Published in final edited form as:

J Speech Lang Hear Res. 2006 August ; 49(4): 823–834. doi:10.1044/1092-4388(2006/058).

Consistency of Sentence Intelligibility Across Difficult Listening Situations

Eric W. Healy and Allen A. Montgomery

Arnold School of Public Health, University of South Carolina, Columbia

Abstract

Purpose—The extent to which a sentence retains its level of spoken intelligibility relative to other sentences in a list under a variety of difficult listening situations was examined.

Method—The strength of this *sentence effect* was studied using the Central Institute for the Deaf Everyday Speech sentences and both generalizability analysis (Experiments 1 and 2) and correlation (Analyses 1 and 2).

Results—Experiments 1 and 2 indicated the presence of a prominent sentence effect (substantial variance accounted for) across a large range of group mean intelligibilities (Experiment 1) and different spectral contents (Experiment 2). In Correlation Analysis 1, individual sentence scores were found to be correlated across listeners in each group producing widely ranging levels of performance. The sentence effect accounted for over half of the variance between listener-ability groups. In Correlation Analysis 2, correlations accounted for an average of 42% of the variance across a variety of listening conditions. However, when the auditory data were compared to speech-reading data, the cross-modal correlations were quite low.

Conclusions—The stability of relative sentence intelligibility (the sentence effect) appears across a wide range of mean intelligibilities, across different spectral compositions, and across different listener performance levels, but not across sensory modalities.

Keywords

sentence intelligibility; generalizability analysis; CID sentences

The intelligibility of sentences is very high when spoken by typical talkers to listeners with normal hearing. However, under difficult listening conditions, such as those imposed by poor signal-to-noise ratio, filtering, or hearing loss, intelligibility may be reduced substantially. Further, under given conditions there may be considerable differences in intelligibility among sentences within a pool of sentences. This is reflected in the significant difficulties encountered when creating lists of sentences that are equal in intelligibility (Bilger, Neutzel, Rabinowitz, & Rzezczkowski, 1984; Giolas & Duffy, 1973; Hood & Dixon, 1969; Rippy, Dancer, & Pittenger, 1983; Webster, 1984). This difficulty exists in creating sentence materials for speech reading as well (Hinkle, 1978; Wilson, Dancer, & Stamper, 1984). Because of learning effects associated with being exposed to a sentence twice, such lists are desirable when the performance of one individual is to be assessed under a variety of conditions.

Differences in intelligibility are even more apparent when individual sentences are considered. Healy and Warren (2003), for example, presented the 100 Central Institute for the Deaf (CID; Davis & Silverman, 1978) Everyday Speech sentences through narrow-band filters and found that the mean sentence intelligibility (key words correct per sentence) from the least intelligible to the most intelligible of the 100 sentences ranged over 80 percentage points within each of several listening conditions.

Many attempts have been made to delineate the factors that contribute to differences in intelligibility among sentences, including word frequency and neighborhood density (Bell & Wilson, 2001), prosodic characteristics (Laures & Weismer, 1999), sentence length (Clauser, 1976), syntactic complexity (Cheung & Kemper, 1992), and semantic predictability (Giolas, Cooker, & Duffy, 1970), among others. Whatever the sources of differences in sentence intelligibility, the question arises concerning how consistent these differences are under varying listening conditions; that is, to what extent do individual sentences retain their intelligibility, relative to other sentences in a list, across a range of listening conditions and listener abilities? This knowledge—the extent to which sentences have an inherent intelligibility—would potentially further both our basic understanding of human speech processing and the tests used to assess it. It could also be of clinical value, for example in constructing tests in which a range of sentence intelligibility is desirable. Such tests could potentially be used to assess an individual's ability to process sentences that vary systematically in auditory-reception difficulty, rather than the traditional use of equivalent lists. Again, the necessity of equivalent lists of sentences for evaluating normal and impaired listeners has forced attention on eliminating the effect rather than studying it.

The consistency of this relative intelligibility of sentences (within a list) across conditions of presentation is the focus of this study and will be referred to as the *sentence effect*. To our knowledge, this phrase has not been used for this concept, and it has not been widely studied. In this study, sentence intelligibility values consisting of the percentage key words correct per sentence per listener for the original 100 CID sentences under eight listening conditions (20 listeners/condition) were used to address this question.

The CID Everyday Speech sentences were developed at the Central Institute for the Deaf to represent “everyday” speech, and the original set of 100 varies widely in length (2–12 words) and sentence structure, including declarative and imperative statements and two forms of questions (see Davis & Silverman, 1978, Appendix, for details, or Alpinier & McCarthy, 2000, pp. 324, 622–623). They have been used in over 50 studies and continue to be used today. Although revised versions exist, the original 100 CID sentences are well suited for the study of the sentence effect because they differ so widely in intelligibility in any given study.

