Published in final edited form as:

J Acoust Soc Am. 2008 March ; 123(3): EL32.

Evoking biphone neighborhoods with verbal transformations: Illusory changes demonstrate both lexical competition and inhibition

James A. Bashford Jr, Richard M. Warren, and Peter W. Lenz

Department of Psychology, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee,
Wisconsin 53201-0413

Abstract

When a recorded verbal stimulus repeats over and over, perceptual changes occur and listeners hear competing forms. These verbal transformations (VTs) were obtained for a phonemically related set of 24 consonant-vowel syllables that varied widely in frequency-weighted neighborhood density (FWND). Listener's initial transformations involving substitution of consonants versus vowels were strongly correlated with the lexical substitution neighborhood [$r = +0.82$, $p < 0.0001$]. Interestingly, as stimulus FWND increased, average time spent hearing illusory forms substantially decreased [$r = -0.75$, $p < 0.0001$]. These results suggest that VTs not only reveal underlying competitors, but also provide a highly sensitive measure of lexical inhibition.

1. Introduction

Modern theories of spoken word recognition typically assume that acoustic-phonetic input activates sets of phonologically similar verbal representations in memory, which compete with one another in the process of word recognition. These activation-competition models of speech perception are supported by evidence from a variety of paradigms (e.g., shadowing, perceptual identification, and lexical decision), which have shown that processing speed and accuracy for a verbal stimulus are influenced by both its neighborhood density (i.e., the number of phonologically similar, lexical neighbors) and by its neighborhood frequency (the sum of the word frequencies of its lexical neighbors). Words having large numbers of high-frequency lexical neighbors are generally processed more slowly and less accurately than words having only a few, low-frequency lexical neighbors (e.g., Goldinger *et al.*, 1989; Luce and Pisoni, 1998; Vitevitch and Luce, 1998, 1999).

Studies dealing with the effects of neighborhood density and frequency typically employ a computational procedure that considers the effective neighbors to be those words in the lexicon that differ from the stimulus by the addition, deletion, or substitution of a single phoneme (Landauer and Streeter, 1973; Luce and Pisoni, 1998). The present study uses an auditory illusion known as the Verbal Transformation Effect (Warren, 1961) as a direct means of identifying salient neighbors, as well as examining their competitive interaction. When listeners are presented with a recorded verbal stimulus, such as a syllable or word, which repeats over and over without change, they typically hear abrupt and compelling illusory changes to other words or syllables (Warren, 1961). These verbal transformations (VTs) appear to result from the operation of two simultaneous processes (Warren, 1996): (1) a repetition-induced adaptation effect, which lowers the activation level of the dominant neural representation best fitting the stimulus; and (2) a repetition-induced summation effect that progressively increases

the activation level of neural representations that are structurally similar to the stimulus.¹ It is considered that VTs occur when the diminished activation level of the stimulus representation is exceeded by that of the most highly activated competing representation. Hence, it is suggested that reports of apparent changes may provide a direct means of accessing the sets of competing representations that have been inferred to influence the speed and accuracy of performance in a variety of psycholinguistic tasks, as well as in everyday spoken word recognition.

