Relative contributions of initial and final similarity to neighborhood density effects on English spoken word recognition

José R. Benkí— Communicative Sciences and Disorders, Michigan State University
Robert Felty— Departments of Linguistics and German, University of Michigan
*Work supported by NIH/NIDCD DC05913

Introduction

Word similarity, also known as phonological neighborhood density, is a significant factor in the speed and accuracy of spoken word recognition (Luce 1986; Luce and Pisoni 1998). Words with few similar-sounding confusors are more quickly and better identified than words with many similar-sounding confusors.

Previous research on English CVC word recognition in noise indicates that phonological neighborhood density effects are overwhelmingly determined by the neighborhood defined by the initial CV of a CVC target (Benkí 2003). The neighborhood defined by the final VC or both consonants (CC) of a target was not found to contribute to the neighborhood density effects. The present study explores two potential nonexclusive explanations for that asymmetry.

- The *acoustic-phonetic* hypothesis: word beginnings are more intelligible than endings (for both acoustic and perceptual reasons), and therefore are simply more available during word recognition.
- The *dynamic* hypothesis: the temporal nature of auditory stimuli and their perception dictate that early-occurring material is available to affect the perception of late-occurring material is perceived, but the reverse is generally not true.

Research Goals

- Determine whether final similarity contributes to neighborhood density effects in spoken word recognition
- Compare any such contributions with those of initial similarity in a quantitative manner
- Determine whether either the acoustic-phonetic or dynamic hypotheses explain previous failures to observe final similarity contributions
- Extend the j-factor model of context effects (Boothroyd and Nittrouer 1988) to stimuli with variable masking and uncertainty

Method

Materials The stimuli consisted of 300 CVC English words, selected under two constraints. The first constraint was that the final consonants /p d t g k s z m n l/ were equally represented (i.e., 30 each) in the stimuli. The second constraint was that the frequency-weighted neighborhood density (Luce and Pisoni 1998) of the stimuli be maximized. Density (FWNP) was calculated using confusion matrices of American English initial consonant, vowel, and final consonant recognition in noise (Benkí and Felty 2005). Thus, the operationalization of density in the present study is a phonetic neighborhood density metric based on an empirical measure of perceptual similarity, rather than a phonological neighborhood density metric based on an edit-distance measure.

Listeners 39 subjects were recruited from the University of Michigan. Subjects reported normal-hearing and were native speakers of English.

Task Listeners were randomly assigned to one of two S/N ratios (-6 or -9 dB) and to one of three stimulus lists (see below). Stimuli were presented over headphones and listeners typed in what they heard using standard orthography. Signal dependent noise was added to the stimuli according to the method described by Schroeder (1968). All of the stimuli were presented to each listener in three different conditions, each consisting of 10 blocks of 10 stimuli:

- Control
- Masked C1: The S/N ratio was lowered by 6 dB during the initial consonant and half of the vowel. This manipulation was designed to evaluate the *acoustic-phonetic* hypothesis.
- Blocked: In the final condition, each block of 10 contained the same final consonant, so listeners knew the identity of C2. This manipulation was designed to evaluate the *dynamic* hypothesis.

Three different stimulus lists were prepared such that each stimulus appeared once in each condition. Each stimulus list contained one instance of each of the 300 stimuli. Presentation order within each condition was random.

Analysis Stimuli were split into two sets by median values of the phonetic FWNP counting contributions from all 1-phoneme neighbors. A similar analysis was done for CV neighbors, VC neighbors, and CC neighbors. The phoneme and syllable identification scores for each subject and density group were calculated and analyzed using the j-factor model of Boothroyd and Nittrouer (1988). The j-factor model assumes that phonemes are the basic unit of speech, and that phonemes are perceived independently (which has been shown to hold true most of the time; see Fletcher (1953); Allen (1994)). The probability of correctly identifying a given CVC word (or nonword) syllable s can be calculated as the product of the probabilities of its constituent phonemes.

$$p_s = p_{C1} p_{V1} p_{C2} \tag{1}$$

where p_s is the probability of correctly identifying a word (or nonword). Assuming that phonemes are perceived independently, (1) can be rewritten as:

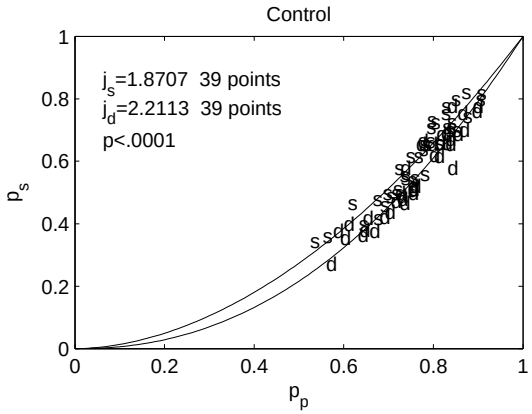
$$p_s = p_p^j \tag{2}$$

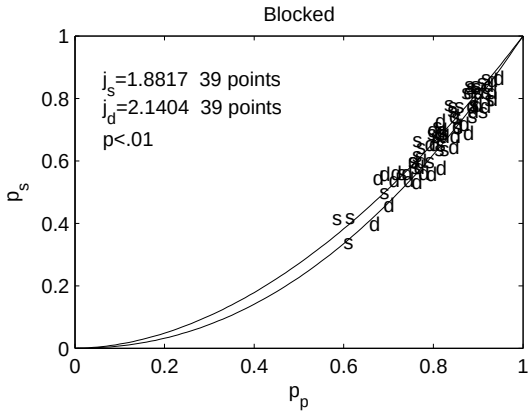
where j is the number of phonemes, and p_p is the geometric mean of the probabilities of each constituent phoneme. Rewriting (2), the quantity j can be empirically determined from syllable and phoneme scores by:

$$j = \frac{log(p_s)}{log(p_p)} \tag{3}$$

Predictions

- Control: CV density effects with a difference in j of ≈ 0.4 (dense $<$ sparse).
- Masked C1: reduction in CV density effects and possible appearance of VC density effects (acoustic-phonetic hypothesis).
- Blocked: appearance of VC density effects (dynamic hypothesis).





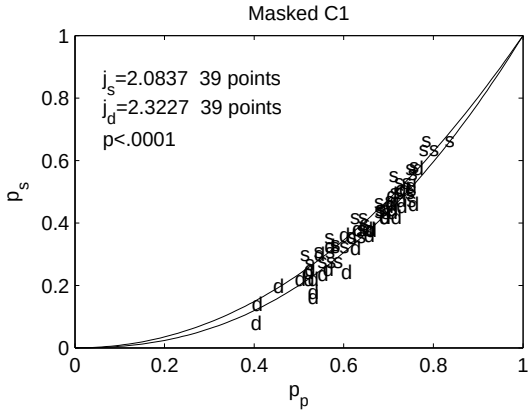
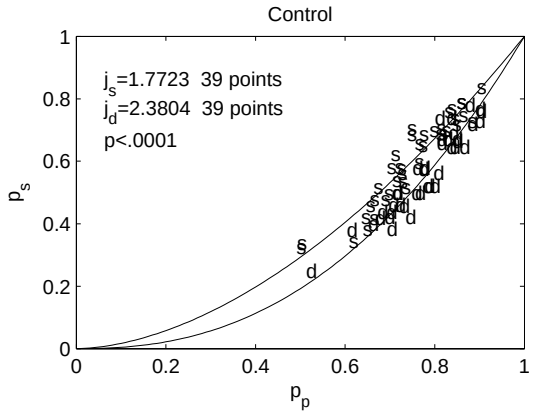
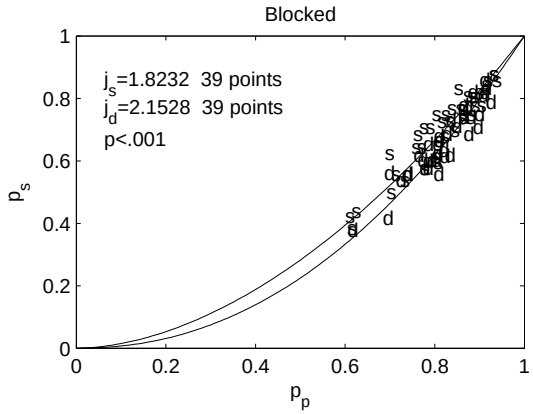


Figure 1: CV density j-factor results - Each plot compares CV density for one of the experimental conditions. Each point represents half of the responses in a single condition for a single listener. Curves represent $y = x^j$, averaged across all listeners for either the dense or sparse subsets. *p*-values given are from 2-sample t-tests.





