Misperceptions of spoken words: Data from a random sample of American English words

Robert Albert Felty[a)]
Nuance Communications, 1198 East Arques Avenue, Sunnyvale, California 94085

Adam Buchwald
Department of Communicative Sciences and Disorders, New York University, 665 Broadway, Suite 910, New York, New York 10012

Thomas M. Gruenenfelder and David B. Pisoni
Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th, Bloomington, Indiana 47405

(Received 17 April 2012; revised 6 February 2013; accepted 20 May 2013)

This study reports a detailed analysis of incorrect responses from an open-set spoken word recognition experiment of 1428 words designed to be a random sample of the entire American English lexicon. The stimuli were presented in six-talker babble to 192 young, normal-hearing listeners at three signal-to-noise ratios (0, +5, and +10 dB). The results revealed several patterns: (1) errors tended to have a higher frequency of occurrence than did the corresponding target word, and frequency of occurrence of error responses was significantly correlated with target frequency of occurrence; (2) incorrect responses were close to the target words in terms of number of phonemes and syllables but had a mean edit distance of 3; (3) for syllables, substitutions were much more frequent than either deletions or additions; for phonemes, deletions were slightly more frequent than substitutions; both were more frequent than additions; and (4) for errors involving just a single segment, substitutions were more frequent than either deletions or additions. The raw data are being made available to other researchers as supplementary material to form the beginnings of a database of speech errors collected under controlled laboratory conditions.
 [http://dx.doi.org/10.1121/1.4809540]

PACS number(s): 43.71.Es, 43.71.Gv [BRM] Pages: 572–585

I. INTRODUCTION

What types of errors do listeners make when they misperceive a spoken word? The purpose of the present study was to develop and make available to other researchers a corpus of misperceptions of a relatively large number of different spoken words. The present study was primarily descriptive and did not test any particular hypothesis or theoretical issue concerning spoken word recognition. Nevertheless, we characterized the errors obtained from a word recognition task on the basis of several formal and lexical variables known to influence spoken word recognition. These analyses and the corpus of errors can be used in future work attempting to further address the question laid out at the top of this paragraph.

Previous studies of misperceptions of spoken words fall along a continuum represented by experimental rigor on the one end, where carefully selected stimuli are presented under well-controlled laboratory conditions, and by ecological validity on the other end, where perceptual errors made in everyday conversations are simply noted. We first review studies that fall more toward the "experimental rigor" end of the continuum and then review studies that have collected errors made in everyday speech. We then describe the present study, which attempted to find more of a midpoint between these two ends of that continuum.

One of the earliest studies to examine spoken word recognition errors in the laboratory was carried out by Miller and Nicely (1955). Listeners in their study were required to identify the initial consonant in /Ca/ syllables (where C indicates one of 16 consonants). Based on their analysis of the resulting confusion matrix, Miller and Nicely concluded that the various articulatory features (voicing, nasality, affrication, place of articulation) of the consonants were perceived independently of one another. Wang and Bilger (1973) later extended the confusion matrix published by Miller and Nicely to include all possible English consonants in both Consonant–Vowel (CV) and Vowel–Consonant (VC) contexts, using three different vowels. They concluded that their observed confusion matrices did not support the hypothesis that there exists a set of natural perceptual features that listeners use to identify phonological segments. This conclusion was based on their observation that several different feature sets could all account equally well for the confusion matrix data.

In terms of addressing our specific question, the Miller and Nicely (1955) and Wang and Bilger (1973) studies share two disadvantages. First, most of the stimuli were nonsense syllables, not real English words. Hence when listeners misperceived a stimulus, they were usually misperceiving a nonsense syllable, not a word. Second, listeners were constrained

a)Author to whom correspondence should be addressed. Electronic mail: robfelty@gmail.com

 0001-4966/2013/134(1)/572/14/$30.00

to respond with one of 16 consonants (the 16 alternatives varied from block-to-block in the Wang and Bilger study), making the task a closed-set task. In a closed-set task, listeners can only make errors pre-determined by the experimenter. Consequently, such studies do not necessarily tell us what errors listeners would make when left to their own devices. Moreover, the processing strategies adopted by listeners in closed-set tasks depend strongly on the nature of the available response alternatives in the closed-set and may not necessarily reflect processing strategies used in natural listening environments (e.g., Clopper *et al.*, 2006).

Pickett (1957) examined confusion matrices for English vowels. One of Pickett's conditions used lists of phonetically balanced (PB) words embedded in a carrier phrase and an open-set identification task in which the listener had to identify the spoken word (although only vowel errors were reported in the analysis). Based on analyses of the resulting vowel confusion matrices, Pickett reached a number of conclusions. For example, vowel confusions were consistent with the hypothesis that listeners use perceived formant frequencies to identify vowels and that duration strongly influences vowel perception when only a single formant is perceived. Pickett's study is perhaps the first to draw conclusions concerning speech perception processes based on the specific errors made when perceiving spoken words rather than relying solely on error rates.

Pollack *et al.* (1959) investigated listeners' ability to identify words at different signal-to-noise (S/N) ratios. Pollack *et al.* were primarily interested in the effects of a word's frequency of occurrence on its intelligibility and hence reported primarily percent correct data. Later, however, Pollack *et al.* (1960) analyzed the error responses made by listeners in a condition where the target word came from a set of 144 monosyllabic English words and in which the listeners were not provided a list of targets when making their responses. As was the case in the original Pollack *et al.* (1959) study, Pollack *et al.* (1960) focused on word frequency effects. They found that the word frequency (i.e., frequency of occurrence of the word in the language) of the error responses was not significantly affected by the word frequency of the target word. The word frequency of error responses, however, did increase as the S/N ratio increased. Subsequently, Gerstman and Bricker (1960) argued that this effect was due to a confounding in the study of Pollack *et al.* of S/N ratio and prior experience with the word lists.

At least on the surface, the condition analyzed by Pollack *et al.* (1960) was an open-set task because listeners were not given an explicit set of alternatives to choose from on each trial. However, the same set of 144 target words was repeated across multiple blocks of trials, giving listeners at least some opportunity to learn what the set of permissible responses was (cf. Gerstman and Bricker, 1960). Further, because all the words were monosyllabic, listeners probably quickly learned to confine their responses to single syllable words (although to be sure, approximately 10% of the error responses were non-words). Consequently, the Pollack *et al.* study is better considered to be a hybrid of open-set and closed-set tasks than to be a pure open-set task.

In terms of helping to answer the question posed at the beginning of this paper, concerning the types of errors that listeners make when they misperceive a word, there are two additional limitations of the Pollack *et al.* study. First, because the study confined itself to monosyllabic words, there is little opportunity to investigate the frequency with which listeners add or delete a syllable when misperceiving a word. Second, the analysis of Pollack *et al.* focused only on the frequency of occurrence in the language of the error responses. They did not report analyses of, for example, the segmental overlap between targets and errors or any other analysis of the similarity of the errors and target words.

More recently, Benkí (2003), using an open-set perceptual identification task, analyzed the error responses to spoken, non-word CVCs presented at various S/N ratios to test several hypotheses regarding the perceptual robustness of various consonantal and vowel contrasts. In a similar study, Cutler *et al.* (2004) had native and non-native speakers of English identify the consonant or vowel in VC and CV syllables. All possible American English VC and CV syllables were used as stimuli. The purpose of the study of Cutler *et al.* was to determine whether noise differentially affected native and non-native listeners. Like the previously cited studies, the work of Benkí is not directly applicable to the question of what errors people make when they misperceive spoken words because Benkí used only non-word CVCs. Similarly, a preponderance of the Cutler *et al.* stimuli would also have been non-words.

All the studies cited in the preceding text were designed to test specific hypotheses regarding speech perception and have made important contributions to our understanding of processes involved in speech perception. Other research, conducted at the opposite end of what might be considered a continuum from hypothesis-testing research to exploratory research, has investigated "slips of the ear," that is, errors listeners make in perceiving speech in their everyday lives. The most extensive corpus of such perceptual errors for English has been collected (and made available) by Bond and her colleagues (Bond, 1999a,b, 2005; Bond and Garnes, 1980; Bond and Robey, 1983; Garnes and Bond, 1980). (See also Browman (1980) for a somewhat smaller corpus. Labov (2010), collected a corpus nearly as large as Bond's over a 14 yr period and reported a number of analyses of this corpus in his study of linguistic change. The corpus is not, however, to our knowledge, generally available.) The Bond corpus consists of nearly 900 misperceptions that occurred in everyday spoken English; what Bond refers to as "casual conversation." A little over 10% of the errors were made by children. Approximately 1/3 of the errors involved misperception of a single consonantal segment in one word, what Bond refers to as a simple misperception. Simple vowel misperceptions accounted for approximately 5% of the errors. The remainder of the errors involved multiple segments and often involved multiple words (Bond, 1999b).