A recent experiment (Bashford, Warren, and Lenz, 2006) explored the use of VTs in the study of neighborhood competition. Listeners were presented with repeating lexical and nonlexical consonant-vowel-consonant (CVC) stimuli that were either very low or very high in frequency-weighted neighborhood density (FWND). About 75% of the initial VTs evoked by high-FWND words and nonwords were lexical neighbors of the stimuli. In contrast, the low-FWND stimuli evoked relatively few lexical neighbors (14%), but did evoke a large percentage (47%) of lexical non-neighbors, which typically differed from the stimuli by 2 phonemes and had very high word frequencies (averaging about 880 wpm). These findings indicate that VTs are strongly influenced by two factors considered to be of central importance in modern theories of word recognition: the structural similarity and word frequency of stimulus competitors. The present study uses VTs to measure the effects of a third factor, lexical inhibition, which is incorporated as an important component of the competitive process in several word-recognition models [e.g., TRACE (McClelland and Elman, 1986), SHORTLIST (Norris, 1994), and PARSYN (Luce *et al.*, 2000)]. The capacity of word representations to inhibit each other, in proportion to their goodness of fit with speech input, could aid word recognition by amplifying activation-level differences among competing representations. However, this mutual inhibition would also lower the activation level of all competitors, including the stimulus representation, thus accounting at least in part for the elevations in response latency and error rate observed for high-density stimuli in a variety of psycholinguistic tasks. Paradoxically, lexical inhibition should improve perceptual accuracy in the VT paradigm. Specifically, it was predicted that the amount of time repeating stimuli are heard accurately should increase with increasing stimulus FWND, producing a decrease in the strength of the VT illusion, since greater numbers of high-frequency neighbors would be expected to exert greater amounts of inhibition, which in turn should lower the activation level of the stimulus representation, and thereby reduce the adapting effects of repeated stimulation. In order to test this prediction, it was decided to use a series of naturally produced stimuli having minimal phonetic complexity: a set of 24 consonant-vowel (CV) syllables, 11 words and 13 nonwords, which represent a factorial pairing of the six English stop consonants [three voiced: /b/, /d/, /g/; and three unvoiced: /p/, /t/, /k/] with each of four steady-state English vowels: /a/, /i/, /u/, and /e/. An additional set of 6 vowel-consonant (VC) syllables paired the series of six English stops with the preceding vowel /u/; this series provided control comparisons to be discussed in the results section. These biphone stimuli cover a large range of FWND (sums of the log frequencies ranging from 1.301 to 90.42), which permits a strong test of the lexical inhibition hypothesis.

The 24 CV stimuli employed in the main experiment not only varied in FWND, but also differed critically in the phonemic composition of their lexical substitution-neighborhoods (i.e., those subsets of neighbors differing from the stimuli via phoneme substitution rather than phoneme addition or deletion). For 12 of the CV syllables, more substitution neighbors differed from the stimulus at the consonant position, and for the remaining 12 syllables, more substitution neighbors differed at the vowel position. Based on the hypothesis that VTs are determined chiefly by lexical neighborhood competition, rather than some form of sublexical processing,

¹For a discussion of repetition-induced verbal summation occurring independently of adaptation, see Warren *et al.* (1996).

it was predicted that initial substitution transformations would generally involve the stimulus phoneme conflicting with the greatest number of lexical substitution neighbors.

2. Method

The 300 listeners in this study (10 groups of 30) were undergraduate students at the University of Wisconsin-Milwaukee who were paid for their participation in sessions lasting about 30 min. All listeners were native monolingual English speakers who reported having no hearing problems and had normal bilateral hearing, as measured by pure tone thresholds of 20 dB hearing level or better at octave frequencies from 250 to 8000 Hz.

The 24 CV syllables and 6 VC syllables described in the introduction were digitally recorded (44.1 kHz sampling, 16 bit quantization) by a highly practiced announcer having an average voicing frequency of approximately 100 Hz and speech patterns typical of southeast Wisconsin. The monosyllables were produced to fit within a 340 ms capture window, and an additional 250 ms segment of digital silence was then added to each capture to emphasize syllable boundaries and minimize any tendency for perceptual resegmentation [e.g., /ti/ heard as /it/] (see Bashford *et al.*, 2006). The 340 ms stimulus capture, along with the added 250 ms silent gap, was digitally iterated to produce a 5 min test stimulus that provided 508 repetitions.

The 240 listeners employed in the main experiment were randomly divided into 8 groups of 30, and each group received a different set of three CV stimuli, having either the three voiced or three unvoiced stop consonants paired with one of the four vowels. The remaining two groups of 30 listeners, who were presented with the VC stimuli, received the vowel /u/ paired with either the three voiced or three unvoiced stop consonants. The order of stimulus presentation within groups was pseudorandom, with the restriction that each stimulus was presented an equal number of times in each serial position across listeners in a group. Testing was performed in a sound-attenuating chamber, with the VT stimuli delivered diotically through Sennheiser HD 250 Linear II Headphones at a slow-rms peak level of 70 dBA sound pressure level. For each stimulus presented, listeners were instructed to call out what the voice was saying at stimulus onset, and then to call out what the voice was saying anytime a change was heard. Listeners' responses during the 5 min stimulation periods were transcribed by the experimenter.