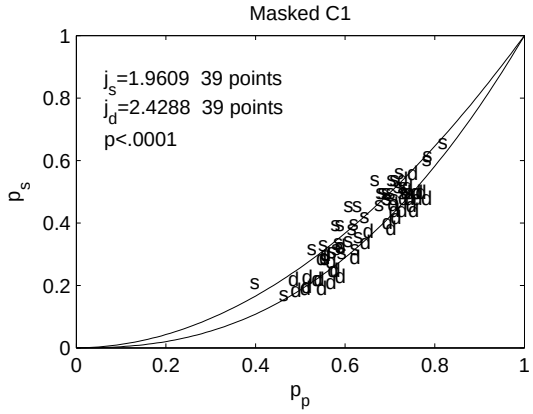


Figure 2: VC density j-factor results - Each plot compares VC density for one of the experimental conditions. Each point represents half of the responses in a single condition for a single listener. Curves represent $y = x^j$, averaged across all listeners for either the dense or sparse subsets. *p*-values given are from 2-sample t-tests.

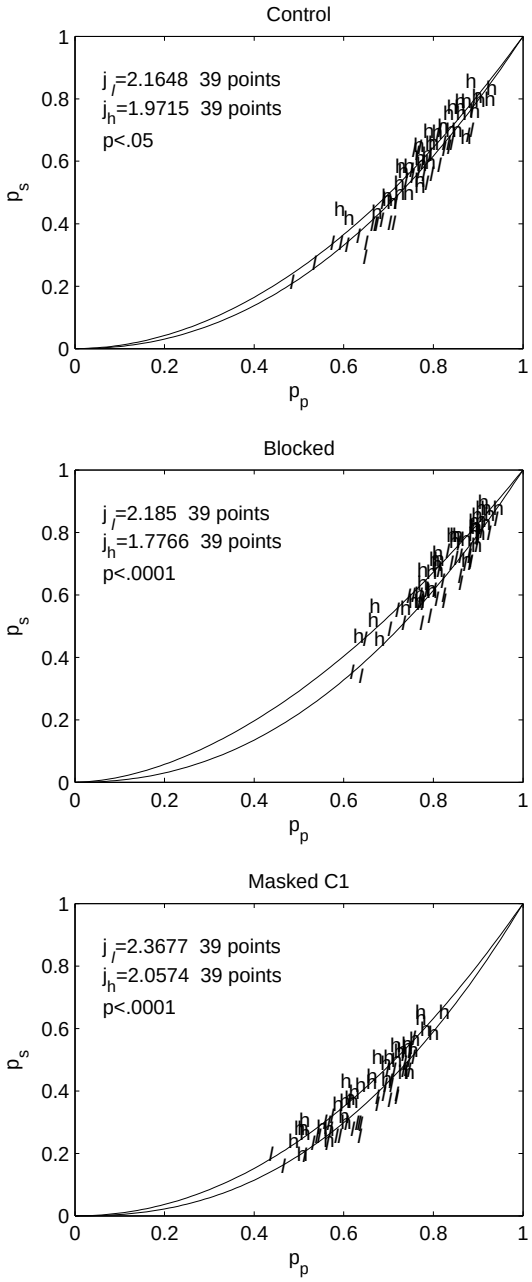


Figure 3: Frequency j-factor results - Each plot compares Kucera-Francis log-frequency for one of the experimental conditions. Each point represents half of the responses in a single condition for a single listener. Curves represent $y = x^j$, averaged across all listeners for either the low or high log-frequency subsets. p-values given are from 2-sample t-tests.

