Measures of segmental lexical statistics

To evaluate the possibility that the preference of items like *mlɟ* reflects only the co-occurrence of their segments in the English lexicon, we calculated several statistical measures of our materials. These measures correspond to factors which have been reported in the literature as modulating perceptual accuracy and speed. The means of these measures, provided in Table 1, reflect averages computed over the twelve items representing each onset type (these means were not used in our analyses; they are presented merely as descriptive statistics). A brief description of these measures is found in the target paper—below we offer a more detailed description of these measures, their calculation and their expected effects on behavior.

The statistical properties included neighborhood measures and measures of segment or letter co-occurrence (for auditory and printed materials, respectively). A final measure concerned the identity of the initial consonants.

1. ***Neighborhood measures***.

A target's lexical neighborhood comprises all words obtained by adding, deleting or substituting one of a target's phonemes (or letters, for printed words). Previous research suggests that words with a large neighborhood consisting of frequent words are recognized more readily in naming and AX tasks (e.g., Carreiras *et al.*, 1997; Perea & Carreiras, 1998; Michael S. Vitevitch & Luce, 1998; Michael S. Vitevitch & Luce, 1999). The better recognition of the items with rising-sonority onsets might thus be due to the structure of their lexical neighborhoods, rather than sonority per se. We evaluated this possibility using two neighborhood measures:

- Neighborhood count: The number of lexical neighbors

- Neighborhood frequency: The summed frequency of a target's neighbors

For example, the item [mlɪf] has one phonological neighbor, /klɪf/, whose frequency is 11 per million. Both measures were calculated form the *Speech and Hearing Lab Database* (http://128.252.27.56/Neighborhood/Home.asp)[1].

2. ***Measures of segment/letter co-occurrence***.

Words whose segments or letters co-occur frequently are better recognized (e.g., Perea & Carreiras, 1998; Michael S. Vitevitch & Luce, 1998; Michael S. Vitevitch & Luce, 1999). To determine whether the advantage of sonority rises is due to these properties, we computed two sets of measures of co-occurrence, one for the whole word and one for the onset specifically.

a. *Whole word measures*. The co-occurrence of elements in the word as a whole was estimated using either segment- or letter co-occurrence, for auditory and printed words, respectively. Segment co-occurrence in auditory words was captured by the following measures based on the *Phonotactic Probability Calculator* (Michael S. Vitevitch & Luce, 2004):

- Position-sensitive phoneme probability: The probability that any given phoneme in the target occurs at the same string position (first through fourth), averaged across the target's four phonemes.
- Position-sensitive bi-phone probability: The probability that any adjacent phoneme-pair in the target occurs at the same string position (averaged across the target's three bigrams).

For printed words (in Experiment 6), we used two measures of letter co-

[1] At present, the *Speech and Hearing Lab Database* incorrectly collapses case-distinctions in phonological inputs (e.g., it fails to distinguish [lif] and [lIf]). In view of this problem, we manually inspected the output of the database and corrected them as necessary.

occurrence:

- Bigram count: The number of words sharing each of the target's adjacent letter-pairs in the same string position (calculated based on Solso & Juel, 1980).
- Bigram frequency: The summed frequency of the words sharing the target's bigrams (from Kucera &Francis, 1967).

For example, the phoneme probability of the auditory [mlɪf] is 0.0382 (averaged across the positional probability of its four phonemes: .0572, .0447, .0350, .0159). In its printed form, *mlɪf* shares its second bigram (*lɪ*) with 10 words whose summed frequency is 85 per million (the initial and final bigrams are not shared with any words), so its bigram count is 10, and its bigram frequency is 85.

b. *Onset-probability measures*. To examine the possibility that the advantage of sonority rises might be due to the individual probabilities of occurrence of their onset consonants in their string positions, we computed the log of the product of the position-specific probabilities of each consonant (i.e., the summed frequency of words sharing that consonant in its position relative to the total summed frequency of the sample) in CCVC words listed in the *Speech & Hearing Lab Neighborhood Database*. Consider the onset *ml*. The database includes a total of 757 CCVC words, three of which begin with an *m*[2] (*mews, mule, mute*) and 199 words that include an *l* in the second position, The probabilities of *m*-first and *l*-second words, respectively, are 0.003963012 and 0.262879789 and what we call the "log cluster probability" for *ml* is the log of their product (0.001041796): 2.982217428.

[2] To avoid biasing the calculations by theory-internal structural assumptions, we based frequency calculations on the segment sequences as indicated in the database (see also Section 4.2).

3. ***The onset's initial consonant***.

Previous research suggests that *n*-initial onsets might be less perceptible than *m*-initial onsets for reasons unrelated to sonority (Byrd, 1992; Surprenant & Goldstein, 1998). Because failure to register the acoustic input as an onset cluster (e.g., registering *nbʧ* as *bʧ*) reduces its markedness, we also entered the identity of the initial consonant into the analysis of auditory stimuli.

Table 1. Some Statistical Properties of the Auditory and Printed Stimuli

	Onset type	
	Sonority Rise	Sonority Fall
Statistical property	Auditory items	
Number of Neighbors	1.58	0.50
Neighbors' Frequency (summed)	581	390
Mean phoneme probability	0.0316	0.0262
Mean biphone probability	0.0016	0.0006
Log cluster probability	-5.4097	-10.7296
	Printed items	
Number of Neighbors	0.33	0.33
Neighbors' Frequency (summed)	1.25	2.00
Bigram Count (summed)	13.67	6.25
Bigram Frequency (summed)	316.92	224.50
Log cluster probability	-5.4097	-10.7296

Note. The figures provided in the table reflect means per onset type (averaged across the 12 items representing each onset type).

References

Byrd, Dani. (1992). Perceptual assimilation in consonant clusters: A gestural model. *Phonetica* **49.** 1-24.

Carreiras, Manuel, Manuel Perea & Jonathan Grainger. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23.** 857-871.

Kucera, Henry & W. Nelson Francis. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Perea, Manuel & Manuel Carreiras. (1998). Effects of syllable frequency and neighborhood frequency in visual word recognition. . *Journal of Experimental Psychology: Human Perception and Performance* **24.** 134-144.

Solso, Robert L & Connie L Juel. (1980). Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods & Instrumentation* **12.** 297-343.

Surprenant, Aimee M. & Louis Goldstein. (1998). The perception of speech gestures. *JASA* **104.** 518-529.

Vitevitch, Michael S. & Paul A. Luce. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods Instruments, & Computers* **36.** 481-487.

Vitevitch, Michael S. & Paul A. Luce. (1998). When words compete: Levels of processing in the perception of spoken words. *Psychological Science* **9.** 325-328.

Vitevitch, Michael S. & Paul A. Luce. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* **40.** 374-408.