Both Bond and Browman were well aware of the limitations of their observational data sets. For instance, for an error to be included in the corpus, it has to be detected by

one of the participants in the conversation. Hence the corpus is likely to include a disproportionately high number of semantically anomalous errors. Errors that preserve the semantic integrity of an utterance are less likely to be noticed (Browman, 1980). Put differently, the corpora probably reflect more bizarre errors disproportionately to the more mundane, hard-to-notice errors (Bond, 1999a). Nevertheless, analyses of such misperceptions can be used to at least provide some converging evidence on basic processes involved in speech perception (cf. Bond and Garnes, 1980). Bond (1999b), for example, used data from her corpus to suggest that the criteria for entertaining lexical candidates is determined primarily by phonological similarity and is little affected by restrictions on valid parts-of-speech or semantic context. Similarly, Browman (1980) used her database of slips of the ear to investigate the effect of word structure on misperceptions of segments in spoken words.

As hinted at in the preceding text, in comparison to the studies reviewed earlier (Benkí, 2003; Cutler *et al.*, 2004; Miller and Nicely, 1955; Pickett, 1957; Pollack *et al.*, 1960; Wang and Bilger, 1973), studies of slips of the ear tend to trade experimental rigor for increased ecological validity. The present study attempted to retain the experimental rigor of those earlier studies while at the same time increasing ecological validity by broadening the set of words used as target stimuli and using an open-set recognition task. In particular, we used as targets 1428 words randomly selected from a large database of American English words. Target words were spoken in isolation and were presented to young, normal-hearing listeners for identification in an open-set task using six-talker babble at three different S/N ratios. Our analyses focused on characterizing the errors and error patterns made by listeners when identifying these stimuli.

II. METHOD

A. Participants

Listeners for this study were 192 native American English speaking undergraduates from Indiana University. Participants received either $10 or course credit and reported no history of speech or hearing impairment at the time of testing.

B. Materials

A random sample of 1428 English words was selected from the Hoosier Mental Lexicon or HML (Nusbaum *et al.*, 1984). The sample was selected randomly to ensure that the words varied widely in both their lexical (e.g., lexical frequency; familiarity) and formal (e.g., number of syllables; syllable structure) properties. The words were randomly selected from the portion of the database containing words ranging in length from one to five syllables and two to 11 phonemes. This subset of the HML has 18 891 words. The full list of words used in this study is available in the supplementary material.[1]

The sample was compared to the HML on a number of lexical and formal dimensions. With respect to lexical variables, the sample did not differ significantly from the relevant portion of the HML with respect to log word frequency [HML: Mean (M) = 1.486; standard deviation (SD) = 0.686; sample: M = 1.492; SD = 0.688; $t(1427) = 0.372$; Cohen's $d = 0.010$, where word frequencies were obtained from CELEX (Baayen *et al.*, 1993)], with respect to familiarity [HML: M = 5.637; SD = 1.593; Sample: M = 5.639; SD = 1.593; $t(1427) = 0.053$; Cohen's $d = 0.001$], or with respect to number of phonological neighbors (words differing from the target word by the substitution, deletion or addition of a single segment) [HML: M = 3.231; SD = 6.386; sample: Mean = 3.300; SD = 6.434; $t(1427) = 0.408$; Cohen's $d = 0.011$].

The sample was also compared to the HML on four formal properties: Number of phonemes, number of syllables, syllable structure, and the first segment of the word. Words in the sample did have fewer phonemes on average than did words in the HML [HML: M = 6.207, SD = 2.138; sample: M = 5.886, SD = 1.955; $t(1427) = -6.20$, Cohen's $d = 0.15$.] Still, the sample did span a range of lengths greater than that covered in most studies of spoken word recognition. We also determined the extent to which the words in our sample were distributed across the different syllable lengths in similar proportions as in the HML. This comparison revealed no difference between the random sample and the HML in the proportion of words that were 1, 2, 3, 4, and 5-syllables long [χ^2 (4, $N = 1428$) = 2.543, $p = 0.64$]. To explore this distribution more fully, the distribution of syllable structures occurring in the random sample was compared to the distribution of the 173 syllable structures of the words in the HML to determine whether the proportions of words with particular syllable structures differed between the two lists. A chi-square test showed that the distribution of syllable structures of the 1428 words in the random sample did not differ from the expected distribution based on the HML [χ^2 (172, $N = 1428$) = 140.971, $p = 0.96$]. The proportion of words beginning with different initial segments in our sample also did not differ from the HML across the 41 initial segments [χ^2 (40, $N = 1{,}428$) = 29.75, $p = 0.88$]. With the exception of length in terms of number of phonemes, these analyses indicate that our sample of words from the HML did not differ from the overall HML in a variety of key measures.

The stimulus materials were recorded by two native speakers of American English (one male, one female) in an IAC booth and digitally sampled at 22.05 kHz. Each word was recorded once by each talker. The words were produced in isolation. Six-talker babble randomly sampled from the Connected Speech Test (Cox *et al.*, 1987) was added to the stimuli at three different signal-to-noise ratios (S/N): 0, +5, and +10 dB, with 500 ms leading and trailing babble surrounding the target word.

Each of the 192 listeners heard a unique list of 357 words, one-quarter of the full set of stimuli. Words were assigned to the lists randomly subject to the following constraints. All words on a given list were spoken by the same talker. One-third of the words on each list were presented at each of the three S/N ratios. Across the 192 lists, each word from the full set of stimuli occurred exactly eight times at each combination of S/N ratio and speaker. The listener heard the list one time; hence, there were 357 trials per listener.

C. Procedure

Listeners heard the recorded materials over Beyer Dynamic D-210 headphones at 70 dB Sound Pressure Level (SPL) and were instructed to identify the English word spoken by the talker. Each listener heard one of the 192 lists described in the preceding text. Listeners were informed that all the stimuli would be English words but that some would be rare words. Listeners entered their responses via keyboard. Stimulus presentation was controlled by software developed at the Speech Research Laboratory at Indiana University using PSYSCRIPT software (Bates and D'Oliveiro, 2003).

D. Analysis

A total of 68 544 trials were presented to listeners. On 592 trials, the listener did not enter a response, or entered a random response such as *asdf*, leaving 67 952 trials for these analyses (In the supplementary material, these 592 trials are classified as NORESP.). Responses were converted into phonetic transcriptions semi-automatically. If the response was in the HML (52 798 of the responses), then the HML transcription was used. Otherwise, if the response was in the Carnegie Mellon University (CMU) pronouncing dictionary (Carnegie Mellon University, 2007) (5085 of the responses), then the CMU pronouncing dictionary was used. If the response was in neither the HML nor in the CMU pronouncing dictionary, CELEX transcriptions were used (410 responses). If the response was in none of those three databases, then transcriptions were created by the first author (RAF) in collaboration with one or more research assistants (9659 of the responses). Frequency of occurrence information was based on the CELEX database (Baayen *et al.*, 1993) and on the SUBTLEXus database (Brysbaert and New, 2009).

For responses that were not in the HML, the CMU pronouncing dictionary, or the CELEX database, the responses were checked manually by a laboratory research assistant and the first author (RAF). We classified the responses into the following categories: (1) misspelling (e.g. *plian* for *plain*), (2) nonword (e.g. *nisc, vicundity*), (3) missing (e.g. *google*), (4) foreign (e.g. bjorn, puedes), (5) multiple (e.g. *and then, both men*), and (6) neologism (e.g. *untypical, righten*). Table I shows the proportion of errors that were classified as nonword, foreign, multiple, or neologism. Responses were categorized as a misspelling if the response was not a real English word and a simple change would make it a real word. For misspelled words, we converted the word into the correct spelling before checking for accuracy. Thus in some cases, fixing the spelling resulted in treating the response as correct, whereas in some cases, the response was still wrong but considered to be a word response, as opposed to a nonword. If there was any doubt, the response was classified as nonword. To determine whether a response was a real English word, we performed a search using our customized version of the CELEX database (i.e., the version using HML and CMU pronouncing dictionary phonetic transcriptions). In some cases, we determined that a response that was missing from the database should be considered a real word, e.g. *google* or *laptop*. These responses were classified as missing. Because the stimulus list contained some proper nouns, we also added other missing proper nouns, e.g. *Michigan* or *Stephen*. For these missing words, phonetic transcriptions were created using native speaker knowledge from the lead author (RAF) as well as one or two research assistants. Other sources such as the American Heritage Dictionary were consulted when needed. When necessary, transcriptions for responses classified as foreign were determined by consulting a dictionary in the appropriate language. Similar to the case for missing words, transcriptions for nonword responses were created by the lead author (RAF) in consultation with a research assistant.

TABLE I. Percent of responses in each error category.