3. Results and discussion

As in our earlier study (Bashford *et al.*, 2006), listeners' initial reported transformations were used for the primary analysis of forms. Nearly all (99.8%) of these initial forms reported for the 24 CV syllables were either English words or phonotactically legal nonwords. They were categorized as one of four types: (1) lexical neighbors (words differing from the stimulus by a single phoneme); (2) nonlexical neighbors; (3) lexical non-neighbors (words differing by more than one phoneme); and (4) nonlexical non-neighbors. Reports of stimulus neighbors comprised 77% of all initial VT responses, with 48% being lexical neighbors and 29% nonlexical neighbors. Reports of non-neighbors as VTs comprised 23% of initial responses, with 13% being lexical and 10% nonlexical.² The frequency of reports for non-neighbors was not correlated with stimulus FWND, either for lexical responses [$r=0.18$, $F(22)=0.74$, $p>0.5$] or for nonlexical responses [$r=0.08$, $F(22)=0.1$, $p>0.70$]. In contrast, report frequencies for lexical and non-lexical neighbors of the stimuli did vary reciprocally with FWND: There was a strong direct correlation between FWND and the report frequency for lexical neighbors [$r=$

²Verbal transformations involving multiple phonemic changes have been found to decrease in frequency with the age of the listener, from childhood through old age (Warren and Warren, 1966), indicating that stimulus neighborhoods become phonetically more restricted throughout the lifespan.

+0.735, $F(22) = 25.9$, $p < .0001$] and a strong inverse correlation with the report frequency for non-lexical neighbors [$r = -0.834$, $F(22) = 50.4$, $p < 0.0001$].

The increase in reports of lexical neighbors with increasing stimulus FWND is consistent with earlier results from this lab (Bashford *et al.*, 2006), and is also consistent with the hypothesis that VTs reveal the most salient neighbors of repeating stimuli. However, reports of lexical neighbors also would be expected to increase with FWND to some extent by chance alone if VTs were due to random errors occurring at a phonetic level of processing. If transformations are indeed determined by interactions between lexical representations, then the phonetic transformations observed for VT stimuli should be consistent with the composition of their lexical neighborhoods. Specifically, it was predicted that substitution transformations of stimulus consonants versus vowels should be correlated with the proportion of lexical neighbors that differ from the stimuli at the consonant versus vowel positions. Strong confirmation of this predicted pattern was obtained through a regression analysis that compared the proportion of consonant versus vowel substitution transformations for the 24 CV syllables with the proportion of lexical neighbors differing from the stimuli by consonant versus vowel substitution [(consonant substitution neighbors – vowel substitution neighbors) / total substitution neighbors]. The proportion of initial transformations involving consonant changes increased in a very strong linear fashion ($r = +0.82$, $F(22) = 45.2$, $p < 0.0001$) as the proportion of consonant-substitution neighbors increased relative to vowel-substitution neighbors. This correlation accounts for about 67% of the variance in consonant versus vowel substitution transformations, despite the inclusion of data for non-neighbor transformations that involved changes in both the consonant and vowel components. Hence, a substantial portion of VTs are clearly not the result of random errors in sublexical processing, but rather are strongly influenced by interactions between lexical representations.

Although substitution transformations were the most frequent initial responses (comprising 63.5% of reported changes), transformations involving the addition of a vowel or consonant (comprising about 30% of responses) occurred with sufficient frequency to permit an examination of the predictability of consonant-addition versus vowel-addition transformations based on the proportion of consonant- versus vowel-addition neighbors. Interestingly, and in contrast with the results obtained for substitution transformations, the latter correlation did not approach significance ($r = 0.163$, $F(22) = 0.6$, $p > 0.4$). This disparity in the effects of substitution- versus addition-neighbors was also found in the analysis of overall illusion strength discussed below.

