Sentence perception in noise by hearing-aid users predicted by syllable-constituent perception and the use of context

James D. Miller,^{1,a)} Charles S. Watson,¹ Marjorie R. Leek,² David J. Wark,³ Pamela E. Souza,⁴ Sandra Gordon-Salant,^{5,b)} Jayne B. Ahlstrom,⁶ and Judy R. Dubno⁶

¹Communication Disorders Technology, Incorporated, Bloomington, Indiana 47408, USA

²VA Loma Linda Healthcare System, Loma Linda, California 92357, USA

³Communication Sciences and Disorders, University of Memphis, Memphis, Tennessee 38105, USA

⁴Communication Sciences and Disorders, Northwestern University, Evanston, Illinois 60208, USA

⁵Hearing and Speech Sciences, University of Maryland, College Park, Maryland 20742, USA

⁶Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, Charleston, South Carolina 29425, USA

ABSTRACT:

Masked sentence perception by hearing-aid users is strongly correlated with three variables: (1) the ability to hear phonetic details as estimated by the identification of syllable constituents in quiet or in noise; (2) the ability to use situational context that is extrinsic to the speech signal; and (3) the ability to use inherent context provided by the speech signal itself. This approach is called “the syllable-constituent, contextual theory of speech perception” and is supported by the performance of 57 hearing-aid users in the identification of 109 syllable constituents presented in a background of 12-talker babble and the identification of words in naturally spoken sentences presented in the same babble. A simple mathematical model, inspired in large part by Boothroyd and Nittrouer [(1988). *J. Acoust. Soc. Am.* **84**, 101–114] and Fletcher [Allen (1996) *J. Acoust. Soc. Am.* **99**, 1825–1834], predicts sentence perception from listeners’ abilities to recognize isolated syllable constituents and to benefit from context. When the identification accuracy of syllable constituents is greater than about 55%, individual differences in context utilization play a minor role in determining the sentence scores. As syllable-constituent scores fall below 55%, individual differences in context utilization play an increasingly greater role in determining sentence scores. Implications for hearing-aid design goals and fitting procedures are discussed. © 2020 Acoustical Society of America.

<https://doi.org/10.1121/10.0000563>

(Received 8 May 2018; revised 4 December 2019; accepted 12 December 2019; published online 28 January 2020)

[Editor: Sarah Hargus Ferguson]

Pages: 273–284

I. INTRODUCTION

A. General

This paper focuses on the relation between syllable-constituent intelligibility and the perception of naturally spoken sentences. Previously, Miller *et al.* (2015) presented the view that sentence intelligibility involves three skills: (1) the ability to perceive phonetic details; (2) the ability to use situational context that is extrinsic to the speech signal; and (3) the ability to use the context that is inherent in the speech signal. This view is explained and given a quantitative formulation in this paper. This approach is described as the syllable-constituent, contextual theory of speech perception and it is applied to data describing the perception of sentences by hearing-aid users. The percentage of correct identifications of 109 syllable constituents (Constituent PC) as described in Miller *et al.* (2017) is a direct measure of the first ability described above. Measures of abilities (2) and (3) above are estimated using the relationships between

sentence scores and syllable-constituent scores obtained at differing signal-to-noise ratios (SNRs). Furthermore, a method to estimate the combined use of the two kinds of context (that is, the overall context utilization) is provided. Data obtained from 57 hearing-aid users are presented that are consistent with and provide preliminary support for the hypothesis that sentence perception can be accounted for by syllable-constituent perception and the ability to use the two types of context mentioned above.

1. Syllable constituents

According to the constituent analysis of syllables, a syllable is considered to have three parts: an onset, a nucleus and a coda or ending (for a general introduction to constituent analysis, see Hall, 2006). The acoustic correlates of syllable constituents appear to be more stable than those of phonemes or features, which vary more with context and syllable position. Because syllable constituents are restricted in their sequential positions, they have a more limited range of phonetic contexts than do phonemes or features. Contributing to this simplification is that each onset, each nucleus, and each

^{a)}Electronic mail: jamdmill@indiana.edu, ORCID: 0000-0002-2297-7870.

^{b)}ORCID: 0000-0001-8037-9724.

coda is treated as a unit, no matter how many phones (speech sounds) it contains.

2. Situational and speech-inherent context

Situational context is given by the situation in which the speech communication occurs and is extrinsic to the speech signal. A common form of situational context is provided by the listener's expectation of what is likely to be said in any particular situation. For example, when a server at a restaurant provides a customer with a menu, she/he might say "Can I get you something to drink?" An aware customer is unlikely to infer that the server said, "Can I get something from the sink?" Other forms of situational context may include non-phonetic gestures such as pantomime and pointing. In contrast, speech-inherent context is carried by the speech signal itself, that is, by the redundancies inherent within a spoken language. Coarticulatory effects, in which the expression of a phoneme is systematically modified by the phonemes that precede or follow it, are examples of speech-inherent context. Other forms of speech-inherent context are provided by allophonic changes in phonemes that indicate syllabic structure and by semantic and grammatical constraints that aid in organization of the speech stream into words, phrases, and meanings.

3. How syllable-constituent perception and context utilization combine to inform speech perception generally

A simple mathematical model, the syllable-constituent, contextual model of speech perception, is given in Sec. II E below. This model, inspired in large part by the work of Boothroyd and Nittrouer (1988) and of Fletcher [see Allen (1996)], describes how syllable-constituent perception combines with context utilization to result in observed speech-perception scores.