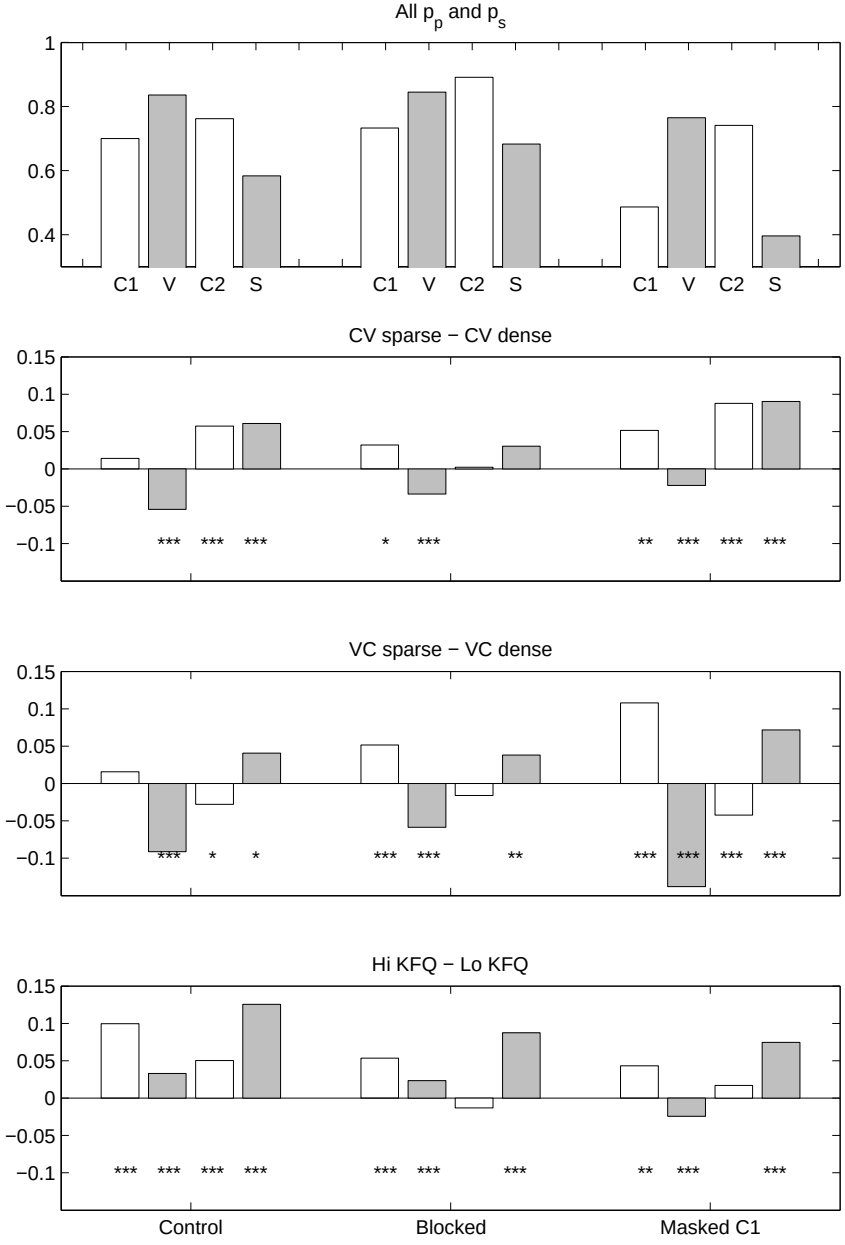


Figure 4: Average phoneme and syllable identification scores. The top panel presents the means for the entire experiment by condition. The subsequent panels present the mean difference between sparse and dense (or high and low frequency) for each score. A simple 2-tailed paired t-test shows statistically significant differences. A single asterisk indicates $p < 0.05$, two asterisks indicate $p < 0.01$, and three asterisks indicate $p < 0.001$. Statistical comparisons are *not* corrected for multiple tests.