In this article, we explore the strength of the sentence effect. In Experiment 1, the effect is examined across large variations of overall intelligibility due to changes in effective bandwidth at a constant center frequency. In Experiment 2, the effect is examined across changes in spectral composition at a constant intelligibility. In addition, the influence of listener ability (Correlation Analysis 1) and sensory mode of presentation (Correlation Analysis 2) on the consistency of individual sentence intelligibility was considered.

Experiments 1 and 2

Experiments 1 and 2 involve the preliminary stages of generalizability analysis. This technique has been applied to many areas of inquiry where test development issues were considered (Shavelson & Webb, 1991). Recently, several applications of generalizability theory to the communication sciences have been made, including studies of visual speech intelligibility (Demorest & Bernstein, 1992), auditory speech intelligibility (Demorest & Bernstein, 1993),

and speech quality judgments (O'Brian, O'Brian, Packman, & Onslow, 2003; O'Brian, Packman, Onslow, & O'Brian, 2003; Scarsellone, 1998).

Upon its introduction, generalizability theory represented a breakthrough in test development. Its power comes from the fact that several factors are under the control of the test developer and that the influence of modification to these factors can be assessed. These factors can include different lists of items, different groups of test-takers and, in the context of auditory testing, different numbers of talkers or listening conditions. The influence of these factors is evaluated by their contribution to the variance of test scores across a population of test-takers and can be expressed as a percentage of the total variance accounted for by each factor (and its interactions). The theory can then be used to assist many aspects of test development, including, for example, the effect of increasing test length. In the present study, generalizability theory was used to assess the contribution of selected aspects of sentence understanding. These factors, including the sentence effect, were assessed using the extent to which each contributed to total variance of test scores.

Generalizability theory was developed in modern form by Cronbach, Gleser, Nanda, and Rajaratnam (1972). In it, they sought to utilize a traditional factorial analysis of variance (ANOVA) model that draws on the Cornfield and Tukey (1956) algorithm to estimate the variance components underlying a testing situation. Several accessible explanations of generalizability theory are available (Demorest & Bernstein, 1992; Di Nocera, Ferlazzo, & Borghi, 2001; O'Brian, O'Brian, et al., 2003; Shavelson & Webb, 1991; VanLeeuwen, Barnes, & Pase, 1998). In this study, we draw on a powerful model developed by Demorest and Bernstein (1992, 1993). However, only the initial stages of their generalizability analysis, estimating the importance of factors within an ANOVA framework, are used.

Demorest and Bernstein (1992) used generalizability theory to estimate the relative importance of three sources of variability in the speech reading of sentences: (a) the talker, (b) the observer, and (c) the sentences. They employed two talkers of differing overall visual intelligibility and presented the 100 CID sentences in two sets to 104 normal-hearing observers without sound. After some preliminary analysis they settled on a model that included factors labeled *Talker*; *Sentence*; *Subject*; *Set* (Lists 1–5 or 6–10); and interactions, including an error term, *Subject* $\times$ *Sentence* (original terminology).

Demorest and Bernstein (1992), using percentage total words correct per sentence, found that the variability in an observer's score on a single sentence is composed of the following variance components: *Subject* $\times$ *Sentence*, 49.4%; *Sentence*, 30.0%; *Subjects*, 9.4%; *Talker* $\times$ *Sentence*, 5.0%; and *Talker*, 4.3% (minor interactions totaling 2% estimated variance are omitted here). These results illustrate the capability of generalizability theory to allow the estimation of the contribution of multiple factors underlying a test-taker's performance. First, there was a substantial interaction between subjects and individual sentences. This component includes residual measurement error as well as the idiosyncratic reaction of each observer to each sentence and is considered to be an estimate of overall measurement error. This seems large at 49.4% but is not surprising considering the difficulty of speech reading sentences without sound or thematic context and the huge uncertainty caused by the missing, ambiguous phonetic information. In addition, the model essentially treats each sentence as a one-item test, which adds to the uncertainty. Of more relevance to the present study is the strength of the *Sentence* factor. At 30.0% variance accounted for, it is much larger than the effect of the two talkers or the subjects. The inherent ease (or difficulty) of a sentence thus appears to be the dominant nonerror factor influencing the speech reading of the CID sentences. Of course, it is possible that the *Sentence* factor is strong only in speech reading, with its great uncertainty and dependence on visual memory. Indeed, the primary motivation for the present study was to

examine the strength of the Sentence factor in comparable degraded auditory listening situations.