II. METHODS

A. General

The general characteristics of a speech perception training project that led to the collection of the data reported here are described in Miller *et al.* (2015). The experimental design and procedures of that project were formulated to determine the efficacy of speech perception training as a means of improving speech perception by hearing-aid users. Because that training project included syllable constituents and meaningful sentences in various levels of speech-babble background, the resulting data provided an opportunity to study systematic relations between the perception of those two classes of stimuli. This paper does not include effects of speech perception training, as the data presented here were collected prior to intensive training. Rather it describes relations between syllable-constituent perception and sentence perception found after initial training, sufficient only to familiarize the participants with the tasks. The data are for 57 of the 59 hearing-aid users who were tested

with the Speech Perception and Assessment Training System (SPATS) as described by Miller *et al.* (2015) and Miller *et al.* (2017). Details of the SPATS testing procedures for constituents and sentences are presented in Secs. II C and II D, respectively. All tests were conducted with the listeners wearing their personal hearing aids with customary use settings. Tests were administered in sound-treated booths or in quiet rooms with stimuli presented at an overall level of 65 dBC via a loudspeaker placed at a 0-degree azimuth approximately 4.5 feet in front of the listener.

B. Listeners

1. General

The hearing-aid users were enrolled in an ongoing study entitled, "Multi-site study of the efficacy of speech-perception training for hearing-aid users," described by Miller *et al.* (2015). Listeners were recruited and tested at five participating sites (see Acknowledgments). The listeners met the following inclusion criteria; each was (1) an adult between 35 and 89 years of age and (2) a native speaker of English. In addition, the listeners (3) had sensorineural hearing loss with pure tone average (PTA) at 500, 1000, and 2000 Hz less than 75 dB HL in each ear, (4) had used a hearing aid in one or both ears for at least 3 months, and (5) had hearing aids that were in working order. Listeners also had (6) no significant conductive hearing loss, (7) no current otologic pathology, (8) no disease or medication that may affect hearing or performance, (9) no significant history of neurologic disease including stroke, (10) normal or corrected normal vision [Snellen (20/50)], (11) a reading level at or above the 4th grade, (12) a score above 85 on the Wechsler Adult Intelligence Scale (WAIS), (13) a score above 20 on the Saint Louis University Mental Status Examination (SLUMS), and (14) the ability to perform project tasks. All listeners signed informed consent forms as approved by each site's Institutional Review Board (IRB). The 57 listeners in this study were the same group that participated in the earlier study by Miller *et al.* (2017), except for two listeners that were not given the sentence tests because of procedural errors and were therefore excluded from the analyses presented in this report.

2. Sexes and ages

Of the 57 listeners, 24 were females and 33 were males; their ages ranged from 51 to 86 years with a mean of 71 years.

3. Audiograms

Pure-tone thresholds were measured by air conduction for the right and left ears prior to participation in the study. An initial evaluation of these thresholds was made by calculating the high-frequency pure tone average (HFPTA), defined by Wilson (2011) as the average hearing level (HL) at the test frequencies of 1000, 2000, and 4000 Hz. The HFPTAs in the listeners' better ears ranged from 25 to 72 dB HL with a mean of 49 dB HL, and in the poorer ears

ranged from 32 to 80 dB HL with a mean of 55 dB HL. The difference in HFPTA between the better and worse ears ranged from 0 to 17 dB with a mean of 6 dB and a mode of 5 dB. The average thresholds of the better ears and of the worse ears fall off at about 9.8 dB per octave over the range 500–4000 Hz and at lower rates between 250 and 500 Hz and above 4000 Hz.

4. Hearing aids

Among the 57 study participants, 55 wore bilateral hearing aids and two wore unilateral aids. Eighty percent wore behind-the-ear aids; the remainder wore in-the-ear or in-the-canal style. Testing was conducted with the hearing aids at the user's customary volume and/or program settings. Consistent hearing-aid function was verified throughout the study, as follows. At each visit, the participant was asked to report any hearing-aid malfunction. If a change in hearing-aid status was noted, testing was postponed, and the participant was referred to the hearing-aid provider for repair. If no concerns were noted, the hearing aids were visually inspected by the tester for damage or blockage. Batteries were checked and replaced if necessary. Functional amplification was evaluated in two ways on entry to the study. First, coupler output was measured as a function of frequency at a range of input levels, using the appropriate coupler (HA1 or HA2) for the style of aid. Second, the real ear aided response was measured for input levels of 50, 65, and 80 dB sound pressure level (SPL). An abbreviated set of coupler measurements was also completed at the beginning of each listening visit. Measurements ± 5 dB re: baseline results were required to demonstrate stable hearing-aid function and to proceed with testing.

C. Syllable-constituent identification

1. General

The measurement of syllable-constituent identification for the individual participants of the present paper is described in detail and justified in [Miller et al. \(2017\)](#) and is briefly summarized here. Each constituent was recorded by eight adult talkers (four males and four females). Half of the males and females were under 40 years of age, and half were over 40. Onsets and codas were recorded with a variety of nuclei as described in [Miller et al. \(2017\)](#). Nuclei were recorded only in the /hNd/ context, that is, in the /h/ (Nucleus) /d/ context.