Category	Number of responses	Percent of errors	Examples
WORD	21 844	72.4	Purse, skew
NONWORD	8061	26.7	Nisc, vicundity
FOREIGN	23	0.1	Bjorn, puedes
MULTIPLE	125	0.4	And then, both men
NEOLOGISM	102	0.3	Untypical, righten

Responses classified as missing accounted for 3% of the errors (1.3% of all responses). Because these responses were real words that were not included in either the HML or CELEX databases, they were treated as words for purposes of analyses and are included in the word category in Table I. A word frequency estimate for these missing words was calculated based on Google page counts. That calculation was made as follows. The Google page count for the 100 most frequent words in the CELEX database was compared with the CELEX frequency. The mean ratio of the Google page count to CELEX frequency was 94 778.56. For each missing word, the Google page count for that word was divided by 94 778.56 and that result used as an estimate of the word's frequency. For example, on June 16, 2008 (the date that the page counts were obtained from Google), the word *google* itself, which was not in the CELEX corpus, had a page count of 2.78×10^9. This was converted to 2.78×10^9 divided by 94 778.56 or 29 332. Despite these being rough estimates, there was a strong correlation between the CELEX frequency (of the 100 most frequent CELEX words) and the Google page count [$r = 0.753$, a result significantly greater than 0, $t(98) = 11.34$, $p < 0.0001$]. We recognize that this method of estimating word frequency is less than ideal (cf. Kilgariff, 2007). However, we felt it would provide at least an approximate estimate of the frequency of occurrence of these missing words as testified to by the strong correlation between CELEX frequencies and Google page counts. Note that this method of estimating frequency was done only for words that had no CELEX frequency. Because missing words comprised only 3% of our error data and because Google page counts do correlate with more conventional measures of frequency, using frequency measures other than the Google page counts is unlikely to have a substantial effect on the pattern of our results.

Subsequent to our initial word frequency measurements Brysbaert and New (2009) have made available the SUBTLEXus database of word frequencies, which they have shown to be a somewhat better predictor of performance in

printed word lexical decision than are CELEX frequencies. However, using SUBTLEXus frequencies for our error responses would have resulted in a larger number of MISSING responses than did using CELEX frequencies (see following text). Hence, when reporting word frequency analyses for errors, we use CELEX frequencies rather than SUBTLEXus frequencies (although we did use SUBTLEXus frequencies in analyses of accuracy data).

III. RESULTS

Of the 67 952 trials for analysis (the 592 NORESP trials were excluded), 37 797 (55.6%) trials were classified as correct responses and 30 155 (44.4%) as incorrect. Of the incorrect responses, 8061 (26.7% of the errors) were non-words. If neologisms and foreign words are counted as non-words, then 8186 errors were non-words (27.1% of the errors). Although participants were told that all the stimuli were English words, because we were using a random sample of American English words, some of the stimuli were low frequency and were perhaps not familiar to some participants. Consequently, on some trials, participants may have misperceived a word as a non-word but thought that they had correctly perceived a rare word with which they were unfamiliar and reported that "word" as their perception.

After providing analyses of the overall percent correct, we report several analyses of the errors made by listeners. Those analyses included all errors, both words and non-words, and examined the edit distance between errors and targets, the difference in number of phonemes and syllables between errors and target, and the frequency of occurrence in the language of errors relative to targets. Finally, we describe an analysis comparing incorrect responses that were words to incorrect responses that were non-words. See supplementary material for the full set of responses,[1] both correct and incorrect (including trials on which the listener made no response or made a random response and were scored as NORESP) as well as pertinent associated information.

A. Percent correct

The main purpose of the present study was to make available to other researchers our collection of spoken word recognition errors and to characterize those errors on several variables that have been of interest to researchers of spoken word recognition. Nevertheless it might be useful to provide at least a summary of performance in terms of overall proportion correct.

We first examined proportion correct across the two speakers used in the present study. The two speakers differed considerably in their intelligibility with the female speaker being more intelligible overall than the male speaker (cf. Bradlow *et al.*, 1996). While the total proportion correct ranged from 26.8% to 78.2% for all subjects, the ranges were actually much smaller when the two speakers were examined individually. Listeners who heard the male talker recognized between 26.8% and 56.6% of the targets correctly, whereas listeners who heard the female talker recognized between 53.9% and 78.2% of the target words correctly (excluding two outliers). This magnitude of difference in intelligibility across the two speakers is not uncommon in the literature comparing different speakers (e.g., Black, 1957; Bond and Moore, 1994; Hood and Poole, 1980; House *et al.*, 1965; Neel *et al.*, 1996). A comparison of our data to that of other investigators using an open-set recognition task and similar S/N ratios as used in the present study indicates that percent correct to our male speaker was somewhat below average whereas percent correct to our female speaker was somewhat above average (cf. Broadbent, 1967; Luce and Pisoni, 1998; Miller *et al.*, 1951; Sommers *et al.*, 1997).

We next performed a logistic multiple regression using the correctness of the listener's response as the dependent variable and the following independent variables: Speaker, the amplitude of the background babble that the word was embedded in (the noise level for short), the number of phonemes in the target word, the number of syllables in the target word, the number of lexical neighbors in the target word (i.e., the number of words in the HML that the target word could be turned into by the deletion, addition, or substitution of a single phoneme), the familiarity of the target (as obtained from the HML), and the logarithm of the frequency of occurrence of the target word (log word frequency for short). (Log word frequency in this document is calculated as $\log_{10}(f+1)$, where f is the number of occurrences per million words. The value of 1 was added to f because when f is 0, the logarithm is undefined and because frequency per million can be less than 1, resulting in a negative logarithm.) All responses, including non-words, were included in this analysis. Two such analyses were conducted, the two differing in terms of the source of word frequencies. The first used CELEX as the source. It found an overall R^2 of 0.310. All the independent variables contributed significant variance to predicting response correctness (all p's < 0.0001), with the exception of number of syllables in the target ($p = 0.698$). Familiarity of the target, log frequency of the target, and the number of phonemes in the target were all positively related to accuracy (e.g., proportion correct was higher for words with more phonemes). The noise level, and the number of lexical neighbors were negatively related to accuracy (e.g., proportion correct was higher for words with fewer lexical neighbors). Accuracy was also lower for the male speaker than for the female speaker, as previously noted.

Subsequent to the collection and initial analyses of the present data, the SUBTLEXus corpus of word frequencies became available (Brysbaert and New, 2009). Brysbaert and New showed that word frequencies based on SUBTLEXus outperformed CELEX word frequencies when predicting performance in a printed word lexical decision task. Accordingly, we repeated the preceding analysis using SUBTLEX as the source for log frequency of the target. (Whereas CELEX frequencies were available for all the target words, SUBTLEXus had no frequency data for 6.7% of the targets.) The pattern of results was the same as that using CELEX word frequencies. The overall R^2 was 0.308 (the same, for all practical purposes, as that obtained using CELEX frequencies). Again, all variables, with the exception of the number of syllables in the target ($p = 0.322$) predicted significant variance (all p's < 0.0001). Familiarity of the target, log frequency of the target, and the number of phonemes in the target were all positively related to

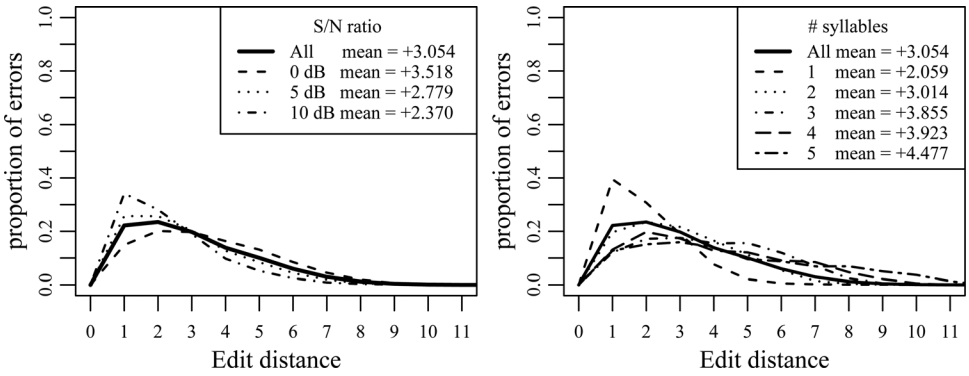

FIG. 1. Frequency distribution of the edit distance of incorrect responses relative to the target word as a function of signal-to-noise ratio (left panel) and of number of syllables in the target word (right panel).

accuracy, whereas the noise level, and the number of neighbors of the target were all negatively related. Performance was again poorer for the male talker.

Both the above analyses were repeated, using only word responses. The pattern of results did not change, with the exception that the number of syllables in the target now also predicted significant variance (both p's < 0.01) Words with more syllables were reported more accurately than words with fewer syllables.

CELEX log word frequencies for targets correlated relatively strongly with SUBTLEXus log frequencies, $r = 0.818$. When all listener responses were taken into account (errors as well as correct responses), the correlation was $r = 0.843$. Considering only errors classified as word responses, frequency counts were unavailable in SUBTLEXus approximately 1.75 times more often than in CELEX. For these reasons (the high correlation between CELEX and SUBTLEXus frequencies, and the larger number of words missing from the SUBTLEXus corpus), when analyzing errors, we only present results using the CELEX word frequencies.

B. Phonetic distance between targets and error responses

A potentially important characteristic of the error responses is their phonetic distance from the intended target word. We quantified this characteristic by using a straightforward simple measure referred to as *edit distance* (also called Levenshtein distance). Edit distance is defined as the number of edits—additions, deletions, and substitutions—needed to change one string of symbols into another. (In our case, the symbols are phonemes.) For example, the edit distance between *cat* and *cats* is 1 (add /s/); between *cat* and *cuts* is 2 (add /s/; change /æ/ to /ʌ/); and between *cat* and *its* is 3 (add /s/, change /æ/ to /ɪ/, delete /k/).