In Experiment 1, the sentence effect was examined using filtered sentences centered on one frequency and adjusted to produce either four (Experiment 1A) or three (Experiment 1B) widely varying mean intelligibilities. In Experiment 2, band pairs were selected that differed in center frequency but yielded similar speech intelligibilities, with either moderate intelligibility (Experiment 2A) or relatively high intelligibility (Experiment 2B).

Experiment 1A

Method—For the current experiment, four conditions were selected from Healy (1999). Although the details of processing are of secondary interest and are available elsewhere (Healy, 1999; Healy & Warren, 2003), of primary interest is the selection of conditions that resided within the same narrow spectral region but provided a wide range of mean intelligibilities. A single digital recording (22-kHz sampling with 16-bit resolution) of the CID sentences was used for all conditions. The sentences were produced using natural rate and intonation by a male speaker having a General American dialect. The four processing conditions all involved narrow-band filtering at a center frequency of 1500 Hz. Two of the conditions were created by filtering to a 1/3-octave band having filter slopes of either 300 dB/octave (a condition that yielded a mean intelligibility of approximately 60%, labeled *60PC*) or 1700 dB/octave (*20PC*). A third condition (*80PC*) employed 2/3-octave filtering at 1700 dB/octave. The fourth condition (*40PC*) was created by amplitude-modulating a pair of tones using the temporal envelopes from a pair of juxtaposed 1/6-octave speech bands, to produce a stimulus centered at 1500 Hz and having an overall bandwidth of approximately 1/3 octave. In each condition, stop-band attenuation was at least 60 dB. The actual group mean intelligibility values for each condition are as follows, *20PC*: 18.40% ($SD = 9.04$); *40PC*: 40.50% ($SD = 6.44$); *60PC*: 59.17% ($SD = 9.34$); *80PC*: 82.69% ($SD = 5.46$). The processed speech stimuli were presented over Sennheiser HD 250 headphones at a slow peak level of 70 dBA calibrated using a flat-plate coupler and Brüel & Kjær 2204 sound-level meter.

Native English speakers between the ages of 18 and 40 years ($Mdn = 20$ years) served as listeners. All reported no known hearing problems, and none had heard the sentence materials previously. Different groups of 20 listeners were randomly assigned to each condition. Following a 10-sentence practice list drawn from the Speech Perception in Noise (SPIN) test (Kalikow, Stevens, & Elliott, 1977), each listener heard all 100 CID sentences in a single processing condition. Listeners heard each sentence only once and were encouraged to guess if unsure of the content. The experimenter was seated with the subject in an audiometric booth and controlled the presentation of each sentence and scored the proportion of key words reported correctly. These data encompass almost the entire range of sentence intelligibility, with some individual listeners producing mean intelligibilities across 100 sentences below 10% in the *20PC* condition and others above 90% in the *80PC* condition.

Analysis—In the ANOVA underlying the generalizability analysis, the Listening Condition is a fixed factor, and Listener is a random factor, nested under Listening Condition. The CID sentences are considered to be a random sample of sentences from a large population and thus also become a random factor. These factors result in the following model (Table 1) for the ANOVA. The analysis differs in Experiments 1A, 1B, 2A, and 2B only in the number of different bandwidth or frequency conditions, designated in Table 1 as “a.” The number of levels for the Listening Condition are as follows ($a = 4$ for Experiment 1A, $a = 3$ for Experiment 1B, $a = 2$ for Experiments 2A and 2B). The values of $b = 100$ sentences and $c = 20$ listeners are the same for all analyses.

An ANOVA was performed to generate the mean squares and calculate the variance components for the factors of the model. The dependent variable was the percentage of key words correct per sentence per listener.

Results—The results of the ANOVA are shown in Table 2. The analysis indicated that the main effects of Condition, Listener, Sentence, and all interactions were all significant with $p < .001$. This is not surprising given the overall power of the ANOVA (total $df = 7,999$) and the large range of mean intelligibility inherent in the listening conditions, CID sentences, and listeners. More importantly, the analysis generated the estimates of variance components and allowed the calculation of the percentage variance attributed to each component.

Discussion—Table 2 shows that the largest component was the Sentence $\times$ Listener interaction, which is considered a measure of experiment-wide error (accounting for 39.83% of the total variance). This seems large, but it is actually small compared with percentage error found for the CID sentences in the two comparable studies of speech reading: Demorest and Bernstein (1992) found error in the range of 50%, as did Adams (2002) in his study of speech reading. The error may reflect the uncertainty and consequent guesswork involved in listening to very incomplete auditory (or visual) versions of sentences but, as noted earlier, it is also a function of the analysis treating each sentence, in essence, as a one-sentence test.

