *splinter*) activates the corresponding lexical item, and this activation spreads to semantically related items. Related words are pre-activated, or *primed,* facilitating recognition of the target. This paradigm enables us to examine responses that are dependent on lexical access, and does not require speech manipulations that may disrupt the recognition process.

2.1 Methods

2.1.1 Participants

Sixty undergraduate students participated in this study. All were monolingual native speakers of American English. No reports of hearing-related issues were made.

2.1.2 Stimuli

Forty-eight critical primes with medial /nt/ were chosen from a free response target-associate mass-testing experiment. 144 filler primes and targets were chosen based on similarity to the target words in terms of length, syllable structure, and lexical frequency. Forty-eight of these fillers were real-real pairs, and 96 were real-pseudoword pairs. Three versions of each critical and filler prime were recorded by a female native speaker of American English using a head-mounted microphone set 2 in. from the speaker's lips. These conditions included: (1) careful productions with [nt]; (2) careful productions with [n_]; and (3) casual productions with [n_]). The first two conditions replicate past work; the third is underexplored in spoken word recognition studies [though degree of reduction independent of pronunciation variants in conversational speech has received a great deal of attention by Ernestus and colleagues (see Ernestus *et al.*, 2002)]. The words were recorded using randomized word lists presented to a single speaker of American English. One stimulus per word per condition was chosen for

the experiment based on word duration [guided by those reported by Pitt (2009) for careful productions and by McLennan *et al.* (2003) for casual productions], presence/absence of [t]-closure and released burst, nasal duration, and word duration and intensity. Figure 1 provides a representative example of the stimuli for the three experimental conditions.

2.1.3 Design and procedure

A within-subject design was used for this study. Trial presentation was blocked by experimental condition as the intensity of the carefully-articulated words was higher than that of the casually articulated words, which may result in better encoding of the louder items. One concern is that listeners adapt to this design, learning that productions without [t] should be interpreted as being a word with an underlying /t/. If so, this should apply equally to both the careful [n_] condition and the casual [n_] condition. Each block consisted of 64 trials: 8 critical prime-target pairs (e.g., center-town; splinter-wood), 8 critical targets preceded by semantically unrelated primes (e.g., bubble-wood), 16 real-real filler pairs (e.g., weather-fool), and 32 real-pseudoword filler pairs (e.g., cheeky-fush). All items were counterbalanced across blocks to avoid