Figure 1 shows the frequency distribution of edit distances of the incorrect responses, as a function of S/N ratio (left panel) and number of syllables in the target word (right panel). As the left panel shows, the edit distance of errors increased as the S/N ratio became less favorable. In other words, as the masking noise increased, not only did listeners misperceive the target word more frequently, but the error responses were also phonetically less similar to the target word. As can be seen in the right panel, the syllable-based edit distance analysis indicated that longer words were more different from the target than were shorter words [$F(4, 30\,150) = 1259$, $p \approx 0$, $\eta^2 = 0.143$]. The same pattern occurs if length is measured in number of phonemes in the target word [$F(9, 30\,145) = 610$, $p \approx 0$, $\eta^2 = 0.154$]. This relationship is reversed, however, and the proportion of variance accounted for considerably reduced if edit distance is normalized by the length of the target word in terms of number of phonemes, as can be seen in the column second from the right in Table II for length measured in number of phonemes [$F(9, 30\,145) = 239$, $p \approx 0$, $\eta^2 = 0.067$], and in the second column from the right in Table III for length measured in number of syllables in the target word [$F(4, 30\,150) = 169$, $p < 0.00001$, $\eta^2 = 0.022$]. Note that if errors shared a constant proportion of phonemes with targets, then the normalized edit distance would be constant across targets of different lengths. The fact that it is not indicates that errors do not share either a constant number of phonemes with targets or a constant proportion of phonemes. On the other hand, length (measured by either the number of phonemes or by the number of syllables) accounts for a very small proportion of the variance in normalized edit distance, suggesting that the proportion of shared phonemes may be a reasonable starting point for determining a word's competitors.

Fewer than 23% of errors (6691 of 30 155) were lexical "neighbors" under the conventional definition of a neighbor having an edit distance of 1 (e.g., Luce and Pisoni, 1998)

TABLE II. Mean normalized edit distance between responses and targets and mean normalized difference in the number of phonemes in responses and targets as a function of the number of phonemes in the target word.

Length	N	ED/Length	Phonemes difference/Length
2	566	0.951413 (0.0264)	0.575972 (0.0242)
3	3263	0.65829 (0.0069)	0.192359 (0.0034)
4	4716	0.589695 (0.0046)	0.050679 (0.0007)
5	5949	0.551118 (0.0038)	−0.00303 (0.0001)
6	5575	0.538236 (0.0037)	−0.05127 (0.0007)
7	4285	0.518153 (0.0039)	−0.10572 (0.0016)
8	2976	0.513567 (0.0046)	−0.13999 (0.0026)
9	1739	0.442655 (0.0058)	−0.12657 (0.0030)
10	813	0.379705 (0.0084)	−0.10258 (0.0036)
11	273	0.418248 (0.0156)	−0.13919 (0.0084)

Length is the number of phonemes in the target word. N is the number of errors that occurred for the corresponding length. ED is the edit distance between the response and the target. Phon. Diff. is the number of phonemes in the response minus the number of phonemes in the target. Numbers in parentheses are standard errors of the means.

TABLE III. Mean normalized edit distance between responses and targets and mean normalized difference in the number of phonemes in responses and targets as a function of the number of syllables in the target word.

Syllables	*N*	ED/Length	Phonemes difference/Length
1	7199	0.610 (0.0043)	0.0139 (0.0040)
2	13 831	0.559 (0.0026)	−0.0134 (0.0023)
3	6410	0.527 (0.0033)	−0.0910 (0.0030)
4	2338	0.440 (0.0052)	−0.1114 (0.0045)
5	377	0.425 (0.0129)	−0.1103 (0.0102)

Syllables are the number of syllables in the target word. Length is the number of phonemes in the target word. *N* is the number of errors that occurred for the corresponding length. ED is the edit distance between the response and the target. Phon. Diff. is the number of phonemes in the response minus the number of phonemes in the target. Numbers in parentheses are standard errors of the means.

(See Luce *et al.*, 2000 for an alternative method of determining lexical neighbors). Although 40% of the incorrect responses for one-syllable words had an edit distance of 1 (2842 of 7199), less than 17% of incorrect responses for polysyllabic target words had an edit distance of 1 (3849 of 22 956). The mean normalized edit distance (edit distance/length in number of phonemes), however, was 0.609 for monosyllabic words and 0.536 for multisyllabic words [$t(30\,153) = 17.61$, $p < 0.0001$, Cohen's $d = 0.239$], suggesting that, in terms of proportion of shared phonemes, multisyllabic target words are in fact somewhat more similar to their errors than are monosyllabic targets.

A multiple regression analysis was carried out using length of the target in phonemes, the density of the target (as obtained from the HML) and the log word frequency of the error responses to a given target to predict the mean edit distance between a given target and its error responses. This analysis involved only errors that were words. The overall R^2 was 0.240 [$F(3, 21\,840) = 2304$, $p \approx 0$]. All three variables explained unique variance in the edit distance (all p's < 0.0001). Length of the target and log word frequency of the error response were positively related to edit distance; the density of the target was negatively related to the edit distance. Similar results occurred when length in number of phonemes was replaced with length in number of syllables [$R^2 = 0.231$, $F(3,21\,840) = 2186$, $p \approx 0$)] All three variables explained unique variance (all p's < 0.0001). Length and log word frequency of the error response were positively related to edit distance; density of the target was negatively related.

C. Phoneme-length difference

To address the possibility that the length of the target word helps to constrain the possible response set, we also analyzed our error corpus with respect to the difference in the number of phonemes between the response and the target word, specifically the number of phonemes in the response minus the number of phonemes in the target. We refer to this measure as the phoneme-length difference. A negative phoneme-length difference means that the response has fewer phonemes than the target; a positive phoneme-length difference means that the response has more phonemes. Figure 2 shows the distribution of the phoneme-length difference between targets and error responses as a function of S/N ratio (left panel) and as a function of number of syllables in the target word (right panel). Overall, incorrect responses tended to have roughly the same number of phonemes as the target word, although there was a tendency for responses to be shorter as indicated by the finding that all but one of the distributions has a negative mean. Collapsing across all syllable lengths, error responses with fewer phonemes than the target (36.6%) were slightly more common than error responses that had the same number of phonemes as the target (34.9%), which were more common than error responses that had more phonemes than the target word (28.5%). (Note that this general pattern did not hold for monosyllabic words, where 20.5% of the errors had fewer phonemes than the target, 44.7% had the same number, and 34.8% had more phonemes than the target.)

This small tendency for errors to be shorter than targets suggests that listeners have a slight response bias toward deleting phonemes as opposed to adding phonemes to their responses.

We further investigated the source of the bias toward shorter responses and determined that this finding is largely attributable to errors with two or more phoneme deletions because roughly the same percentage of errors had either one phoneme addition or deletion. Collapsing across all errors, 10 533 of the 30 155 error responses (34.9%) had the same number of phonemes as the target, 6067 (20.1%) of the error responses were one phoneme shorter than the target, whereas 6284 (20.8%) were one phoneme longer. However, 2625 (8.7%) were two phonemes shorter than the target, whereas only 1576 (5.2%) were two phonemes longer. This observed response bias to delete phonemes increased with word length; however, it remains true even for five-syllable

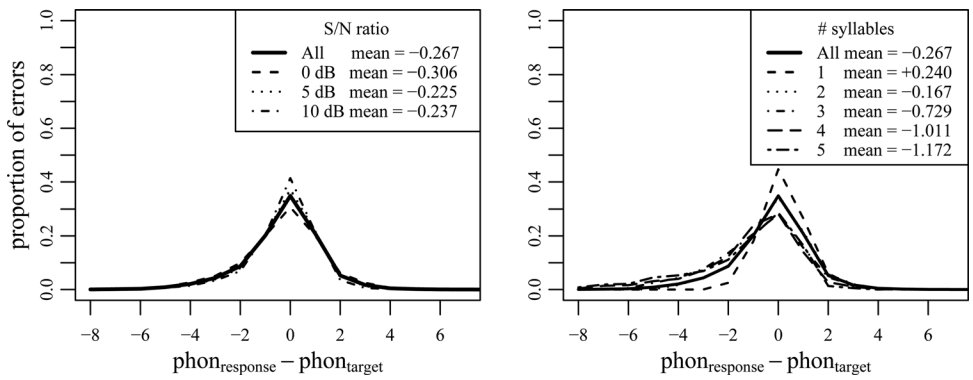

FIG. 2. Frequency distribution of the number of phonemes in incorrect responses minus the number of phonemes in the corresponding target response as a function of signal-to-noise ratio (left panel) and of number of syllables in the target word (right panel).

words. For these words, 76 of the 377 errors (20.2%) were one phoneme shorter than the target and 42 (11.4%) were two phonemes shorter, whereas 62 (16.4%) were one phoneme longer and only 5 (1.3%) were two phonemes longer. (We should perhaps note that the bias to delete phonemes as opposed to adding phonemes is not evident when only errors with an edit distance of 1 are examined. In fact there seems to be a bias to add phonemes in this case. Of the 6691 such errors, 4,076 (60.9%) involved a substitution, 1,147 (17.1%) involved a deletion, and 1468 (21.9%) involved an addition.)