The second largest component of the variance accounted for in Table 2 (31.22%) is due to Listening Condition. This is a function of the variance of the means of the four listening conditions, which ranged from 18.40% to 82.69% and was “built in” to the analysis to provide an extreme test of the sentence effect. Despite this, the sentence effect accounted for 18.83% of the variance. This can be considered a substantial effect considering the magnitude of the error and the variance of the listening conditions. The effect of Listener (within each group), however, is small (3.02%) relative to the other sources of variation, as is the Condition $\times$ Sentence interaction (7.09%). This latter finding suggests that the changes in processing required to produce the four mean intelligibilities did not differentially affect the intelligibility of the sentences to a large degree, and the sentence effect held across a wide range of overall intelligibilities.

Experiment 1B

Another consideration, illustrated in Demorest and Bernstein (1992), is the interdependent (relative) nature of the variance percentages that generalizability theory generates. For example, in their study the talkers differed by about 9% in visual intelligibility. Had they differed by 20% or 30%, the talker contribution would have been larger, at the expense of the other factors, because the percentages are constrained to sum to 100. The purpose of Experiment 2 was to examine the effect of reducing the variance of the condition means on the Sentence factor and the other components.

Method and analysis—The method, the data sets, and the ANOVA model were identical to Experiment 1A with the exception that the four listening conditions were reduced to three by removing the 80PC condition. Either the condition with the highest overall mean percentage correct (80PC) or the lowest (20PC) could have been chosen for removal for this demonstration. The primary effect (and intent) of this was to reduce the variance due to the mean intelligibility of the listening conditions, as seen in Table 3.

Results—The results of the ANOVA are shown in Table 3. All effects were again significant. The most apparent change, compared with Table 2, is the anticipated reduction of the Condition factor (now based only on the 20PC, 40PC, and 60PC means) from 31.22% to 16.83% variance accounted for, and a corresponding increase in the Sentence factor (18.83% to 25.28%). The

increase in error (Sentence $\times$ Listener) from 39.83% to 47.42%, with the removal of the highest intelligibility group and its associated mean, also illustrates the interdependent nature of the values of percentage variance accounted for. The variance that was removed with the removal of the highest intelligibility condition caused a “redistribution” of the percentage variance attributable to the various effects. It is interesting that only minor changes in the effects of Listener and Condition $\times$ Sentence were noted.

Discussion—The range of intelligibility in these data is still large by any standard, however, with overall means for conditions ranging from 18.40% for the 20PC group to 59.17% for the 60PC group. Furthermore, mean performance across 100 sentences for individual listeners still ranged from 2.28% to 75.31% in the 60 listeners remaining in the three conditions. This value of 25.28% variance accounted for is consistent with a substantial sentence effect despite considerable error and large range in mean percentage-correct performance among listeners. As illustrated above, however, this strength is dependent on the strength of other factors chosen to vary in the experiment. At this point, it is concluded that the difficulty understanding sentences presented with spectral content restricted to one frequency region is influenced to a substantial extent by the characteristics specific to each sentence token, including its semantic content, its grammatical structure, and the acoustics of the talker's production.

Experiment 2A

In Experiment 1, it was found that large variations in mean intelligibility did not eliminate the sentence effect. It is possible, however, that the strength of the sentence effect found in Experiment 1 is due in part to the narrow frequency content common to the stimuli in the four listening conditions. The purpose of Experiment 2 was to examine the sentence effect with intelligibility held approximately constant and spectral content varied. Two analyses were performed on data from Healy (1999), the first on two listening conditions in which the center frequencies of two narrow-band pairs were selected to have mean intelligibility of approximately 56% (Experiment 2A) and the second on conditions in which the mean intelligibility of the two conditions was near 80% (Experiment 2B).

Method—Narrow speech-modulated bands were created by amplitude-modulating pure tones by frequency-matched 1/3-octave (96 dB/octave) speech. One band pair was created by mixing at equal amplitudes bands centered at 1336 and 1684 Hz (adjacent 1/3-octave bands, labeled *ADJ*) and a second pair consisted of bands centered at 530 and 4200 Hz (having a separation of approximately three octaves, labeled *3-OCT*). As before, separate groups of 20 listeners each were assigned to the two conditions, and each listener heard all 100 CID sentences using procedures identical to those of Experiment 1. The *ADJ* condition produced a mean intelligibility score of 57.28% ($SD = 9.23\%$), and the *3-OCT* condition produced a score of 55.52% ($SD = 10.58\%$).