2. Measurement of syllable-constituent identification

Each participant identified, in separate blocks, 45 onsets, 28 nuclei, and 36 codas in quiet (SNR = 40 dB). A percent correct score in quiet (PCq) was found for each constituent type. The listeners were then tested adaptively to determine the SNRs associated with two other percent correct scores associated with similar loci on upper branches of their logistic functions relating PC to SNR. The

data for the three constituent types were then combined, and the best-fitting logistic that described the relation of the overall syllable-constituent PC with SNR was found for each participant. These fitted functions enabled the calculation by interpolation of each participant's percent-correct-constituent identification at each of the SNRs (−5, 0, 5, 10, 15 dB) used in the study of sentence perception. For additional details, see [Miller et al. \(2017\)](#).

D. Sentence perception

1. SPATS sentence task

a. Sentences and SNRs. This task involves the recognition of words in short, meaningful sentences, presented in a background of 12-talker speech babble. The sound level of the combined speech and noise was held constant at an overall level of 65 dBC, and the SNR was varied randomly from sentence to sentence from −5 to +15 dB in 5 dB steps. The SNRs were intended¹ to represent the levels commonly encountered in everyday situations in which people try to communicate with speech and are described as “noisy” ([Pearsons et al., 1977](#); [Olsen, 1998](#); [Smeds et al., 2015](#)). The sentences use a simple vocabulary and are four to eight words in length. Typical sentences are “Let's go to the store” and “Hurry, or we will miss the train.”

b. Talkers. A corpus of over 1000 sentences was recorded by 14 native speakers of middle American English (also called the “The South Midland Dialect”; [Labov et al., 1997](#)). This group of talkers included males and females, with ages that ranged from 21 to 80 years.

c. Trial procedures. A sentence trial begins with the listener hearing a sentence at one of the five SNRs. When the recorded sentence ends, an alphabetized list of words is displayed on the listener's computer screen as exemplified on the top panel of Fig. 1.

This list includes all the words in the spoken sentence, plus three foils for each target word. The foils share syllabic constituents with the target words and sometimes differ by only a single vowel or consonant. Additional situational context for each N-word sentence is provided by displaying N numbered positions in a line at the top of the display. As words are correctly identified, they are “greyed out” in the list and are entered in their proper sequential positions in the line at the top of the display as schematically illustrated in the middle panel of Fig. 1. The listeners' task is to click on the words heard in the sentence, in any order, making their best guesses when uncertain.

d. Scoring. One measure of performance on this task is the number of responses required to identify all the target words, ranging from a minimum of N responses for an N-word sentence to a maximum of 4xN responses at worst. Two additional features of the task encourage listeners not to devote a great deal of time attempting to use the list of words and foils to deduce the possible words in a sentence.

	(1)	(2)	(3)	(4)
	BELT	HEAT	PELT	
Click the mouse	BUCKLE	LOCAL	PEOPLE	
on the words you	DEALT	MELT	SEAT	
heard	FEET	MORE	SHEET	
	FOR	OR	UNCLE	
			YOUR	

	(BUCKLE)	(2)	(SEAT)	(4)
	BELT	HEAT	PELT	
Temporal	BUCKLE	LOCAL	PEOPLE	
Penalties	DEALT	MELT	SEAT	
4	FEET	MORE	SHEET	
Errors	FOR	OR	UNCLE	
1			YOUR	

	(BUCKLE)	(YOUR)	(SEAT)	(BELT)
	BELT	HEAT	PELT	
Temporal	BUCKLE	LOCAL	PEOPLE	
Penalties	DEALT	MELT	SEAT	
8	FEET	MORE	SHEET	
Errors	FOR	OR	UNCLE	
1			YOUR	

FIG. 1. Schematic examples of sentence screens. The top panel shows a sentence screen that appears immediately after the listener hears the sentence, “Buckle your seat belt!” The middle panel is an example of a screen a listener sees during the process of sentence identification. At this stage there are four temporal penalties [see Sec. IID 1 d and Eq. (1), below], one word selected in error, and two words correctly selected. Correctly selected items have been “greyed out,” correct words have replaced the item numbers in the parentheses above, and the incorrect selection is lined out, e.g., ~~SHEET~~. The bottom panel exemplifies a sentence screen shown after all words in the sentence have been selected. In this case there have been four words correctly selected, one word selected in error, and eight temporal penalties. The effective percent correct (ePC) as defined by Eq. (1) of the text is $100 \times (4 / (4 + 1 + 8)) = 30.8\%$. Note that screens are updated after each event in the sentence identification process, that is, after each temporal penalty, each correct word selection, and each error.

First, whenever a selection error is made (and only then), the recording is replayed. Second, every time more than five seconds elapse between successive responses, a temporal penalty point is assessed. Sentences are scored using a statistic termed the “effective Percent Correct” (ePC). As shown in Eq. (1), the ePC is computed using the number of words presented (#w), the number of errors (#e), and the number of temporal penalties (#tp):

$$ePC = 100 * [\#w / (\#w + \#e + \#tp)]. \quad (1)$$

If a listener has no errors and no temporal penalties, the ePC is 100%. Thus, the ePC is a joint measure of the accuracy and speed of recognition. This procedure is a form of “Dynamic Assessment” that embeds intervention within an assessment procedure (Feuerstein *et al.*, 1979; Sternberg and Grigorenko, 2002). Note that by showing the correctly identified words in their correct sequential positions in the sentence at the top of the sentence response screen and by

greying out selected words and showing which of these are errors, this procedure converts some of what would be speech-inherent context into situational context. In real life, the use of this kind of information often relies on the listener’s working memory, the listener’s self-evaluations of the accuracy of perception, and the listener’s criteria for requests for repetitions and clarifications. In the SPATS sentence task, many of these aspects of sentence perception are made more structured and more secure by converting them to on-screen-situational context as indicated on Fig. 1 and by replaying the sentence after an error.