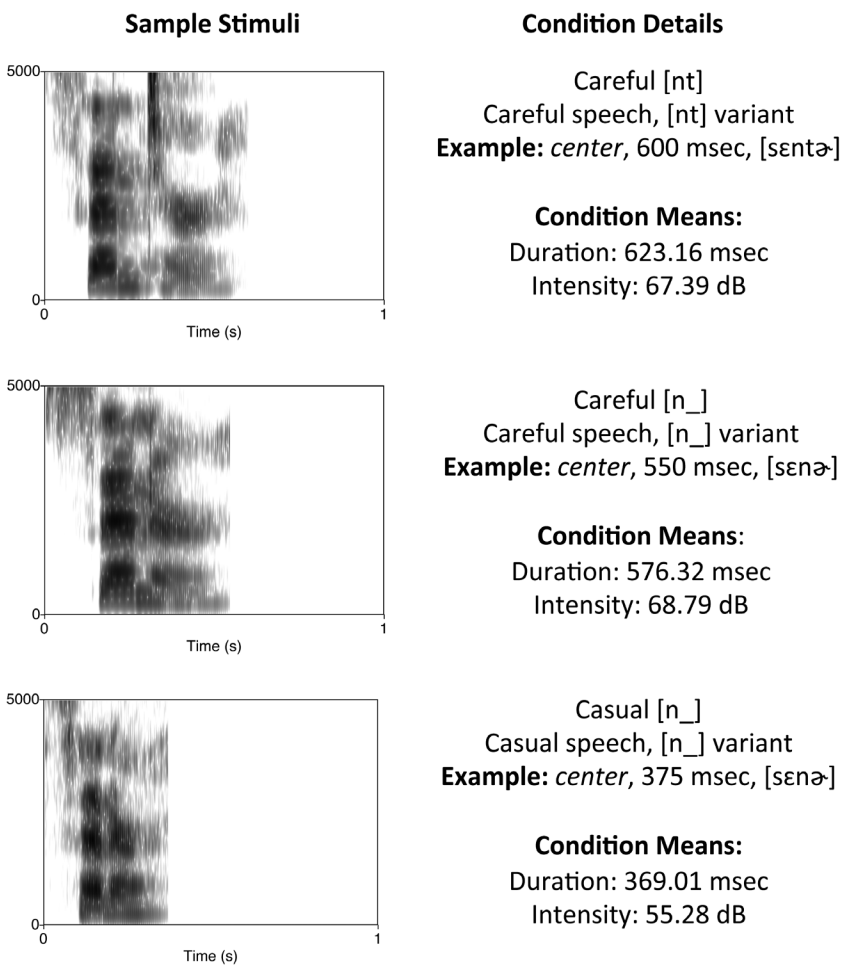

Fig. 1. Example stimuli representative of the three experimental conditions, with corresponding condition means for duration and intensity.

repetition. Each session included all 192 prime-target pairs. Presentation of trials within each block was randomized, and block order was randomized across participants.

The 60 participants were tested individually or in groups up to three in a sound-attenuated booth. A trial consisted of an auditory prime, a 100 ms interstimulus interval (ISI), presentation of a visual target, followed by a response. A new trial began 1000 ms after a response, or after 3000 ms if no response was made. Participants were instructed that they would be presented with auditory-visual pairs of words, and they needed to decide whether the word on their monitors was a real word or a pseudoword. The experiment lasted 12 min.

If the canonical bias is independent of word-level phonetics, costs should be apparent for both the careful [n_] and casual [n_] conditions. If word-level phonetics drive the difference rather than the presence/absence of a particular variant, both the careful [nt] and casual [n_] primes should facilitate recognition to semantically related compared to unrelated targets.

2.2 Results and discussion

No differences in error rates were found across conditions. Reaction times shorter than 300 ms and greater than 1200 ms were excluded from all analyses, along with trials with incorrect responses (2.8% of the data were excluded). Mean reaction times to semantically related and unrelated targets by prime condition are provided in Table 1.

A repeated-measures analysis of variance was conducted by relatedness (related × unrelated) and prime condition (careful/[nt]; careful/[n_]; casual/[n_]). Main effects were found for relatedness [$F1(1,59) = 9.461$, $p < 0.01$; $F2\ (1,47) = 6.162$, $p < 0.05$) and prime condition [$F1(2,59) = 5.121$, $p < 0.01$; $F2\ (2,47) = 3.618$, $p < 0.05$). The interaction was significant by subjects and by items [$F1(2,59) = 4.005$, $p < 0.05$; $F2\ (2,47) = 3.986$, $p < 0.05$). Planned comparisons support an interpretation that the interaction is driven by the careful [n_] condition, as that condition was found to be different from both the careful [nt] condition (careful [nt]-careful [n_]: $F1(1,59) = 13.451$, $p < 0.01$; $F2(1,47) = 9.775$, $p < 0.01$; casual [n_]-careful [n_]: $F1(1,59) = 15.863$, $p < 0.01$; $F2(1,47) = 8.696$, $p < 0.01$), and no difference was found between the careful [nt] and casual [n_] condition [$F1(1,59) < 1$; $F2(1,47) = F < 1$].

The results replicate two main findings: (1) priming effects are strong for carefully articulated primes with canonical variants and (2) priming effects are absent for carefully articulated primes with a non-canonical variant. Were it not for the casual [n_] condition, we might take this as further support for the canonical bias. However,

Table 1. Mean response times (ms) to semantically related and unrelated targets by prime condition for Experiment 1 (top) and to old and new targets for Experiment 2 (bottom). Percent incorrect responses are provided in parentheses.