An ANOVA was performed to generate the mean squares and calculate the variance components for the factors of the model. The model again involved Listening Condition, Sentence, Listener, and interactions as shown in Table 1, this time with the effect of intelligibility of listening condition designed to be negligible. The primary question of Experiment 2 is the size of the Sentence $\times$ Condition interaction. A large effect would indicate that particular sentences become more or less intelligible relative to others in the same list as frequency content is varied.

Results and discussion—The results of the ANOVA are shown in Table 4. The analysis indicated that the main effects of Listener and Sentence, and the interactions, were all significant with $p < .001$. However, the effect of Condition was not significant, reflecting the fact that the overall means of the two listening conditions were selected to be very similar, thus

allowing the effect of two different frequency regions to be examined. More importantly, the variance components showed that the effect of condition had fallen to zero, consistent with the near equality of the means. The sentence effect, freed of variance-sharing with mean intelligibility of listening condition, was a substantial 27.50%. Except for the overall error, which was again large, the sentence effect dominated the contributions of Listener within group (5.45%) and the Condition $\times$ Sentence interaction (7.90%). The small interaction suggests that the rankings of the individual sentences did not differ substantially in the two listening conditions, despite the different frequency regions involved. The sentence effect appears to be stronger than the differential contributions of the widely separated frequency regions. This may potentially be attributed to acoustic cues that are distributed across the spectrum, such as prosody (Grant & Walden, 1996), and to general sentence characteristics such as syntax and word frequency, which have no specific spectral affiliations.

Experiment 2B

An ANOVA was performed to generate the mean squares and calculate the variance components for the factors of the model. The model again involved Listening Condition, Sentence, Listener, and interactions using the data from two listening conditions that resulted in approximately equal but high overall mean intelligibility. Again, the primary question here was the size of the Sentence $\times$ Condition interaction and its contribution relative to the sentence effect.

Method—Two additional amplitude-modulated band pairs were employed using bands centered at 1100 and 2100 Hz (approximately one octave separation, labeled *1-OCT*) and 750 and 3000 Hz (*2-OCT*). Two additional groups of 20 listeners each participated, and all processing and procedures were identical to those used in Experiment 2A. The *1-OCT* condition produced a mean intelligibility score of 82.68% ($SD = 6.38\%$), and the *2-OCT* condition produced a score of 75.31% ($SD = 6.30\%$).

Results and discussion—Table 5 shows the results of this analysis. The findings are entirely consistent with the analysis of the moderate-intelligibility data: The effect of Listening Condition was near zero, the sentence effect was large (27.66%), error was large (60.90%), and the Sentence $\times$ Condition interaction was small (7.14%). The sentence effect, due in part to the consistency of the mean intelligibility of the sentences in the lists, again accounts for considerably more variance than the Sentence $\times$ Condition interaction. These results confirm the robustness of the sentence effect in the presence of disparate spectral information.

It must be noted, however, that the results in Experiments 1 and 2 are based on relative variance components and that there was a large error term. The error term has been ignored to this point, but it obviously exerts an influence on the stability of the rankings of the sentences based on mean intelligibility within each listening condition (see SENT effect, Table 1). The error that was present in each sentence's intelligibility would tend to lower correlations between mean sentence intelligibilities from two different listening conditions. We turn to such correlations in Correlation Analyses 1 and 2 as another measure of the strength of the sentence effect in the presence of substantial experimental error. The use of correlations is also consistent with our definition of the sentence effect as a measure of the stability of the *ranking* of a sentence's intelligibility relative to other sentences in a list under various intelligibility-degrading conditions. That is, because the Pearson product-moment correlation is very sensitive to the ranking of data in corresponding sets of data (McNemar, 1969), it is an appropriate descriptor of the strength of the sentence effect. The extent to which the intelligibilities of individual sentences are correlated across different listeners or different listening conditions reflects the extent to which they retain their intelligibility relative to other sentences in the list, the stability of their rankings, and the strength of the sentence effect.

Correlation Analysis 1: Good and Poor Listeners

It is well known that listeners vary in their ability to recognize speech even under identical listening conditions. This was apparent in the data sets used in Experiments 1 and 2, where mean intelligibility across 100 CID sentences for individual listeners differed by as much as 37% in a given condition (see Table 6 for the ranges of mean recognition scores among listeners in selected conditions). The focus of Correlation Analysis 1 was to test the strength of the sentence effect by comparing the performance of the best and worst listeners within a listening condition, labeled *good* and *poor* groups. Within-condition comparisons were chosen to avoid any possible confounding of sentence effect with listening condition. Correlations of the mean intelligibilities of the 100 CID sentences from the good and the poor listeners within each of the eight listening conditions used in Experiments 1 and 2 were calculated.