2. The sentence test

After preliminary training and familiarization with the sentence task, the sentence test was administered. It consisted of 60 sentences, with 12 presented at each of the following SNRs: −5, 0, 5, 10, and 15 dB. The order of the sentences and their SNRs were randomized for each listener. Each listener’s ePC scores were calculated for each SNR and averaged across SNRs for an overall sentence score in noise.

E. Sentence scores as related to syllable-constituent scores

For each of the 57 listeners, the percent of correct syllable-constituent identifications, Constituent PC, was calculated using the fitted logistic functions (Miller *et al.*, 2017) for each of the five SNRs used in the sentence task, which could then be related to the ePCs at the same five SNRs. Next, Eq. (2) was fitted to these data for each subject individually:

$$ePC' = 100 * [p(\text{txt}) + ((1 - p(\text{txt})) * (1 - (1 - ((\text{Constituent PC})/100)^i)))] \quad (2)$$

The first parameter in the expression, $p(\text{txt})$, is a measure of a listener’s usage of situational context to arrive at the words in the sentence. It can be interpreted as the proportion of the sentence that the listener infers from the context provided by the situation in which the sentence is heard, which in our case includes all the information on the sentence screens as illustrated in Fig. 1. The second term $((1 - p(\text{txt})) * (1 - (1 - ((\text{Constituent PC})/100)^i)))$ is meant to reflect the proportion of the sentence information **not** given by the situational context, **but** based on perception of the syllable constituents and use of speech-inherent context. The variable “i” is interpreted as a measure of a listener’s usage of the context inherent in the spoken sentence in arriving at the words in a sentence.² The Solver algorithm of Microsoft Excel was used to fit Eq. (2) to each listener’s data relating syllable-constituent scores to sentence scores, with the parameter $p(\text{txt})$ constrained to $(0 \leq p(\text{txt}) \leq 1)$ and the parameter “i” to $(1 \leq “i”)$. The resulting values of $p(\text{txt})$ and “i” are shown on the inset tables of Fig. 2 and on the supplementary material.³

FIG. 2. Examples of fits of Eq. (2) to the data of individual hearing-aid users. Syllable-constituent scores are shown on the x axes and sentence scores on the y axes. The thick curved line is the fitted Eq. (2). The thin diagonal line is the special case where constituent scores equal sentence scores. The dashed vertical line shows chance performance for the syllable-constituent task. The inset-tables show the Pearson R for the observed and fitted sentence scores; the associated standard error of the estimate (SE); the value of the fitted parameters of Eq. (2), "p(txt)" and "i"; and the listener's estimated overall percent context utilization (CU). Graphs in the leftmost column (Panels 1 and 2) are representative of the best fits; those in the middle column (Panels 3 and 4) are representative of the modal fits; and those in the right-most column (Panels 5 and 6) are representative of the poorest fits.

III. RESULTS

A. Fits of Eq. (2) to the data of individual hearing-aid users

1. Examples

Six examples of the fits of Eq. (2) to the results for individual listeners are shown in Fig. 2.

These six were chosen as a representative sample of the results observed for 53 of the 57 hearing-aid users included in the present study. Recall that each listener had five syllable-constituent scores and five sentence scores, one pair for each of the five SNRs. Each panel of Fig. 2 has syllable-constituent scores (Constituent PCs) on the x axis and sentence scores (ePC) on the y axis. The curve defined by Eq. (2) is shown as the thick curved line on each plot, while the thin diagonal line represents the case where the sentence score equals the constituent score. The dashed vertical line represents chance performance on syllable-constituent identification. The inset-tables show, for each hearing-aid user, the Pearson R for the five observed and fitted sentence scores; the associated standard error of the estimate (SE); the value of the fitted parameters of Eq. (2), "p(txt)" and "i"; and the listener's estimated overall percent context utilization (CU). (The definition and calculation of CU is described in Sec. IIIC, below.) Note that when the area between the curved fitted line and the thin diagonal line is larger, the use of context is assumed to be greater, as

is explicated in the discussion of Fig. 8. Graphs in the left-most column (Panels 1 and 2) represent the best fits of Eq. (2) to the data, those in the middle column (Panels 3 and 4) represent the modal fits, and those in the right-most column (Panels 5 and 6) represent the poorer fits. Four of the 57 hearing-aid users had negative correlations between their observed and fitted sentence scores. In these cases, the average of the sentence scores at all five SNRs was quite high (e.g., 85%, 92%, 94%, and 94%), and did not always systematically increase as the SNRs increased.

2. Goodness of fits of Eq. (2) to the observed data

Two measures of the goodness-of-fit to each listener's data were examined. One measure is the linear correlation between each listener's five observed ePCs and five fitted ePCs (ePC's). The other measure is the standard error of the estimate of y on x calculated as shown in Eq. (3),

$$SE = \sqrt{\sum_{i=1}^5 \frac{(y - y')^2}{5}}, \quad (3)$$

where y is the observed ePC and y' is the fitted ePC' for each listener at each of the five SNRs.

The distribution of R 's is shown in Fig. 3 and the distribution of SEs is shown on Fig. 4.