Experiment 1: Target type			
Prime condition	Related	Unrelated	Priming effect
Careful [nt]	595 (2.1)	629 (2.4)	−34 ms
Careful [n_]	637 (1.9)	635 (2.2)	+2 ms
Casual [n_]	601 (2.2)	628 (2.0)	−27 ms
Experiment 2: Block 2 Target type			
Example: Target SENNER	New (Prime not presented in B1)	Old (Prime presented in B1)	Priming effect
Block 1 Prime condition			
Careful [nt]: [sɛntɚ] 623 ms	644 (3.4)	691 (3.3)	+47 ms
Careful [n_]: [sɛnɚ] 576 ms	639 (3.3)	632 (3.4)	−7 ms
Casual [n_]: [sɛnɚ] (369 ms)	638 (3.5)	682 (3.3)	+44 ms

when a visual target is preceded by a casual production of a word with a non-canonical variant, facilitation is robust.

The careful [n_] condition may be costly because the careful frame without [t] is a better match for a pseudoword string than the intended word (which is how we may understand a carefully articulated *winner* as *winner* and not *winter*).

3. Experiment 2: Long-term repetition priming of pseudowords

Experiment 2 uses long-term repetition priming to examine (1) whether the careful [n_] string is a better acoustic match to a pseudoword (e.g., *senner*) than to the real word *center* and (2) whether the careful [nt] and casual [n_] productions result in equally robust lexical activation over the long-term. Pseudoword repetition oftentimes results in a null effect, as two opposing processes are at work (the familiarity of having previously heard a string before (bias toward a "word" response) and the lack of lexical support for a string (bias toward a "pseudoword" response); Wagenmakers *et al.*, 2004). Building on this finding, Sumner and Samuel (2007) have shown that hearing a word inhibits the ability of a listener to subsequently reject a similar pseudoword. The two findings together make this methodology well suited to the goals of this study.

3.1 Methods

3.1.1 Participants and stimuli

Forty eight native speakers of American English participated in the study for pay. None reported any hearing difficulties. Participants were split into three groups, corresponding to three experimental conditions. Using the 48 critical primes with medial /nt/ from experiment 1, the three critical prime conditions remained the same. The target for each prime was an orthographic medial NN pseudoword (e.g., SPLINNER from *splinter*). 48 bisyllabic fillers that did not contain NT clusters were used (e.g., *fluffy*), as were 48 pseudowords created by changing one sound from an existing bisyllabic word [e.g., *fotion* (from *lotion*)].

3.1.2 Design and procedure

A between-subject design was used. Within each condition, two counterbalanced lists were created. Each list contained two blocks. The first block consisted of 24 critical primes (careful [nt]; careful [n_]; or casual [n_]), 24 pseudoword fillers, and an additional 36 real word fillers. Each target block contained all 48 critical targets [half corresponding to a prime in block 1 (old), half with no block 1 correspondent (new)], 48 pseudowords (half old, half new), 36 repeated, and 36 new filler words.

Auditory primes were presented one at a time. Lexical decisions were avoided as listeners show facilitatory effects in pseudoword repetition when using this task (see Zeelenberg *et al.*, 2004). This facilitation is linked to form-based, not lexical, activation. Participants pressed a button at item completion. As analyses are dependent on block 2 performance, this task ensured that listeners were attending to the critical block 1 primes, but also avoided explicit decisions about them. Participants made lexical decisions to the visual items in the target block.

3.1.3 Results and discussion

No differences in error rates were found across conditions. Response times shorter than 300 ms and greater than 1200 ms were excluded from all analyses, as were trials with incorrect responses (3.5% of the data). Mean lexical decision response times to correctly recognized pseudoword block 2 targets are provided in Table 1.