Method

First, for any one condition the mean recognition performance for each listener across the 100 sentences was calculated. Then the best 5 and worst 5 listeners were selected and formed two subgroups from the original 20. The mean percentage-correct score for each sentence was then calculated within each subgroup. For example, the good group in the 60PC condition had a mean score of 64% on the sentence “There’s a good ball-game this afternoon,” whereas the poor group scored 20%. Furthermore, the good group’s overall mean was 70.60%, compared to 48.06% for the poor group in that condition. Finally, the two sets of 100 mean sentence scores within each condition were correlated.

Analysis—Selection of conditions—The listening conditions chosen for the analysis of good and poor groups were selected from the eight original conditions on the basis of several criteria designed to include only correlations of high statistical quality. First, the scatter plots of 100 points for each correlation were examined for approximate normality, linearity, and homoscedasticity. Second, subgroups that had more than 10 mean sentence scores of less than 2.5% or greater than 97.5% (floor or ceiling effects) were excluded. Third, listening conditions in which the means of the good and poor listeners differed by less than 15% were eliminated. Finally, the range of individual sentence means within the set of 100 for each subgroup was examined and required to be greater than 80.0 percentage points, and skewness and kurtosis values required to be less than 1.5 in magnitude, in each selected data set. The correlations of the sentence scores for good and poor subgroups in the remaining four conditions—40PC, 60PC, 3-OCT, and ADJ—were then examined. This selection process resulted in using those listening conditions where the sentence effect had the opportunity to manifest itself most fully: data sets for conditions where the overall intelligibility was in the mid-range (40%–60%), the individual sentence means ranged over 80 percentage points, and floor and ceiling effects were minimal. Moreover, these selected conditions still represent a wide range of listener abilities and spectral content.

Results and Discussion

The correlations across 100 CID sentences between the good and poor listeners in the selected conditions were as follows (99% confidence intervals are in parentheses): 40PC, $r = .84$ (.74, .90); 60PC, $r = .71$ (.56, .82); 3-OCT, $r = .76$ (.63, .85); and ADJ, $r = .67$ (.50, .79). The square of these values indicates that the good and poor listeners shared from 44.9% to 70.6% of the variance of the sentence means, with a mean shared variance of 55.9%. It seems clear that more than half the variance may be attributed to the sentence effect in these selected conditions. It thus appears that good and poor listeners (differing by approximately 22% in average performance level) find the same sentences to be generally easy or difficult, regardless of the listening condition. Because these findings are based on a variety of listening conditions with

a range of performance levels and frequency components, this is interpreted as a strong affirmation of the sentence effect at least for these selected, optimal conditions.

Correlation Analysis 2: Listening Conditions and Modalities

The first purpose of Correlation Analysis 2 was to assess the Sentence factor again using correlations, but this time with all listeners included, across selected listening conditions. Finally, a second comparison of the strength of the sentence effect was made between the auditory and visual modes of perception (audition vs. speech reading).

Method

The mean percentage-correct scores for the CID sentences from the eight listening conditions used in Experiments 1 and 2 were again employed in Correlation Analysis 2. In addition, intelligibility data for speech reading CID sentences were available from one published study (Demorest & Bernstein, 1992), which employed 1 male and 1 female talker and 104 observers with normal hearing, and also one unpublished experiment (R. R. Hinkle, personal communication, 1982), who employed 1 female talker and 20 observers with normal hearing. The experimental procedures were generally similar to those of the listening conditions in the present study; however, differences in response collection existed. In Demorest and Bernstein's (1992) study, observers typed responses on a computer keyboard and were scored on the basis of total words correct, whereas Hinkle required observers to hand-write responses scored for key words correct. Demorest and Bernstein's (1992, Appendix B, pp. 889–891) results were recast as percentage total words correct per sentence.

Thus, the effects of degrading intelligibility of sentences through a variety of band-pass filters was compared to the effects of the low intelligibility present in speech reading. High correlations, again, would indicate that the ranking of mean sentence intelligibilities was similar across presentation modes. The effect of using percentage-correct words per sentence as the primary measure of sentence difficulty is to include sentence length in the calculation as a contributor to difficulty. That is, if a long sentence is difficult it may still yield more words (or key words) correct than a short sentence, but as a percentage of total length it will have an appropriately lower score. Although Demorest and Bernstein (1992) used percentage of total words correct per sentence as their measure, whereas the present study used percentage of key words correct, the correlation of total to key words in the CID sentences is .93, so they measure very similar aspects of sentence recognition and were considered acceptable for correlation.