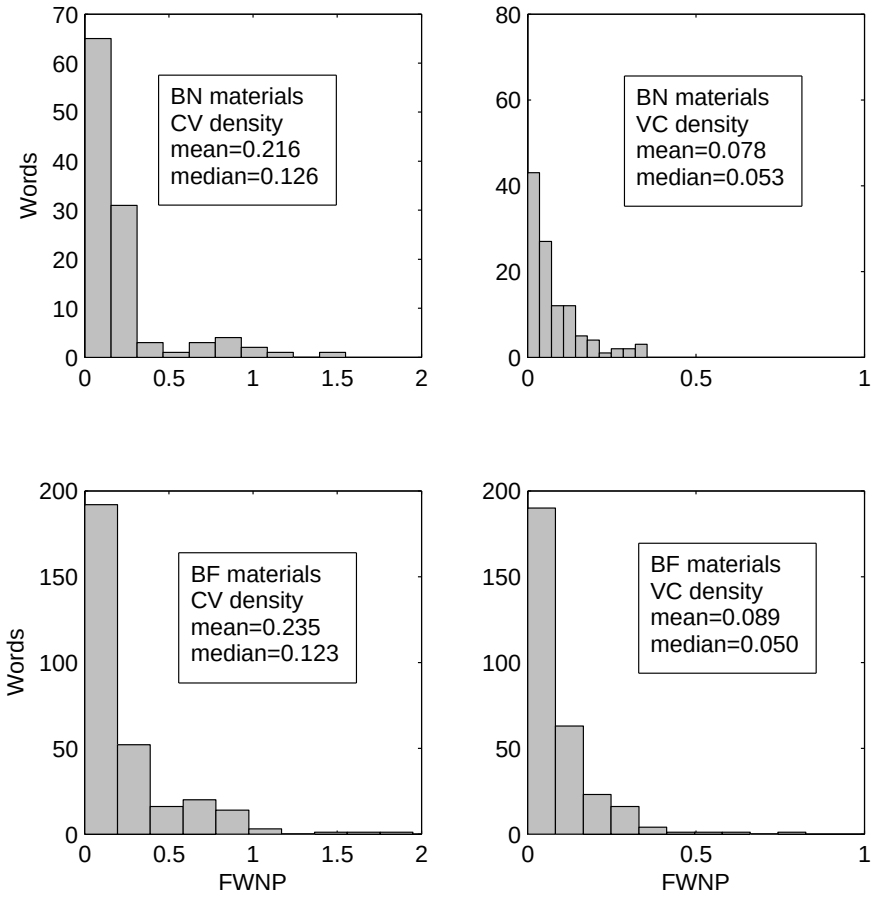


Figure 5: Histograms of the CV and VC frequency-weighted neighborhood probability for Boothroyd and Nittrouer (1988) / Benkí (2003) materials and the materials in the present study.

Results

The results are shown in Figures 1-4. Figures 1-3 provide j-factor subjects analyses, while Figure 4 presents the average phoneme and syllable scores by experimental condition. There were no significant differences in the j-factor analyses by S/N ratio, and no significant effect of CC density. The mean j-factor for the whole experiment is ≈ 2.1, compared with 2.35 in Benkí (2003).

Both CV and VC density effects are present in all three experimental conditions. Unexpectedly, VC density effects are present in the control condition, and CV density effects are present in the masked C1 condition. The magnitude of the CV density effect is consistent with Benkí (2003).

A frequency effect is also present in all three conditions, again consistent with Benkí (2003), but the magnitude of the effect appears to be larger in the present study.

Discussion

The acoustic-phonetic and dynamic hypotheses

The unexpected presence of VC density effects in the control (and to some extent in the masked C1 condition) might be explained in part by the higher identification scores for C2 than C1, providing partial support for the acoustic-phonetic hypothesis. The materials used by Benkí (2003) (from Boothroyd and Nittrouer (1988)) were well-matched phonemically for C1 and C2, with the consequence that identification scores were generally higher for C1. In the present study, because of the constraints in stimulus selection, C1 was much more varied and contained a number of fricatives and sonorants that are more confusable on average. The histograms in Figure 5 indicate that the densities in both studies are similar enough, however, that further explanation is needed.

The persistence of the CV density effects in the masked C1 condition (though reduced) provides support for the dynamic hypothesis in that degradation of C1 intelligibility was not sufficient to eliminate CV density effects. A coherent interpretation of all results is that the stimulus selection in the present study enabled VC density effects to emerge not only in the blocked and masked C1 conditions, but the control condition as well. A key prediction is that a C2 masked condition should reduce or eliminate the VC density effects observed in the control condition. In contrast, it may not be possible to strongly reduce or eliminate CV density effects given the temporal nature of speech stimuli as explained by the dynamic hypothesis.

A final unexpected finding was the smaller value of j relative to previous studies of CVC English words, which have found $j \approx 2.4$. At least two factors may be contributing to this reduction. Firstly, the present study uses the geometric mean of the phoneme scores, which while more mathematically sound, will always provide a smaller estimate of j given the same raw scores. Secondly, the mean Kucera-Francis frequency of the stimuli in the present study (18) is higher than the previous study (10), and frequency is known to be inversely correlated with j.

References

Allen, Jont. 1994. How do humans process and recognize speech?, *IEEE Transactions of Speech and Audio Processing*, 2(4), 567–577.

Benkí, José. 2003. Quantitative evaluation of lexical status, word frequency and neighborhood density as context effects in spoken word recognition, *Journal of the Acoustical Society of America*, 113(3), 1689–1705.

Benkí, José and Robert Felty. 2005. Recognition of english phonemes in noise, *Journal of the Acoustical Society of America*, 117, 2568.

Boothroyd, A. and S. Nittrouer. 1988. Mathematical treatment of context effects in phoneme and word recognition, *Journal of the Acoustical Society of America*, 84, 101–114.

Fletcher, Harvey. 1953. *Speech and Hearing in Communication*, New York: Krieger.

Luce, Paul. 1986. *Neighborhoods of words in the mental lexicon*, Ph.D. thesis, Indiana University.

Luce, Paul and David Pisoni. 1998. Recognizing spoken words: The neighborhood activation model, *Ear and Hearing*, 19, 1–36.

Schroeder, M. 1968. Reference signal for signal quality studies, *Journal of the Acoustical Society of America*, 44, 1735 – 1736.