FIG. 3. The distribution of linear correlations between observed and fitted sentence scores for 57 hearing-aid users. The fitted scores were obtained by finding the best fit of Eq. (2) to the observed scores for each of the 57 hearing-aid users.

Note that most of correlations are positive and reasonably high, and that 86% of SE's fall within a range from 2.0% to 10.0%. Thus, many, but not all, of the fits of the equation to the data are quite good.

3. Relation between fitted and observed sentence scores for 57 hearing-aid users

Another demonstration of the goodness of fit is the relations between the five observed and five fitted points for all 57 hearing-aid users. As shown in Fig. 5, the overall linear correlation between the 285 observed and fitted scores is 0.94, and the corresponding SE is 5.71. One can also gain an impression of the adequacy of Eq. (2) for describing the

FIG. 4. The distribution of the standard errors of the estimate of observed sentence scores for 57 hearing-aid users. The fitted scores were obtained by finding the best fit of Eq. (2) to the observed scores for each of the 57 hearing-aid users.

FIG. 5. The relation between fitted and observed sentence scores for 57 hearing-aid users (each contributing five data points). The overall linear correlation between the 285 observed and fitted scores is 0.94, and the corresponding SE is 5.71. This demonstrates that for each listener, the sentence scores (ePCs) tend to be strongly positively correlated with the syllable-constituent scores (Constituent PCs) when context utilization is taken into account.

data by examining the 57 fitted functions in order of decreasing SEs (footnote 1). In summary, it appears that Eq. (2) approximates the observed data well enough to encourage continued study of the proposed model.

B. The fitted parameters: $p(\text{txt})$ and "i"

1. The distributions of $p(\text{txt})$ and "i"

The distributions of the fitted values of $p(\text{txt})$ and "i" are shown on the left and right panels of Fig. 6.

We see on the left panel that 28 listeners appear to have made little use of the situational context, having fitted values of the $p(\text{txt})$ parameter less than 0.15. The remaining 29 listeners are scattered over the range from 0.15 to 0.95, with about a third of all listeners appearing to make substantial use of the on-screen situational context (i.e., they have $p(\text{txt})$ values between 0.5 and 0.9). The right panel shows the distribution of the "i" parameter. This distribution may be bimodal, with one mode near 1.0 and a weak second, broader mode between 2.5 and 3.5. The listeners at the lower mode appear to be making little use of speech-inherent context, while those at and above the upper mode appear to be making substantial use of speech-inherent context. However, more data are needed before decisive interpretations can be made.

2. The relation between $p(\text{txt})$ and "i"

The relation between $p(\text{txt})$ and "i" is shown in Fig. 7.

This scatter plot indicates that the parameter "i" falls from a value of about 3.3 when $p(\text{txt})$ is 0 to a value of close to 1 when $p(\text{txt})$ exceeds 0.8. Notice the large scatter in the values of the parameter "i" for intermediate values of $p(\text{txt})$.

FIG. 6. The distribution of $p(\text{txt})$ is shown on the left panel; the distribution of “i” is shown on the right panel.

We tentatively interpret this to mean that when $p(\text{txt})$ is very large, listeners rely mostly on the situational context and the information given by constituent identification to determine the words in the sentence. In this case, the listeners do not appear to use, or need to use, speech-inherent context to determine the words in the sentence. When $p(\text{txt})$ is less than about 0.8, the values of $p(\text{txt})$ and “i” appear to be nearly independent.

C. Calculation and distribution of an approximate estimate of overall context utilization

While $p(\text{txt})$ and “i” are interpreted as measuring aspects of each listener’s use of context, it is not obvious how they combine to give an overall estimate of each listener’s total use of context. Such an estimate, Context Utilization, is derived as follows and illustrated in Fig. 8. First, we assume

that if a listener makes no use of context, the sentence score will be a simple monotonic function of the syllable-constituent score. Since the exact form of the hypothesized monotonic function is unknown, the diagonal line on Fig. 8 is used as an approximation. We then represent the listener’s overall sentence perception by the fitted line defined by Eq. (2), i.e., the thick curved line on Fig. 8. Finally, we approximate the listener’s overall use of context by the area between the fitted curve and the diagonal line, expressed as a percent of the shaded area. That is, we use numerical integration to find the area between the curved and the diagonal lines, divide it by the shaded area, and multiply by 100 to give an estimate of the overall Context Utilization. For the listener shown in Fig. 8, the Context Utilization estimate is

FIG. 7. The relation between $p(\text{txt})$ and “i” is shown. Values of $p(\text{txt})$ are shown on the x axis, while values of “i” are shown on the y axis.

FIG. 8. Illustration of the calculation of percent context utilization (CU). The thin diagonal line shows sentence perception as a linear function of Constituent PC. The curved thick line shows the fitted sentence scores for the data points shown. The dashed vertical line to the left shows chance performance for the syllable-constituent task.

62.75%. Figure 9 shows the distribution of individual differences in overall Context Utilization.

Note that the listeners, all native speakers of American English, are at or above 30% context utilization. The three poorest users of context fall in the range from 30% to 40%. Most of the listeners fall in the range from 40% to 90%. Nineteen of the listeners appear to be exceptionally good at the utilization of context with estimates of context utilization that range from 70% to 90%. The supplementary material³ shows the individual fits of Eq. (2) for each of the 57 listeners ordered from least to most context utilization.