Analysis—Selection of conditions—Strict criteria for inclusion of correlations were again applied to the data sets. The range of mean sentence scores was greater than 80 percentage points, skewness and kurtosis were constrained to be less than 1.5 in magnitude, scatter plot examination was performed, and only data sets having fewer than 10 sentences with scores less than 2.5% or greater than 97.5% were included. The four auditory conditions selected were again 40PC, 60PC, 3-OCT, and ADJ. The two speech-reading conditions were (a) Demorest and Bernstein (1992), male talker (mean intelligibility = 28.53%), and (b) Hinkle (1978), female talker (mean intelligibility = 46.36%).

Results and Discussion

The correlations between pairs of selected listening conditions, with all 20 listeners contributing to the sentence intelligibility means, are shown in Table 7. One can see that the correlations are somewhat lower than in Correlation Analysis 1, with values ranging from .47 to .84. The median correlation, however, was .63, and the mean of the squared correlations accounted for 42.4% of the variance. In general, the auditory comparisons again suggest that

the sentence effect is one of the significant factors in sentence recognition when listening conditions without substantial floor or ceiling effects are considered.

The selected correlations between the auditory and the visual mode of speech perception were quite low, on the other hand, ranging from .22 to .45, with a median of .36 and a mean shared variance of 13.2%. In addition, an overall correlation was calculated between the mean percentage-correct data for each sentence averaged across all the selected listening conditions and the corresponding sentence means across all three speech-reading conditions. This correlation was .41 (approximately 18% of the variance) and also reflects the modest extent to which sentence difficulty (relative to other sentences in the list) remains constant across the auditory and visual modes of perception.

The linguistic aspects of the sentences, of course, are identical in the two modes of presentation, but the physical representation of the phonemes and prosodic characteristics is vastly different. For example, the visual appearance of /p/, /b/, and /m/ is virtually identical, but their acoustic manifestations are very different. It appears that these differences in the phonetic and prosodic representations in the auditory and visual modalities, as well as possible differences in talker productions, prevent the presence of any substantial cross-modal sentence effect. Note that the two selected talkers in the speech-reading experiments were moderately correlated ($r = .74$). This correlation supports the existence of a substantial sentence effect within the speech-reading mode, as shown by Demorest and Bernstein's (1992) and Adams's (2002) generalizability analyses.

General Discussion

It is obvious that the linguistic characteristics of an isolated sentence and their acoustic manifestations through the talker's production determine the intelligibility of that sentence when listening conditions are difficult. Furthermore, it is apparent that sentences differ considerably in their intelligibility under constant listening conditions. The purpose of the experiments and analyses reported here, however, was to assess the extent to which sentences retain their intelligibilities, relative to the other sentences in a list, across a wide range of acoustic manipulations. That is, how strong is the sentence effect in circumstances where overall intelligibility, frequency content, and listener ability varied?

Generalizability analysis (as used here to obtain variance components) and simple Pearson product-moment correlations were used to provide answers to that question. They allow somewhat different questions to be asked of the same data sets. In this study, the generalizability analyses gave a broad picture of the magnitude of the sentence effect when contributions of listener and general intelligibility were included and when spectral content of the sentences was varied. The correlations, on the other hand, to a large extent reflected the similarity of the ranking of mean sentence intelligibilities across conditions.

The primary finding of the two generalizability experiments and two correlation analyses was that the sentence effect is a prominent component of sentence intelligibility, especially when conditions are conducive to its appearance. Both types of analysis supported the presence of the sentence effect. Experiments 1 and 2 showed the effect was strong in the presence of a wide range of intelligibilities and spectral content, respectively. Correlation Analysis 1 indicated that more than 50% of the variance shared between good and poor listeners within a listening condition was attributable to the sentence effect, whereas Correlation Analysis 2 showed an effect greater than 40% across listening conditions. In general, the strength and consistency of the sentence effect across conditions and listeners indicate that the specific set of linguistic characteristics of a sentence and their acoustic manifestations establish, to an important extent, its intelligibility relative to other sentences. However, these acoustic manifestations do not

appear to be restricted in spectral frequency, as indicated by the presence of the sentence effect across conditions in Experiment 2 where spectral content varied.