The final prediction tested in this study was that the strength of the VT illusion, measured in terms of the amount of time stimuli were heard nonveridically during the 300 s repetition period, would decrease with increasing stimulus FWND. This prediction is based on the hypothesis that increasing FWND should result in increased lexical inhibition, which should reduce the activation levels of stimulus representations and thereby attenuate the adapting effects of their repeated activation in the VT paradigm. The results obtained for the 24 CV stimuli clearly conform to this prediction: As FWND increased, there was a highly reliable linear decrease in the average amount of time listeners reported hearing the stimuli nonveridically. This strong negative correlation ($r = -0.746$, $F(22) = 27.5$, $p < 0.0001$) accounted for approximately 56% of the variance in illusion strength. Additional regression analyses separately examining the data for lexical and nonlexical stimuli confirmed that the negative correlation between FWND and illusion strength was reliable for both the 11 lexical CVs ($r = -0.67$, $F(9) = 6.3$, $p < 0.03$) and the 13 nonlexical CVs ($r = -0.57$, $F(11) = 5.3$, $p < 0.05$).

Although the decline in VT strength observed with increasing FWND is consistent with the predicted inhibitory effect of lexical competition, it is necessary to consider a possible sublexical effect of phonotactic probability. Previous studies have shown that speech

processing is influenced not only by the density and frequency of lexical neighbors (FWND), but also by the position-specific frequency with which stimulus phonemes and phoneme sequences occur within words and syllables of the listener's language (e.g., Vitevitch and Luce, 1998, 1999). Accordingly, these phonotactic probabilities were obtained for the 24 CV stimuli (see Vitevitch and Luce, 2004) and submitted to regression analyses, which indicated that the strength of the VT effect (i.e., nonveridical percept time) was not correlated with the probability of either the consonants ($r = -0.144$, $F(22) = 0.46$, $p > 0.50$) or the CV biphones ($r = +0.11$, $F(22) = 0.264$, $p > 0.6$). Correspondingly, there was no correlation between FWND and the probability of the consonants ($r = +0.03$, $F(22) = 0.02$, $p > 0.90$) or biphones ($r = +0.02$, $F(22) = 0.01$, $p > 0.90$). However, there was a strong negative correlation between FWND and vowel probability for this set of stimuli ($r = -0.91$, $F(22) = 93.3$, $p < 0.0001$) and, correspondingly, there was a strong correlation between vowel probability and VT strength ($r = +0.79$, $F(22) = 35.7$, $p < 0.0001$). Given the absence of correlation between VT strength and the probabilities of either the consonants or biphones, it appeared likely that the significant correlation with vowel probability was the spurious consequence of its very high correlation with FWND. This hypothesis was tested by combining the data for the 6 CV syllables ending with the vowel /u/ with those obtained for the 6 VC syllables that were comprised of the same phonemes in reverse order. For this set of 12 syllables, the correlation between vowel probability and FWND was also very high, but was positive rather than negative in slope ($r = +0.917$, $F(10) = 52.9$, $p < 0.0001$). As expected, the correlation between vowel probability and VT strength was also reversed in direction ($r = -0.69$, $F(10) = 8.9$, $p < 0.02$), while, in contrast, the correlation between FWND and VT strength was unchanged in direction. Stimulus FWND, the apparent causal factor, accounted for approximately 68% of the variance in VT strength for the matched set of CV and VC syllables ($r = -0.827$, $F(10) = 21.6$, $p < 0.001$).

The dominance of FWND over phonotactic probability in this study is not surprising given the extended stimulation period employed in the VT paradigm. Effects of phonotactic probability may override those of FWND when stimuli, especially nonwords, are presented in speeded tasks that do not require lexical analysis (e.g., phoneme identification); but FWND has been found to dominate processing for both words and nonwords when lexical analysis (e.g., lexical decision) is required (e.g., Vitevitch and Luce, 1998, 1999). The 5 min stimulus-repetition periods in the present study certainly provided ample time for the activation of lexical representations and the consequent buildup of inhibition.