The difference in number of phonemes between error responses and target words was highly consistent across all S/N ratios with a slight tendency for shorter responses as the noise level increased. For all S/N ratios, the phoneme-length difference frequency distribution peaks at 0, indicating that the most common errors have the same number of phonemes as the target word.

The finding that errors are more likely to have fewer phonemes than more phonemes than targets could simply reflect the structure of the English language. Suppose that when listeners incompletely decode the acoustic signal, they randomly choose a word from the lexicon consistent with whatever decoding of the acoustic signal they have accomplished. Although we have no particular reason to believe that such random sampling would produce errors shorter than the target, we also have no particular reason to believe that they would not. Accordingly, we performed a Monte-Carlo style simulation to determine whether observed errors were more likely to have fewer phonemes than the target than would be predicted by chance. To generate these chance estimates, we generated "pseudo-errors" for each of the word errors we obtained. The pseudo-errors were generated such that they differed from the target by the same edit distance as the actual error as we considered the edit distance to be the best available measure of the proportion of acoustic information correctly decoded by the listener. Thus we asked whether the errors we obtained were more likely to have fewer phonemes than the target than random errors differing from the target by the same edit distance. The following describes the algorithm we used in our simulation:

(1) For each word error, pick a random word from the CELEX database that has the same edit distance from the target word as the real error. (Note that the simulation only included errors that were words; non-word errors were not included. The question asked in the simulation is whether, given a signal with incomplete acoustic information, a randomly generated word error would tend to have more or fewer phonemes than the target. Because non-word errors cannot be randomly selected from the lexicon, they were not included in the simulation.) For example, one of the errors to the target word *zipper* was *scissor*, which has an edit distance of two from *zipper*. We constrained the algorithm to only choose from the set of words with the same edit distance as the actual error to account at least partially for the extent to which the listener correctly perceived the acoustic signal. In the case of *scissor*, this constrained the possible choices to 160 words, including words such as *bigger*, *liver* and so forth. This randomly selected error is referred to as a pseudo-error.
(2) Calculate the difference between the length of the target and the length of the pseudo-error in terms of number of phonemes For example, the difference in the number of phonemes between *zipper* and *scissor* is 0.
(3) Difference scores were computed for each of the pseudo-errors and their mean calculated.
(4) Steps 1-3 were repeated 10 000 times.

Because our simulation executed this algorithm 10 000 times, we obtained a range of possible values for the mean phoneme-length difference. Table IV shows the minimum and maximum mean phoneme-length difference obtained in the 10 000 simulation runs as well as the actual phoneme-length difference observed in our results for word errors. All the simulation runs produced a mean difference that was less negative (closer to 0) than the mean observed in the actual results. Thus our simulations revealed that the probability of achieving the actual results under the assumption that error word length was randomly related to target word length, to be less than 1 in 10 000. Thus the results from the simulation indicate that listeners were selectively biased and more likely than predicted by chance to generate word-error responses with fewer phonemes than the target word.

Do listeners perhaps delete a constant proportion of phonemes? If so, then the function relating the phoneme-length difference as a function of target length would be flat once that difference is normalized by target length. To investigate this possibility, we first divided the phoneme length by the length of the target as measured by the number of phonemes. We then examined how this normalized difference changed as a function of the number of phonemes in the target (see the rightmost column in Table II) and as a function of the number of syllables in the target (see the rightmost column in Table III). In both cases, the tendency to delete phonemes clearly increases as the length of the target word increases [$F(8, 29\,580) = 528$, $p \approx 0$, $\eta^2 = 0.125$ for target length measured in number of phonemes (This analysis did not include targets with just two phonemes as there are few opportunities to delete a single phoneme from a two-phoneme word and still have a word.); $F(4, 30\,150) = 530$, $p \approx 0$, $\eta^2 = 0.066$ for length measured in number of syllables]. Note, however, that

TABLE IV. Monte-Carlo simulation results for the mean difference in the number of phonemes, in the number of syllables, and in the log frequency of occurrence of errors and targets.

Monte-Carlo results			
	Simulation		
	Minimum	Maximum	Observed
Mean phoneme difference	−0.2455	−0.1964	−0.317*
Mean syllable difference	−0.1807	−0.1591	−0.191*
Mean frequency difference	−0.1197	−0.0842	0.562*

All measures are the mean for the errors minus the mean for the targets. Min is the smallest mean difference that occurred across all the simulation runs; Max is the largest such mean difference. Observed is the mean difference observed in the data for word errors (non-word errors were not included).
*$p < 0.0001$

the function does become considerably flatter once length reaches a value of approximately seven or eight phonemes, as can be seen in Table III, suggesting that once length reaches some critical value, listeners do in fact have a tendency to delete a constant proportion of phonemes.

D. Syllable-length difference

Error analyses parallel to those described above for word length in terms of the number of phonemes were also performed for length in terms of the number of syllables in the error response and the target word. Figure 3 displays the syllable-length difference, i.e., the number of syllables in the error response minus the number of syllables in the target, as a function of S/N ratio (left panel) and of the number of syllables in the target (right panel). Similar to the phoneme-length difference analysis reported in the previous section, responses largely had the same number of syllables as the target word with a slight bias toward deleting syllables over adding syllables. The number of errors with the same number of syllables as the target was 74% (22 309 of 30 155). Only 1744 (5.8%) errors were one syllable longer than the target, while 4942 (16.4%) were one syllable shorter. One hundred thirty-six (0.5%) of the errors had two additional syllables, while 865 (2.9%) of the errors had two fewer syllables. The tendency to delete syllables increased with word length, although even for five syllable words most responses had the same number of syllables as the target word. This tendency was apparent even when the syllable-length difference was normalized by the number of syllables in the target, although the amount of variance explained was low. The normalized differences were −0.043, −0.139, −0.153, and −0.177 for syllable lengths of two through five, respectively [$F(3,\ 22\,952) = 321$, $p \approx 0$, $\eta^2 = 0.04$]. The syllable-length difference was also highly consistent across all S/N ratios, with a slight tendency for listeners to respond with shorter words as the S/N ratio decreased. This pattern of results indicates that listeners are very unlikely to add or delete syllables when misperceiving a spoken word. When they do add or delete syllables, deletions are more common than additions.

We conducted a second simulation, similar to the one described in the previous section, comparing the difference in the number of syllables between pseudo-errors and the target. Again, only word errors were included in the simulation. This simulation was done to assess whether the patterns observed in Fig. 3 simply reflect the pattern expected by chance given the observed distribution of edit distances between errors and target words. Table IV shows the minimum and maximum mean syllable-length difference obtained in the 10 000 simulation runs, as well as the actual syllable-length difference observed in our results for word errors. As was the case for the difference in number of phonemes, all the simulation runs produced a mean syllable-length difference that was less negative (closer to 0) than the mean observed in the actual results ($p < 0.0001$). Thus, the results from the simulation indicate that listeners were selectively biased and more likely than predicted by chance to generate word-error responses with fewer syllables than the target word.

E. Lexical frequency

We also assessed the differences in frequency between the response and target and then performed simulations to determine whether those differences were greater than we would expect by chance, as this difference can be relevant to discriminating different models of the frequency effect (e.g., Broadbent, 1967).

When analyzing the effects of lexical frequency, as explained earlier, we used a logarithmic transformation of the raw CELEX frequency, $\log_{10}(f+1)$, where f is the number of occurrences per million words in the CELEX corpus.

Figure 4 shows the frequency distribution of errors as a function of the log frequency of the error response minus the log frequency of the target, as a function of S/N ratio in the left panel, and of the number of syllables in the target in the right panel. These distributions peak at a value of +1 and all have positive means, indicating that error responses were generally higher in frequency than target words. This result was consistent across both S/N ratios and word lengths. (Such a result could simply reflect the fact that errors tend to be shorter than targets, and shorter words are of higher frequency. Alternatively, the direction of causation may be reversed—errors are shorter than targets because errors tend to be of higher frequency than the target.) In addition, considering only errors that were classified as word, there was a moderate correlation between the log frequency of the target words and the log frequency of incorrect responses [$r = 0.154$, $t(21\,842) = 26.58$, $p < 0.0001$], a result that

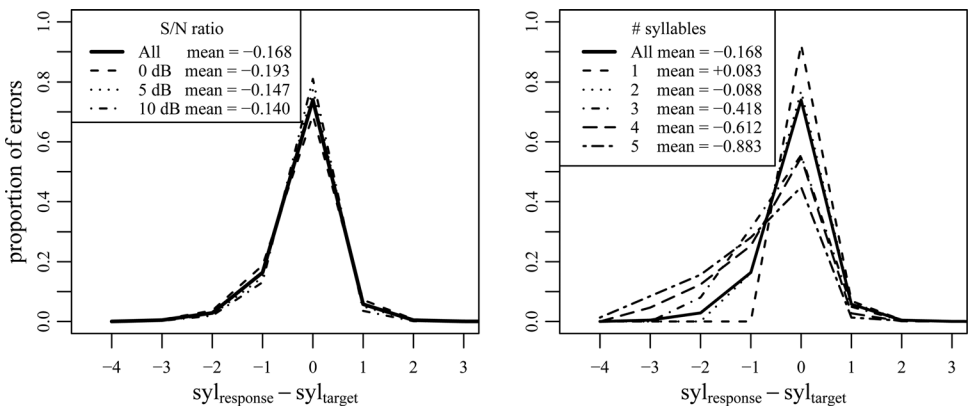

FIG. 3. Frequency distribution of the number of syllables in incorrect responses minus the number of syllables in the corresponding target responses as a function of signal-to-noise ratio (left panel) and of number of syllables in the target word (right panel).