D. How syllable-constituent perception interacts with context utilization to determine sentence scores

In this section, we explore the relation between syllable-constituent identification and sentence recognition. With a few exceptions, the listeners exhibit a clear trend: The better the syllable-constituent identification, the better the sentence score, which is illustrated in Fig. 10. The few exceptions are listeners with uniformly high sentence recognition scores (ePC's). Depending on context utilization, there are large individual differences in sentence scores when syllable-constituent identification scores are low, below 55% (vertical line, Fig. 10). However, when syllable-constituent scores are high, above 55%, individual differences in sentence scores based on context utilization are much smaller.

On Fig. 10, each listener contributes five data points, one for each SNR used. Each data point shows a listener's fitted sentence score (ePC') on the y axis and the corresponding syllable-constituent identification score, Constituent PC, on the x axis. Additionally, to help elucidate the role of context utilization in sentence perception, we arbitrarily placed the listeners into three groups along the context utilization dimension, as defined in Sec. III C above. The top and bottom groups (upward- and downward-pointing black

triangles, respectively) represent the tails of the distribution, while the middle group (grey diamonds) represents the modal center of the distribution; context utilization scores for these three groups ranged from 78.6% to 89.7% for the top group, from 30.2% to 50.6% for the bottom group, and from 50.8% to 78.5% for the middle group. Note that syllable-constituent scores greater than about 55% are associated with excellent sentence-perception scores (greater than about 87.5%). However, when syllable-constituent scores are lower than 55%, there is wide scatter among the sentence scores (ePC's). It appears that the ability to use context is very important when syllable-constituent identification is poor but less important when syllable-constituent identification is good.

It is also possible that listeners with poor syllable-constituent identification skills and high sentence scores devote more cognitive resources to sentence perception than those with good syllable-constituent identification skills. As a result, it seems reasonable to expect that listeners who must rely on extraordinary context utilization will find the perception of spoken material in challenging environments, such as in multitalker babble, to be a demanding and effortful task. It follows that those listeners with both poor syllable-constituent identification skills and low context utilization skills are likely to find their perception of spoken material to be poor, and perhaps unrewarding. Finally, it seems likely that listeners with very high syllable-constituent scores will find speech perception to be better, perhaps automatic and nearly effortless. Obviously, additional research is needed to confirm these speculations.

E. Sentence scores averaged across SNRs predicted by syllable-constituent scores and each hearing-aid user's parameters $p(\text{txt})$ and "i"

1. General

For each subject, the average sentence score, $[ePC(\text{ave})]$, is defined as the average of the ePCs at the five SNRs. Similarly, the average syllable-constituent score in noise, $(\text{Constituent PC})_N$, is defined as the mean of the Constituent PC's scores at the five SNRs.

2. Linear multiple regression analyses

The regression of $ePC(\text{ave})$ on the best linear combination of $(\text{Constituent PC})_N$, $p(\text{txt})$, and "i" was determined. For comparison, the regression of $ePC(\text{ave})$ on the best linear combination of the constituent score in quiet, $(\text{Constituent PC})_Q$, $p(\text{txt})$ and "i" was also determined. The results are shown in Fig. 11.

The left panel of Fig. 11 shows the relation between observed and fitted sentence scores when average syllable-constituent scores in noise are used to calculate ePC' , while the right panel shows the result for the case where ePC' is calculated using syllable-constituent scores in quiet. Clearly, optimal linear combinations of the three predictors account for most of the variance in the observed sentence scores. We can see that, except for the weights on the

FIG. 9. The distribution of estimated context utilization for 57 hearing-aid users.

FIG. 10. The relation between syllable-constituent perception and sentence perception (ePC) as conditioned by percent context utilization. The upper panels separate the top and bottom groups from the middle group so that more of the individual points can be seen.

predictor variables, the result on the right is nearly identical to the result on the left. Otherwise said, syllable-constituent scores in quiet correlate with sentence perception in noise almost as well as syllable-constituent scores obtained at the same SNRs as the sentence scores. Taken as a whole, the results support the proposition that hearing-aid users' sentence perception in noise is strongly related to syllable-constituent perception when the latter is combined with estimates of the ability to use situational context extrinsic to the speech signal and the ability to use the context inherent to the speech signal. Since syllable-constituent

identification in noise for these listeners is highly correlated with that in quiet (Miller *et al.*, 2017), it makes very little difference whether syllable-constituent perception is measured in noise or in quiet.

3. Application of Eq. (2) to average sentence scores in noise

The same issues as those addressed in Sec. III E 2 above can be addressed by applying Eq. (2) to the syllable-constituent and sentence scores averaged across

FIG. 11. The best linear combination of each listener's average Constituent PC, the parameter $p(\text{txt})$, and the parameter "i" for predicting the average ePC in noise for each of 57 hearing-aid users. The left panel displays the results for the case where Constituent PCs are measured at the same SNRs as used for the sentences. The right panel displays the results when each listener's average Constituent PC in quiet is used. The values of $p(\text{txt})$ and "i" are found by fitting Eq. (2) to each listener's data as described in the text. The weights for the linear combinations of the predictor variables are shown in the inset tables and the linear regression equations are shown above the x axes at the tops of the graphs.

the five SNRs. First, consider the case wherein each listener's average sentence score is predicted by combining the average (Constituent PC)_N, $p(\text{txt})$, and "i" as prescribed by Eq. (2). The result is shown on the left panel of Fig. 12.