It should be noted that the effect of FWND on VT strength was not only highly reliable but also substantial: The five CV stimuli having the lowest FWND values (sums of the log frequencies of lexical neighbors averaging 20.8) were heard nonveridically for an average of 248 s (standard error = 15.3 s) during the 300 s repetition interval, whereas the five CVs highest in FWND (averaging 86.2) were heard nonveridically for an average of only 155 s (standard error = 11.1 s). This mean shift of nearly 40% suggests that the VT strength (i.e., illusion time) measure may provide an especially sensitive index of lexical inhibition.

A final analysis of the VT-strength data for the complete set of 30 biphones examined the relative salience of competition effects exerted by substitution neighbors versus addition neighbors. When the FWND scores for these two types of lexical neighbors were used as predictors in separate regression analyses, significant negative correlations between VT strength and FWND were obtained for both the stimulus substitution-neighborhoods [$r = -0.733$, $F(28) = 32.6$, $p < 0.0001$] and addition neighborhoods [$r = -0.460$, $F(28) = 7.54$, $p < 0.02$]. However, the substitution- and addition-neighborhood FWND scores for this set of stimuli were also correlated [$r = 0.398$, $F(28) = 5.26$, $p < 0.03$], so first-order partial correlation coefficients were computed to better determine the individual contributions of the different neighbor types to VT inhibition. When variance in common between substitution- and addition-neighborhood FWND values was partialled out, the correlation between VT strength and

substitution-neighborhood FWND remained highly significant [$pr = -0.675$, $t(27) = -4.76$, $p < 0.0001$] while that involving addition-neighborhood FWND was nonsignificant [$pr = -0.27$, $t(27) = -1.46$, $p > 0.15$]. These results obtained with the VT-strength measure converge with those from listeners' initial VT reports, which showed that the relative frequencies of specific phoneme transformations were strongly correlated with the lexical substitution neighborhoods of the stimuli, but not with their addition neighborhoods. This disparity can also be seen in the types of forms reported in this study. About 44% of addition transformations yielded non-neighbors of the stimuli, either through multiple phoneme additions or through addition combined with a phoneme substitution. In contrast, only about 18% of substitutions yielded non-neighbors, and in most instances these transformations involved a single phoneme substitution combined with one or more phoneme additions. Hence, while substitution transformations conform strongly to the similarity neighborhood of the stimulus, addition transformations do not.

At present it is not clear whether the null effects obtained for addition-neighborhood composition in this study reflect a general lack of salience of addition neighbors in speech processing. It is possible that addition neighbors do effectively compete with substitution neighbors under normal listening conditions but become ineffective when stimuli are repeated for an extended period. It should be noted that the VT paradigm does appear to be inappropriate for the study of possible deletion-neighborhood effects: Pitt and Shoaf (2001, 2002) have shown that word repetition, especially with a very brief interstimulus interval (ISI), evokes the general auditory process of stream segregation (Bregman, 1990), which can cause spectrally distinct consonants such as stops (especially unvoiced stops) to split from the remainder of an utterance and be heard as a separate background stream. This segregation effect can also be incomplete (Pitt and Shoaf, 2002), so that a portion of stop-consonant aspiration may remain grouped with a following vowel, resulting in the substitution of /h/ (e.g., /pi/ heard as /hi/). In the present study, deletion transformations were reported infrequently (6% of initial changes), probably due to the 250 ms ISI separating stimulus repetitions. Consistent with Pitt and Shoaf (2002), simple deletion responses occurred most frequently for the voiceless stops, and most often when they were syllable final (VC stimuli). Moreover, about 1% of transformations involved substitution of /h/ for the stop consonants, all in syllable-initial position and most unvoiced.

Deletion transformations may be too heavily confounded with auditory streaming to provide useful information about speech-specific processing. However, substitution transformations are clearly very sensitive to a lexical level of analysis: In the present study, stimulus variation in substitution-neighborhood composition accounted for nearly 70% of the variance in phonetic substitution transformations, and 50% of the variance in overall VT illusion time. These values compare favorably with those obtained using other psycholinguistic tasks, in which neighborhood FWND, the most potent stimulus variable, seldom accounts for more than 20% of performance variance (e.g., Luce and Pisoni, 1998).