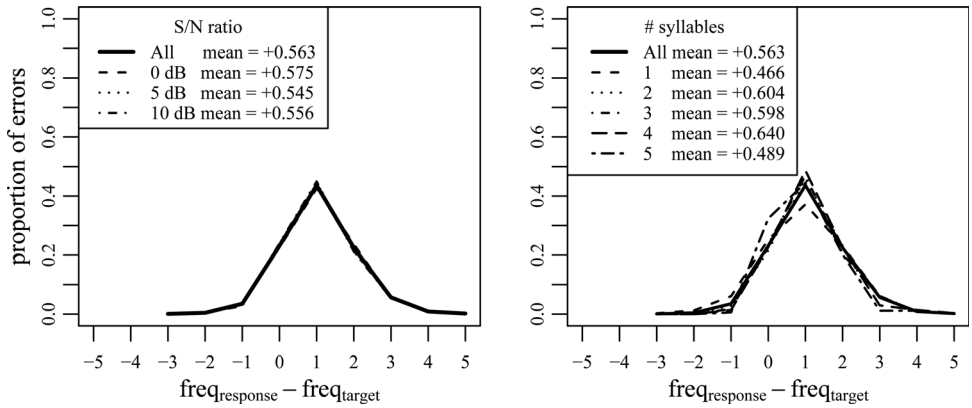

FIG. 4. Frequency distribution of the log frequency of occurrence in error responses minus the log frequency of occurrence in the corresponding target responses as a function of signal-to-noise ratio (left panel) and of number of syllables in the target word (right panel). See the text for an explanation of how log frequency of occurrence was calculated. In the left panel, the distributions lie virtually on top of one another, making it extremely difficult to see the individual curves for the different S/N ratios.

conflicts with the earlier findings reported by Pollack *et al.* (1960), who found no correlation. The discrepancy with the findings of Pollack *et al.* is partially due to the fact that the present study included words ranging in length from one to five syllables. Given that shorter words are higher in frequency than longer words and that errors to short words tend to be short and errors to long words tend to be long, this positive correlation naturally emerged; Pollack *et al.* (1960), on the other hand, used only monosyllabic words in their study. Hence this factor would not have played a role in their study. However, even when we considered only monosyllabic target words in the present study, a weak correlation was still observed between the log frequency of target and response [$r = 0.108$, $t(6546) = 17.56$, $p < 0.0001$]. A possible reason for the discrepancy between our results and those of Pollack *et al.* is that in Pollack *et al.*, all words were drawn from a closed-set of items that were used repeatedly. The methodology used by Pollack *et al.* may have increased the proportion of words coming from that closed response set and thereby masked frequency effects in error responses, thereby eliminating the small correlation that we observed for monosyllabic words. It is also the case that our sample size is much larger than that of Pollack *et al.*, making it more likely that we would uncover any statistical relation that did exist between the log word frequency of errors and targets.

We also performed a simulation, like those described in the preceding text and involving word errors only, to assess whether the pattern in Fig. 4 indicating that responses have a higher frequency than their target would be obtained by chance given random errors of the same edit distance from the target. The results are summarized in Table IV. All the mean frequency differences generated by the simulation were slightly negative. In contrast, the mean frequency difference observed in the data was moderately positive. Hence listeners were more likely than chance to select error responses that were higher in frequency than the target words.

F. Word vs non-word responses

The phoneme- and syllable-level analyses treated all incorrect responses equally. However, it is also informative to distinguish between real word responses and non-word responses. As indicated earlier, we divided the incorrect responses into five broad categories; the proportion of responses per category is shown in Table I. The foreign, multiple, and neologism categories, which comprised only 250 responses, or less than 1%, of the total errors, have been grouped into the non-word category for the remaining analyses.

Figure 5 shows the edit distance distribution separately for word and non-word responses broken down by the number of syllables in the target word. The mean edit distance for word and non-word responses did not differ substantially (3.073 vs 3.002 with standard deviations of 1.824 and 1.686, respectively). Because of the large number of observations, this effect is statistically significant [$t(30153) = 3.083$, $p < 0.01$]. The size of the effect, however, is relatively small (Cohen's $d = 0.040$). There were, though, some notable differences in the edit distance distributions as a function of target word length. For monosyllabic words, non-word responses had a higher edit distance than word responses, whereas for multi-syllabic target words the opposite pattern was obtained. We also analyzed the difference in number of

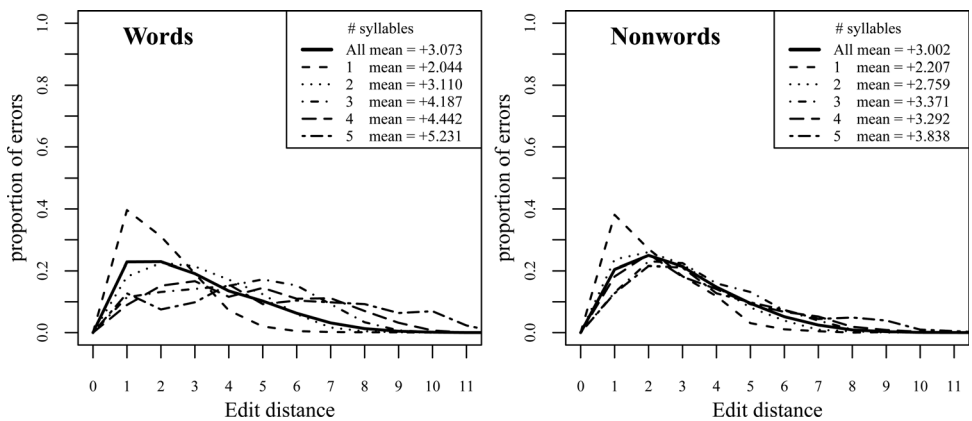

FIG. 5. Frequency distribution of the edit distance of incorrect responses relative to the target word for words (left panel) and non-words (right panel) as a function of number of syllables in the target word.

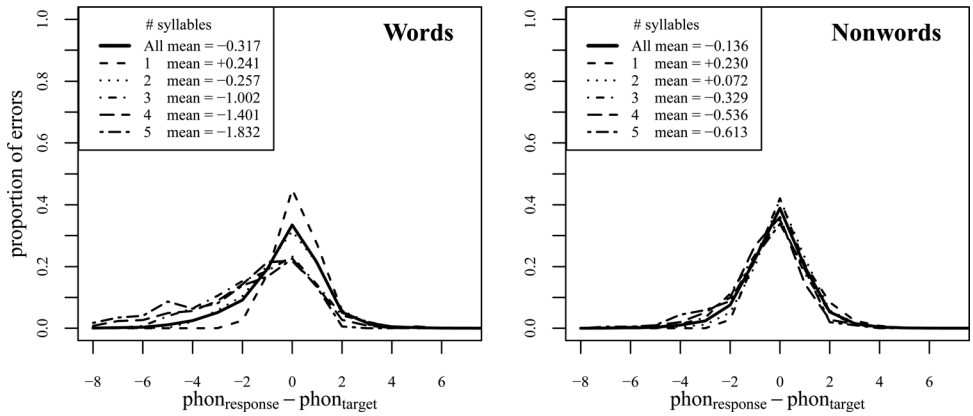

FIG. 6. Frequency distribution of the number of phonemes in incorrect responses minus the number of phonemes in the target word for words (left panel) and non-words (right panel) as a function of number of syllables in the target word.

phonemes separately for word and non-word responses. These results are shown in Fig. 6. In terms of the number of phonemes, word error responses differed more from the target word than non-word responses at all word lengths. This result suggests that when listeners made non-word responses, the errors reflected responses based on bottom-up, sensory-based processes and hence shared the overall acoustic-phonetic structure of the target word. For instance, suppose that the listener misperceives a single phoneme but that misperception forms a non-word. A listener relying only on bottom-up processing would report that non-word as the heard stimulus. A listener who was also using top-down knowledge of the lexicon might remove an additional phoneme to ensure that a word was reported as the response but also resulting in the word error having fewer phonemes than the target.