Two-factor analyses of variance [prime condition (careful [nt], careful [n_], casual [n_]) × target type (old or new)] of correct responses to critical targets were performed for subjects (F1) and items (F2). Response times were significantly slower for old targets than for new targets [$F1(1, 47) = 21.386$, $p < 0.001$; $F2(1, 47) = 7.623$, $p < 0.01$].

A main effect of the prime condition was also found [F1(2, 46) = 5.993, $p < 0.01$; F2(2, 46) = 3.386, $p < 0.05$]. An interaction between prime condition and target type was also found [F1(2, 45) = 5.371, $p < 0.01$; F2(2, 47) = 4.116, $p < 0.05$]. The interaction reflects the inhibition effect for conditions that include pronunciation variants paired with congruent word-level phonetic patterns, along with a lack of inhibition for Careful [n_] items. Planned comparisons support this interpretation. Pseudoword targets preceded by careful [nt] or casual [n_] primes in block 1 were recognized more slowly than the same targets with no block 1 correspondent. (Careful [nt]: F1(1, 15) = 18.293, $p < 0.001$, F2(1, 47) = 9.657, $p < 0.01$; Casual [n_]: (1, 15) = 15.331, $p < 0.001$, F2(1, 47) = 6.89, $p < 0.05$.) For the careful [n_] condition, old pseudoword targets were not recognized differently from new pseudoword targets [F1(1, 15) < 1, F2(1, 47) < 1]. The data suggest that (1) the bias toward canonical variants is not due to an idealized lexical representation, (2) a reduced, surface-based variant is not costly to spoken word recognition, (3) the recognition of words with different pronunciation variants depends heavily on the phonetic composition of the entire word, and (4) variants that are experienced with different frequencies result in strikingly similar patterns.

4. Discussion

This study examined the semantic and long-term repetition priming of words and pseudowords by auditory primes that were carefully or casually articulated and contained either a canonical [nt] or a non-canonical [n_] variant. The experiments were designed to address the canonical bias and reported equivalence across word forms with different variant frequencies.

While we may develop a probabilistic account, dependent on the acoustic cues in the first syllable, that approach may predict some graded difference between the two variants given vastly different usage patterns. One possibility is that these effects are less about a particular variant and more about processing differences between careful and casual speech [see also Mattys *et al.* (2009) for an analogous argument pertaining to word segmentation]. In this case, careful speech is processed as a clear (and less ambiguous) signal and casual speech is processed as a noisy, (more) ambiguous signal. In other words, the default processing mode for casual speech may involve a stronger reliance on top-down information than that of careful speech (which itself relies more on the bottom-up signal). This account centers effects of pronunciation variants on word-level phonetic variation, while avoiding issues any frequency-based theory of spoken word recognition might face—specifically, the equivalence of two strings with different global likelihoods (though certainly bound by social factors). This explanation accounts for the difference between the careful [nt] and careful [n_] conditions (where the former matches existing representations and the latter is a clear signal with no matching representation), as well as the equivalence between the careful [n_] and casual [n_] conditions. In one case, careful [n_], a clear signal is less likely to induce top-down processing. The clear speech stream contains ideal examples of the sounds [s] [ɛ] [n] [ɚ]. The absence of [t] is neither cued by sufficient residual phonetic cues nor by the top-down influence of the lexicon. In the other case, casual [n_], a casual signal is processed as a noisy signal, inducing top-down processing, with one match—*center*. This account extends nicely to pairs like *winner-winter* where some amount of semantic ambiguity *may* be present (even with subtle residual cues) when articulated casually, but predicts that neither form should be ambiguous when articulated clearly.

This study showed that multiple pronunciation variants are recognized equally well by listeners when produced in a congruent phonetic word frame. The data support the view that the canonical bias may be partially bolstered by our comparisons, and not by an idealized representation of a word. A process-based account might best capture the understudied behavioral similarities across variants that differ greatly in global production rates. Stronger bottom-up influences may lead to socially- or linguistically-

weighted encoding differences between spoken words, explaining long-term memory benefits for perceived standard forms.