The fit is nearly perfect, with the correlation between the observed and fitted values being 0.99 and the standard error of the fit being 2.00%. The right panel is for the case of the syllable-constituent score in noise, (Constituent PC)_N, calculated from the syllable-constituent score in quiet, (Constituent PC)_Q, by means of the polynomial

equation relating the two as reported by Miller *et al.* (2017). The fit obtained by this second procedure is almost as good as that in the left panel, with a correlation between the observed and fitted values of 0.96 and a standard error of the fit of 3.84%. Sentence scores in noise are highly correlated with the syllable-constituent scores in the same noise as well as with the syllable-constituent scores in quiet. Importantly, the panels of Fig. 12 demonstrate that when data are stabilized by using averages, the quantitative model of Eq. (2) appears to describe the observed data very well.

FIG. 12. Showing the fits of the model of Eq. (2) to the average syllable-constituent scores and the average sentence scores for all five SNRs for each hearing-aid user. The left panel shows the results when the syllable-constituent scores are the average score in noise for each individual listener, i.e., for the average for the same five SNRs as used for the sentence scores. The right panel shows the results when the syllable-constituent score in noise is calculated from the syllable-constituent score in quiet by the equation reported in Miller *et al.* (2017).

IV. GENERAL DISCUSSION

A. Review of basic concepts: Speech signal, phonetic detail, speech-inherent context, situational context

1. The speech signal

We define the speech signal as having several parts: One part is the speech waveform generated by the talker that reaches the listener's ears. Other parts are the observable positions and movements of the talker's speech articulators, including observable indicators of these articulatory positions and movements.

2. Phonetic detail

We take the ability to recognize the segments of the speech signal, which we define as syllable constituents, to be a very important aspect of the recognition of phonetic detail. Of course, many other phonetic details likely contribute to the overall processing of sentence perception. These include perception of the many cues that help listeners segment the speech stream into syllables, words, and phrases. Listeners' abilities to recognize prosodic cues that determine the word, phrase, and sentence stress patterns also have been implicated in sentence recognition (Cutler, 2015). In the present paper, some of these unmeasured skills may be confounded with the use of context. Additional research is needed to clarify these issues.

3. Speech-inherent context

Speech-inherent context is carried by the speech signal itself. It is carried by the redundancies in the speech signal and includes not only phonetic details but also articulatory, grammatical and semantic constraints.

4. Situational context

Situational context includes *a priori* knowledge the listener has as to what talkers are likely to say in any given situation. However, it could also include non-speech communicative gestures such as pointing, pantomime, and air drawing. In the special case of the SPATS sentence task, the words and phrases perceived during sentence recognition are presented on the sentence screen and provide situational context.

B. Comments on the syllable-constituent, contextual theory of speech perception: Caveats

As embodied in Eq. (2), this theory is not intended to describe the details of the very complicated moment-to-moment interactions between various sources of information that occur during the process of recognition of words in sentences. Rather, we suggest that the theory offers a way to estimate, after completion of the sentence recognition process, the relative contributions of phonetic detail and contextual information to the recognition process. Also, as discussed above in Sec. IV A 2, phonetic details other than those captured by syllable-constituent identification can

contribute to the process of sentence perception. Noteworthy are those details associated with speech prosody and other cues that aid the process of segmenting the speech stream into syllables, words and phrases (see Cutler, 2015, for examples). In the present case, wherein all listeners were native speakers of American English, these seemingly varied skills are probably highly correlated with the ability to identify syllable constituents. Finally, we note that much of the situational context provided by the SPATS sentence task is not available in commonly occurring speech communications. Indeed, the textual material provided as part of the SPATS sentence task probably favors listeners that are rapid readers, although additional research is needed to test this hypothesis. In most everyday situations, listeners with high social awareness and knowledge of relevant jargon will likely benefit most from situational context.

V. SUMMARY AND CONCLUSIONS

Sentence perception in noise appears to be highly correlated with the ability to perceive phonetic detail, whether in noise or in quiet, in combination with the abilities to use situational and speech-inherent context. This suggests that designers of prosthetic devices such as hearing aids or cochlear implants should focus not only on noise reduction but also on the optimization of the transmission of phonetic details in quiet, as this may optimize perception in quiet and in noise. In choosing prosthetic devices, it appears that the optimal device for speech perception will be the one providing the best syllable-constituent perception in quiet. In fact, a reasonable goal for listeners with age-related hearing loss might be for such listeners to reach or exceed 71% identification of syllable constituents in quiet, as this implies syllable-constituent identification in noisy situations equal to or greater than 55%. Our data suggest that listeners who reach syllable-recognition scores equal to or greater than 71% in quiet may be more likely to have satisfactory sentence perception in quiet and in noise. (Of course, as more data are collected using a variety of sentence tests, the exact target value may shift away from 71%.) Most importantly the data presented suggest that those interested in hearing-aid design, fitting, and speech-perception training should focus not only on methods to reduce noise but also on methods to improve the perception of phonetic detail and on methods to improve the use of situational and speech-inherent context. Finally, the data presented here appear to support further investigation and development of the syllable-constituent, contextual model of speech perception.