As shown in Fig. 7, the difference between errors and targets in terms of the number of syllables for word and non-word errors was consistent with the results for the difference in edit distance and with the results for the difference in number of phonemes. Listeners' non-word responses tended to be phonetically more similar to the target word (i.e., had smaller edit distances) than were word responses and were also closer to the target word in terms of the number of phonemes and, with the exception of monosyllabic words, the number of syllables than were word responses.

IV. DISCUSSION

The present study examined a subset of the formal and lexical characteristics of incorrect responses in an open-set spoken word recognition paradigm. In contrast to earlier studies reported in the literature over the years, many of which used only monosyllabic words, our set of target words was a random sample of American English. Some of the earlier results obtained from the study of the recognition of monosyllabic words in noise generalized to our broader sample; others did not.

Consistent with previous studies such as Wiener and Miller (1946), we found that longer words (where word length is measured either in terms of number of phonemes or number of syllables) were perceived more accurately than shorter words. We also found that errors tended to be slightly higher in lexical frequency than the target words and that the lexical frequency of errors and targets was significantly correlated in contrast with the earlier findings of Pollack *et al.* (1960), who used only monosyllabic words in a paradigm that was a hybrid of open and closed-set. In addition, our results indicate that as word length increased, errors became increasingly different from the targets (as estimated by edit distance) but that the difference in the number of phonemes and syllables between the response and the target only increased slightly. The relationship between edit distance and length of the target reversed when edit distance was normalized by target length (in number of phonemes): Normalized edit distance between errors and targets decreased with increasing length of the target (as measured by either number of phonemes or number of syllables). Finally, we observed that listeners were more likely to generate non-word responses as S/N ratio decreased and as target word frequency decreased.

As noted in Sec. I, the intent of the present study was to make available a corpus of spoken word recognition errors

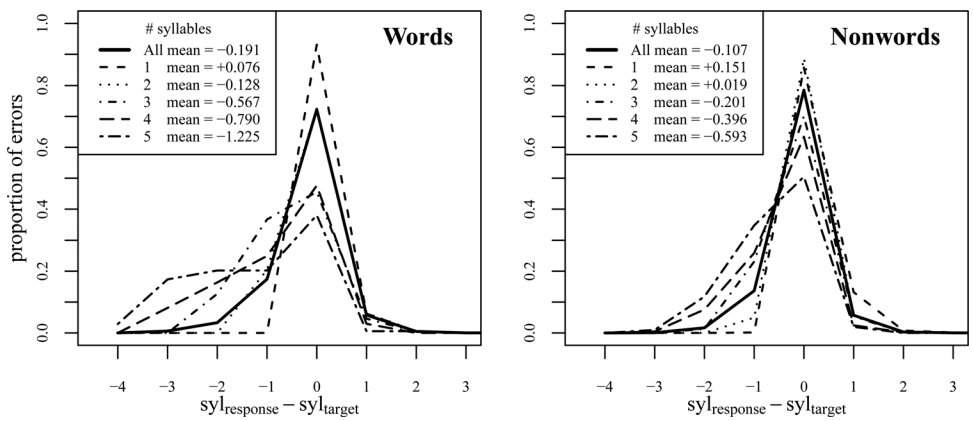

FIG. 7. Frequency distribution of the number of syllables in incorrect responses minus the number of syllables in the target word for words (left panel) and non-words (right panel) as a function of number of syllables in the target word.

rather than to serve as a critical test of various models of spoken word recognition. Nevertheless, our results do address several aspects of such models. For instance, Pollack *et al.* (1960) viewed word recognition as akin to choosing balls randomly from an urn in which each ball represents a word token. The number of balls corresponding to a given word type is determined by that word's frequency of occurrence in the language. That is, there will be 779 balls representing *cat* out of a total of 17.9×10^6 balls (the total word count in the COBUILD corpus used by the CELEX database). In this model of spoken word recognition, acoustic input effectively decreases the total number of words types in the urn. Only balls representing word types that are reasonable acoustic matches for the input are included in the final decision set. Pollack *et al.* (1960) argued that this model predicts that word frequency of the incorrect responses will be independent of the target word frequency, an important prediction that was supported by the results that Pollack *et al.* reported in 1960 but not by the present results. As previously explained, we believe that this discrepancy stems primarily from methodological factors between the two studies. In addition, our finding that 74% of incorrect responses had the same number of syllables as the target word (in contrast, only 35% had the same number of phonemes) suggests that the initial narrowing down of lexical candidates may very well start on a more global, syllabic basis and then become narrowed down and refined further based on a phonological and perhaps even a feature analysis. Such a pattern suggests (though by no means proves) that the candidate pool of words is not constructed in a completely left-to-right, linear fashion using sequences of phonemes or phonetic segments as assumed in some contemporary models of spoken word recognition (e.g., Marslen-Wilson and Tyler, 1980; Marslen-Wilson and Zwitserlood, 1989). Instead listeners may first extract more salient information from the acoustic signal, such as the number of syllables, and use that information to help construct the pool of candidate targets. Indeed, over the years, a number of proposals have been made for the syllable as the primary unit of speech perception (e.g., Goldinger, 2003; Savin and Bever, 1970). More recently, Poeppel (2003) has reported neurolinguistic evidence suggesting that listeners monitor for both phonemes and syllables simultaneously.

In terms of activation-plus-competition models of human speech perception (Elman and McClelland, 1986; Goldinger *et al.*, 1989; Luce and Pisoni, 1998; McClelland and Elman, 1986; Norris, 1994; Norris *et al.*, 2000), if we assume that error responses indicate what words compete with the target word, then our study can be viewed as an initial attempt to identify what factors determine the extent to which two words compete with one another for recognition, i.e., which words are neighbors of one another. Most, if not all, researchers working in spoken word recognition have recognized that the single phoneme deletion, addition, and substitution (DAS) definition of a lexical neighbor is an overly simplistic similarity metric. Nevertheless, the DAS rule does remarkably well in predicting what words will be hard to identify and what words will be easy to identify when those words are restricted to monosyllabic words and especially CVCs (e.g., Luce and Pisoni, 1998; Pisoni *et al.*, 1985). Obviously, edit distances beyond 1 need to be considered when defining lexical neighborhoods of longer, more complex spoken words. However, our results indicate that the nature of the edit also appears to make a difference. Words competing with a given target tend to have the same number of or fewer phonemes than the target. Words with more phonemes than the target tend to be weaker competitors. Words involving syllable substitutions are more likely to compete with a target word than are words involving syllable deletions or additions. Errors were also more likely to result in deleted syllables than in added syllables, suggesting that words with deleted syllables are stronger competitors for a given target than are words with added syllables.

The present study used words spoken in isolation. Would similar results occur with connected speech and, most important, in everyday conversational speech? As noted earlier, Bond and her colleagues (Bond, 1999a, 2005; Bond and Garnes, 1980; Bond and Robey, 1983; Garnes and Bond, 1980) have collected a corpus of slips of the ear or misperceptions of speech occurring in everyday conversations. Many of the errors in that corpus occurred across multiple words in an utterance (e.g., "I am going up for my office hours" is misheard as "I am going up for my vodka sours."). In such cases, our results cannot be directly compared to her findings. In everyday speech, once an error occurs, top-down, knowledge-based semantic factors may cause additional errors. In the preceding example, once "hours" is misheard as "sours," "office" may be misheard as "vodka" simply to maintain semantic coherence and keep the perceived utterance halfway sensible. In our case, because isolated words were used, there does not seem to be a place for such semantically determined errors.

Where comparisons can be made, however, our results and Bond's findings (1999b) parallel one another. For simple errors, errors involving a single phonetic segment (i.e., errors with an edit distance of 1), both Bond (1999b) and Labov (2010) found that substitutions were more common than deletion and addition errors as we also observed in our analysis of the difference in the number of phonemes in error responses and targets for this class of errors. For an edit distance of 1, our results showed that 59.4% of the errors were substitutions. We also found that additions (21.9%) were somewhat more common than deletions (18.6%) for errors with an edit distance of 1. In contrast, for single segment errors, Bond found approximately the same number of consonant deletions (29) as consonant additions (28). These results, though, involve a very small number of observations and hence may not be entirely reliable. In a laboratory study involving spoken Harvard sentences, Bond *et al.* (1996) found that single-segment substitutions were more common than deletions. Deletions, in turn, were more common than additions as was also the case in our data. Like us, Bond (1999b) observed relatively few syllable additions or deletions—the incorrectly perceived word tended to have the same number of syllables as the target. Bond *et al.* (1996) observed a similar pattern of errors in their study using Harvard sentences. Bond *et al.* also observed a higher

absolute number of syllable deletions than syllable additions as did we. In comparing her results for errors in everyday conversations to those of the laboratory study of Bond *et al*., Bond concluded that the similarities in results greatly outweighed the differences. Likewise, at least at a coarse level of analysis, our results with isolated words in the laboratory parallel those of Bond *et al*. for sentences and those of Bond for everyday speech, suggesting that our results have at least some ecological validity. Similarly, Vitevitch (2002) observed that the errors in Bond's corpus had higher neighborhood densities and neighborhood frequencies than expected by chance as would be predicted based on laboratory studies of speech perception. The parallels observed here and in the study of Vitevitch between laboratory results and naturally occurring slips of the ear are encouraging to those wishing to study the errors made in perceiving speech because errors are much easier to elicit and record in the laboratory than in everyday conversations (cf. Bond, 1999b).