Acknowledgments

I am grateful to Reiko Kataoka and Ed King for helpful comments and to Christopher Frederick for help with data collection. This material is based upon work supported by NSF Grant No. BCS 1226963 to M.S.

References and links

Andruski, J. E., Blumstein, S., and Burton, M. (**1994**). "The effect of subphonetic differences on lexical access," Cognition **52**, 163–187.
de Jong, K. (**1998**). "Stress-related variation in the articulation of coda alveolar stops: Flapping revisited," J. Phonetics **26**, 283–310.
Ernestus, M., Baayen, H., and Schreuder, R. (**2002**). "The recognition of reduced word forms," Brain Lang. **81**, 162–173.
Gaskell, M. G., and Marslen-Wilson, W. D. (**1996**). "Phonological variation and inference in lexical access," J. Exp. Psychol. Hum. Percept. Perform. **22**, 144–158.
Goldinger, S. D. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psychol. Rev. **105**, 251–279.
Gow, D. W. (**2003**). "Feature parsing: Feature cue mapping in spoken word recognition," Perception Psychophys. **65**, 575–590.
Gow, D. W., Jr. (**2001**). "Assimilation and anticipation in continuous spoken word recognition," J. Mem. Lang. **45**, 133–159.
Johnson, K. (**2006**). "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," J. Phonetics **34**, 485–499.
Marslen-Wilson, W., and Warren, P. (**1994**). "Levels of perceptual representation and process in lexical access: Words, phonemes, and features," Psychol. Rev. **101**, 653–675.
Mattys, S. L., Brooks, J., and Cooke, M. P. (**2009**). "Recognizing speech under a processing load: Dissociating energetic from informational factors," Cognitive Psychol. **59**(3), 203–243.
McLennan, C. T., Luce, P. A., and Charles-Luce, J. (**2003**). "Representation of lexical form," J. Exp. Psychol. Learn. Mem. Cogn. **29**, 539–553.
McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (**2002**). "Gradient effects of within-category phonetic variation on lexical access," Cognition **86**, B33–B42.
Mitterer, H., and McQueen, J. (**2009**). "Processing reduced word-forms in speech perception using probabilistic knowledge about speech production," J. Exp. Psychol. Hum. Percept. Perform. **35**, 244–263.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker contingent processes," Psychol. Sci. **5**, 42–46.
Pitt, M. A. (**2009**). "The strength and time course of lexical activation of pronunciation variants," J. Exp. Psychol. Hum. Percept. Perform. **35**, 896–910.
Ranbom, L. J., and Connine, C. M. (**2007**). "Lexical representation of phonological variation in spoken word recognition," J. Mem. Lang. **57**, 273–298.
Sumner, M., and Samuel, A. G. (**2005**). "Perception and representation of regular variation: The case of final-/t/," J. Mem. Lang. **52**, 322–338.
Sumner, M., and Samuel, A. G. (**2007**). "Lexical inhibition and sublexical facilitation are surprisingly long lasting," J. Exp. Psychol. Learn. Mem. Cogn. **33**, 769–790.
Tucker, B. V. (**2011**). "The effect of reduction on the processing of flaps and /g/ in isolated words," J. Phonetics **39**, 312–318.
Tucker, B. V., and Warner, N. (**2011**). "Inhibition of processing due to reduction of the American English flap," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken (August 2007).
Wagenmakers, E.-J., Zeelenberg, R., Steyvers, M., Shiffrin, R. M., and Raaijmakers, J. G. W. (**2004**). "Nonword repetition in lexical decision: Support for two opposing processes," Q. J. Exp. Psychol. **57**, 1191–1210.
Zeelenberg, R., Wagenmakers, E.-J. , and Shiffrin, R. M. (**2004**). "Nonword repetition priming in lexical decision reverses as a function of study task and speed stress," J. Exp. Psychol. Learn. Mem. Cogn. **30**, 270–277.