ACKNOWLEDGMENTS

This work was supported by NIH-NIDCD Grant No. R21/33 DC011174 entitled, "Multi-Site Study of Speech Perception Training for Hearing-Aid Users," C.S.W., PI. M.R.L.'s participation in this project is supported by VA Rehabilitation Research and Development Service Senior

Research Career Scientist Award #C4942L. The data were collected at the Medical University of South Carolina, J.R.D., Site PI; Northwestern University, P.E.S., Site PI; the University of Maryland, S.G.-S., Site PI; the University of Memphis, D.J.W., Site PI; and the VA Loma Linda Healthcare System, M.R.L., Site PI. The authors wish to acknowledge contributions to this paper and the multi-site study by Heather Belding, Holley-Marie Biggs, Megan Espinosa, Emily Franko-Tobin, Sara Fultz, Sarah Hall, Gary R. Kidd, Rachel Lieberman, Laura Mathews, Roy Sillings, Laura Taliferro, and Erin Wilbanks. Daniel Maki, CFO of CDT, Inc. and Professor Emeritus of Mathematics, made helpful suggestions concerning issues related to fitting the logistic functions to the data. Maki, Miller, and Watson are all stockholders in CDT, Inc. and may profit from sales of SPATS software.

¹One might argue that most hearing aid users would not consider an SNR of 15 dB to be “noisy.” A more appropriate range of SNRs for hearing-aid users would be from about -10 to $+10$ dB. If still less advantageous SNRs were used, many hearing-aid users would have scores approaching chance, and the correlation between syllable-constituent scores in quiet and noise would tend toward zero. The conclusion of Miller *et al.* (2017) that there is a high correlation between syllable-constituent scores in quiet and in noise is supported only when scores in noise are significantly above chance.

²We note that the term $(1 - (1 - p(\text{const}))^i)$ was suggested by the previous work of Boothroyd and Nittrouer (1988). However, here, “ i ” is a fitted parameter that depends on relations between Constituent PCs and sentence scores as described by Eq. (2). Thus, our parameter “ i ” is intended to be much more global than the “ k ” of Boothroyd and Nittrouer. Boothroyd and Nittrouer estimated their k by comparing the identification accuracy for CVC nonsense syllables and CVC words. They also compared the accuracy of the identification of isolated words and the same words spoken in sentences to obtain j -values as a measure of the context provided by sentences for the perception of the words contained in them. For a more detailed presentation of how to model the use of context in sentence recognition, see Bronkhorst *et al.* (1993), Bronkhorst *et al.* (2002), and the references cited therein.

³See supplementary material at <https://doi.org/10.1121/10.0000563> for fits of Eq. (2) for all 57 hearing-aid users. The graphs are ordered with respect to goodness of fit from worst to best.

- Allen, J. (1996). “Harvey Fletcher’s role in the creation of communication acoustics,” *J. Acoust. Soc. Am.* **99**, 1825–1839.
- Boothroyd, A., and Nittrouer, S. (1988). “Mathematical treatment of context effects in speech recognition,” *J. Acoust. Soc. Am.* **84**, 101–114.
- Bronkhorst, A., Bosman, A., and Smoorenberg, G. (1993). “A model for context effects in speech recognition,” *J. Acoust. Soc. Am.* **93**, 499–509.
- Bronkhorst, A., Brand, T., and Wagener, K. (2002). “Evaluation of context effects in sentence recognition,” *J. Acoust. Soc. Am.* **111**, 2874–2886.
- Cutler, A. (2015). *Native Listening: Language Experience and the Recognition of Spoken Words* (MIT Press, Cambridge, MA).
- Feuerstein, R., Feuerstein, S., Falik, L., and Rand, Y. (1979). *Dynamic Assessments of Cognitive Modifiability* (ICELP Press, Jerusalem, Israel).
- Hall, T. A. (2006). “Syllable phonology,” in *Encyclopedia of Language & Linguistics*, 2nd ed., edited by K. Brown (Elsevier, Oxford, UK), Vol. 12, pp. 329–333.
- Labov, W., Ash, S., and Boberg, C. (1997). A national map of the regional dialects of American English. http://www.ling.upenn.edu/phono_atlas/NationalMap/NationalMap.html#Heading6.
- Miller, J. D., Watson, C. S., Dubno, J. R., and Leek, M. R. (2015). “Evaluation of speech-perception training for hearing-aid users: A multi-site study in progress,” *Semin. Hear.* **36**(4), 273–278.
- Miller, J. D., Watson, C. S., Leek, M. R., Dubno, J. R., Wark, D. J., Souza, P. E., Gordon-Salant, S., and Ahlstrom, J. B. (2017). “Syllable-constituent perception by hearing-aid users: Common factors in quiet and noise,” *J. Acoust. Soc. Am.* **141**(2), 2933–2946.
- Olsen, W. O. (1998). “Average speech levels and spectra in various speaking/listening conditions: A summary of the Pearson, Bennett, & Fidell report,” *Am. J. Audiol.* **7**, 21–25.
- Pearsons, K. S., Bennett, R. L., and Fidel, S. (1977). “Speech levels in various noise environments: Project Report on contract 01-2468,” Washington, DC, pp. 82.
- Smeds, K., Wolter, F., and Rung, M. (2015). “Estimation of speech-to-noise ratios in realistic sound scenarios,” *J. Am. Acad. Audiol.* **26**(2), 183–196.
- Sternberg, R. J., and Grigorenko, E. L. (2002). *Dynamic Testing: The Nature and Measurement of Learning Potential* (University of Cambridge, Cambridge, UK).
- Wilson, R. H. (2011). “Clinical experience with the words-in-noise test of 3430 veterans: Comparisons with pure-tone thresholds and word recognition in quiet,” *J. Am. Acad. Audiol.* **22**(7), 405–423.