In summary, we examined the errors people made when recognizing isolated spoken words drawn from a random sample of American English words. The characteristics of those errors paralleled the characteristics of errors made when recognizing words in everyday speech. Such a finding suggests that the processes involved in laboratory studies of spoken word recognition are also involved in recognizing spoken words in everyday conversations. Analyses of the errors can also, in terms of activation-plus-competition models of spoken word recognition, indicate what words compete with one another for recognition. For example, our findings indicate that competing words tend to have the same number of syllables as the target word. Words with fewer phonemes than a given target word appear to be stronger competitors of that target than do words with more phonemes—errors with fewer phonemes than the target were more frequent than errors with more phonemes. Our finding that the edit distance between a target word and errors made when perceiving that target increases with the length of the target suggests that it is not the *absolute number* of phonological contrasts or segmental differences between two words that determines the extent to which they compete with one another. Rather the *proportion* of phonological differences between two words may be a better starting point for determining the extent to which they compete with one another for recognition. Our finding, though, that edit distance normalized by word length is negatively related to word length indicates that additional variables will also need to be taken into account. The extent to which each of these effects of competition reflects bottom-up perceptual processes and top-down, post-perceptual, decision processes awaits further research using words with different lengths and syllable structures rather than relying upon a small number of monosyllabic words as researchers have done in the past.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grant No. DC00111-34. The first three authors contributed equally to this paper.

[1]See supplementary material at http://dx.doi.org/10.1121/1.4809540 for a list of the stimulus words used in the experiment and listeners' responses to each stimulus presentation.

Baayen, H. R., Piepenbrock, R., and van Rijn, H. (**1993**). *The CELEX Lexical Database* (CD-ROM), (Linguistics Data Consortium, University of Pennsylvania, Philadelphia).

Bates, T. C., and D'Oliveiro, L. (**2003**). "PSYSCRIPT: A Macintosh application for scripting experiments," Behav. Res. Methods, Instrum. Comput. **35**, 565–576.

Benkí, J. R. (**2003**). "Analysis of English nonsense syllable recognition in noise," Phonetica **60**, 129–157.

Black, J. W. (**1957**). "Multiple-choice tests of intelligibility," J. Speech Hear. Disord. **22**, 213–235.

Bond, Z. S. (**1999a**). "Morphological errors in casual conversation," Brain Lang. **68**, 144–150.

Bond, Z. S. (**1999b**). *Slips of the Ear: Errors in the Perception of Casual Conversation* (Academic, New York), pp. 1–212.

Bond, Z. S. (**2005**). "Slips of the ear," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishing, Malden, MA), pp. 290–310.

Bond, Z. S., and Garnes, S. (**1980**). "Misperceptions of fluent speech," in *Perception and Production of Fluent Speech*, edited by R. A. Cole (Erlbaum, Hillsdale, NJ), pp. 115–132.

Bond, Z. S., and Moore, T. J. (**1994**). "A note on the acoustic-phonetic characteristics of inadvertently clear speech," Speech Commun. **14**, 325–337.

Bond, Z. S., Moore, T. J., and Gable, B. (**1996**). "Listening in a second language," *Paper Presented at the Fourth International Conference on Spoken Language Processing*, Philadelphia.

Bond, Z. S., and Robey, R. R. (**1983**). "The phonetic structure of errors in the perception of fluent speech," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), pp. 249–283.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Broadbent, D. E. (**1967**). "Word-frequency effect and response bias," Psychol. Rev. **74**, 1–15.

Browman, C. P. (**1980**). "Perceptual processing: Evidence from slips of the ear," in *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*, edited by V. A. Fromkin (Academic, New York), pp. 213–230.

Brysbaert, M., and New, B. (**2009**). "Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," Behav. Res. Methods **41**, 977–990.

Carnegie Mellon University (**2007**). *The Carnegie Mellon University Pronouncing Dictionary,* available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Last viewed January 19, 2010).

Clopper, C. G., Pisoni, D. B., and Tierney, A. T. (**2006**). "Effects of open-set and closed-set task demands on spoken word recognition," J. Am. Acad. Audiol. **17**, 331–349.

Cox, R. M., Alexander, G. C., and Gilmore, C. (**1987**). "Development of the Connected Speech Test (CST)," Ear Hear. **8**, 119S–126S.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," J. Acoust. Soc. Am. **116**, 3668–3678.

Elman, J. L., and McClelland, J. L. (**1986**). "Exploiting lawful variability in the speech waveform," in *Invariance and Variability in Speech Processing*, edited by D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 360–385.

Garnes, S., and Bond, Z. S. (**1980**). "A slip of the ear: A snip of the ear? A slip of the year"? in *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*, edited by V. A. Fromkin (Academic, New York), pp. 231–239.

Gerstman, L. J., and Bricker, P. D. (**1960**). "Word frequency effects in learning unknown messages sets," J. Acoust. Soc. Am. **32**, 1078–1079.

Goldinger, S. D. (**2003**). "Puzzle-solving science: The quixotic quest for units in speech perception," J. Phonetics **31**, 305–320.

Goldinger, S. D., Luce, P. A., and Pisoni, D. B. (**1989**). "Priming lexical neighbors of spoken words: Effects of competition and inhibition," J. Mem. Lang. **28**, 501–518.

Hood, J. D., and Poole, J. P. (**1980**). "Influence of the speaker and other factors affecting speech intelligibility," Audiology **19**, 434–455.

House, A. A., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (**1965**). "Articulation-testing methods: Consonantal differentiation with a closed set response," J. Acoust. Soc. Am. **37**, 158–166.
Kilgariff, A. (**2007**). "Googleology is bad science," Comput. Ling. **33**, 147–151.
Labov, G. (**2010**). *Principles of Linguistic Change: Cognitive and Cultural Factors* (Wiley-Blackwell, Malden, MA), Vol. 3, pp. 21–48.
Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., and Vitevitch, M. S. (**2000**). "Phonetic priming, neighborhood activation, and PARSYN," Percept. Psychophys. **62**, 615–625.
Luce, P. A., and Pisoni, D. B. (**1998**). "Recognzing spoken words: The neighborhood activation model," Ear Hear. **19**, 1–36.
Marslen-Wilson, W. D., and Tyler, L. K. (**1980**). "The temporal structure of spoken language understanding," Cognition **8**, 1–71.
Marslen-Wilson, W. D., and Zwitserlood, P. (**1989**). "Accessing spoken words: The importance of word onsets," J. Exp. Psychol. Hum. Percept. Perform. **15**, 576–585.
McClelland, J. L., and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cognit. Psychol. **18**, 1–86.
Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. **41**, 329–335.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Neel, A. T., Bradlow, A. R., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. II: Analysis of transcription errors," Research on Spoken Language Processing: Progress Report No. 21 (Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN), pp. 421–437.
Norris, D. (**1994**). "Shortlist: A connectionist model of continuous speech recognition," Cognition **52**, 189–234.
Norris, D., McQueen, J. M., and Cutler, A. (**2000**). "Merging information in speech recognition: Feedback is never necessary," Behav. Brain Sci. **23**, 299–370.
Nusbaum, H. C., Pisoni, D. B., and Davis, C. K. (**1984**). "Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words," Research on Speech Perception Progress Report No. 10 (Speech Research Laboratory, Psychology Department, Indiana University, Bloomington, IN), pp. 357–376.
Pickett, J. M. (**1957**). "Perception of vowels heard in noises of various spectra," J. Acoust. Soc. Am. **29**, 613–620.
Pisoni, D. B., Nusbaum, H. C., Luce, P. A., and Slowiaczek, L. M. (**1985**). "Speech perception, word recognition and the structure of the lexicon," Speech Commun. **4**, 75–95.
Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as asymmetric sampling in time," Speech Commun. **41**, 245–255.
Pollack, I., Rubenstein, H., and Decker, L. (**1959**). "Intelligibility of known and unknown message sets," J. Acoust. Soc. Am. **31**, 273–279.
Pollack, I., Rubenstein, H., and Decker, L. (**1960**). "Analysis of incorrect responses to an unknown message set," J. Acoust. Soc. Am. **32**, 454–457.
Savin, H. B., and Bever, T. G. (**1970**). "The nonperceptual reality of the phoneme," J. Verbal Learn. Verbal Behav. **9**, 295–302.
Sommers, M. M., Kirk, K. I., and Pisoni, D. B. (**1997**). "Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format," Ear Hear. **18**, 89–99.
Vitevitch, M. S. (**2002**). "Naturalistic and experimental analyses of word frequency and neighborhood density effect in slips of the ear," Lang. Speech **45**, 407–434.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Wiener, F., and Miller, G. (**1946**). "Some characteristics of human speech," Trans. Recept. Sounds Combat Cond. **3**, 58–68.