The large inhibitory effect of neighborhood density on VTs was not restricted to the illusion-time measure, but also was observed for the number of different forms reported by individual listeners. Forms decreased in a strong linear fashion with increasing FWND [$r = -0.74$, $F(22) = 28.5$, $p < 0.0001$]. Interestingly, a paper by Yin and MacKay (1992, cited by MacKay *et al.*, 1993) reported the opposite result for a set of 20 words varying in neighborhood density, which suggests that the dominance of neighborhood inhibition observed with simple biphone stimuli in the present study may give way to dominance by activated competitors—producing a greater range of transformations—when stimulus complexity is increased. This possibility, which is consistent with the present finding that neighborhood effects upon VTs are chiefly manifested through substitution transformations, will be examined in further planned work that will

systematically increase syllabic complexity, beginning with CVCs composed of the same stops and vowels used for the biphones of this study.

Acknowledgements

This work was supported in part by Grant No. R01DC000208 from the National Institute on Deafness and Other Communication Disorders. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Deafness and Other Communication Disorders or the National Institutes of Health.

References and links

- Bashford JA Jr, Warren RM, Lenz PW. Polling the effective neighborhoods of spoken words with the verbal transformation effect. *J Acoust Soc Am* 2006;119:EL55–EL59. [PubMed: 16642865]
- Bregman, AS. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press; Cambridge, MA: 1990.
- Goldinger SD, Luce PA, Pisoni DB. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *J Mem Lang* 1989;28:501–518.
- Landauer T, Streeter LA. Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *J Verbal Learn Verbal Behav* 1973;12:119–131.
- Luce PA, Goldinger SD, Auer ETJ, Vitevitch MS. Phonetic priming, neighborhood activation, and PARSYN. *Percept Psychophys* 2000;62:615–662. [PubMed: 10909252]
- Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. *Ear Hear* 1998;19:1–36. [PubMed: 9504270]
- MacKay DG, Wulf G, Yin C, Abrams L. Relations between word perception and production: New theory and data on the verbal transformation effect. *J Mem Lang* 1993;32:624–646.
- McClelland JL, Elman JL. The TRACE model of speech perception. *Cognit Psychol* 1986;18:1–86. [PubMed: 3753912]
- Norris D. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 1994;52:189–234.
- Pitt MA, Shoaf L. Linking verbal transformations to their causes. *J Exp Psychol Hum Percept Perform* 2002;28:150–162.
- Pitt MA, Shoaf L. The source of lexical bias in the verbal transformation effect. *Lang Cognit Processes* 2001;16:715–721.
- Vitevitch MS, Luce PA. When words compete: Levels of processing in perception of spoken words. *Psychol Sci* 1998;9:325–329.
- Vitevitch MS, Luce PA. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *J Mem Lang* 1999;40:374–408.
- Vitevitch MS, Luce PA. A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behav Res Methods Instrum Comput* 2004;36:481–487. [PubMed: 15641436]
- Warren RM. Illusory changes of distinct speech upon repetition—the verbal transformation effect. *Br J Psychol* 1961;52:249–258. [PubMed: 13783239]
- Warren, RM. Auditory illusions and perceptual processing of speech. In: Lass, NJ., editor. *Principles of Experimental Phonetics*. Mosby; St Louis: 1996. p. 435–466.
- Warren RM, Healy EW, Chalikia MH. The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. *J Acoust Soc Am* 1996;100:2452–2461. [PubMed: 8865650]
- Warren RM, Warren RP. A comparison of speech perception in childhood, maturity, and old age by means of the verbal transformation effect. *J Verbal Learn Verbal Behav* 1966;5:142–146.
- Yin, C.; MacKay, DG. Auditory illusions and aging: Transmission of priming in the verbal transformation paradigm. Poster presented at the Cognitive Aging Conference; Atlanta. 1992.