The large error components in Experiments 1 and 2 remain to be explored, however, as does the unexplained variance in the correlations. Furthermore, the consistency of the relative intelligibility of unrelated sentences is clearly going to be influenced by the set of sentences involved. The magnitude of the sentence effect, or even its presence at all, is dependent on the constraints applied in the generation of the set of sentences. Materials that conform to a narrow range of sentence lengths, like the Harris, Haines, Kelsey, and Clack (1961) revision of the CID sentences, may not be expected to show as much sentence effect as in this study. Similarly, the sentences in the SPIN test (Kalikow et al., 1977), with its scoring of only the final word in the sentence, could not show a true sentence effect but would show a ranking of intelligibilities based largely on the acoustics of the final word and its predictability from the context in the high-predictability list. The original CID sentences, on the other hand, form a much more heterogeneous set that would presumably allow the sentence effect to manifest itself more fully, although with a concomitant increase in error variance. In general, conditions conducive to the sentence effect would appear to include (a) listening conditions that yield overall performance (mean percentage correct across all sentences in the list) in the 40% to 60% range and (b) a variety of sentence lengths and grammatical structures. A study is under way to determine the extent to which the relative sentence intelligibilities are the same when noise produces the listening difficulty compared with the filtering used in the present study.

The cross-modal comparisons in Correlation Analysis 2, in contrast to the listening experiments, indicated that the acoustic properties of the CID sentences did not produce a hierarchy of sentence intelligibilities similar to that derived through speech reading. This can be at least partly explained by the complementary nature of audible and visible speech. As Miller and Nicely (1955) pointed out, the most confusable acoustic information (place of articulation) is the most easily seen on the talker's lips. In addition, the prosodic information, so prominent in the acoustic spectrum of sentences, is virtually absent in speech reading.

The demonstration of the existence and strength of the sentence effect, in those conditions favorable to its appearance, supports the construction of tests of auditory sentence understanding in which the difficulty of sentences is varied through sentence selection rather than by altering sentence intensity adaptively (e.g., Plomp & Mimpen, 1979) or signal-to-noise ratio during test construction (e.g., Kalikow et al., 1977). Indeed, the fact that sentences vary in inherent intelligibility is conducive—in fact, vital—to the development of tests that go beyond the need for equivalent lists and probe more central, cognitive abilities, as dictated by classical test theory or item response theory (McDonald, 1999).

Conclusions

The sentence effect, in which a sentence retains its relative intelligibility within a large set of sentences across a variety of conditions, appears to be strong in listening experiments where the signal is degraded through filtering, strong enough to warrant consideration in test construction. The effect is weaker, however, when comparing listening and speech-reading performance on the same set of sentences. It would potentially be useful, especially if the sentence effect is robust with respect to different sentence materials and distortions other than filtering, to explore in more detail which set of linguistic features, manifested acoustically, contribute to the inherent intelligibility of a particular spoken sentence.

Acknowledgments

This work was supported in part by Grant DC05795 from the National Institute on Deafness and Other Communication Disorders. We thank Marilyn Demorest for helpful comments and Robert Hinkle for providing the speech-reading data used in Correlation Analysis 2.

References

- Adams CF. Blame the message, not the messenger: Probing the difficulty of speechreading sentences. *Dissertation Abstracts International* 2002;63-07A:2414.
- Alpiner, JG.; McCarthy, PA., editors. *Rehabilitative audiology: Children and adults*. Lippincott Williams & Wilkins; Philadelphia: 2000.
- Bell TS, Wilson RH. Sentence recognition materials based on frequency of word use and lexical confusability. *Journal of the American Academy of Audiology* 2001;12:514–522. [PubMed: 11791938]
- Bilger RC, Neutzel J, Rabinowitz WM, Rzezchowski C. Standardization of a test of speech in noise. *Journal of Speech and Hearing Research* 1984;27:32–48. [PubMed: 6717005]
- Cheung H, Kemper S. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics* 1992;13:53–76.
- Clauser RA. The effect of vowel consonant ratio and sentence length on lipreading ability. *American Annals of the Deaf* 1976;121:513–518. [PubMed: 983911]
- Cornfield J, Tukey JW. Average values of mean squares in factorials. *Annals of Mathematical Statistics* 1956;27:907–949.
- Cronbach, LJ.; Gleser, GC.; Nanda, H.; Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. Wiley; New York: 1972.
- Davis, H.; Silverman, SR., editors. *Hearing and deafness*. 4th ed.. Holt, Rinehart & Winston; New York: 1978.
- Demorest ME, Bernstein LE. Sources of variability of speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research* 1992;35:876–891. [PubMed: 1405543]
- Demorest ME, Bernstein LE. Applications of generalizability theory to measurement of individual differences in speech perception. *Journal of the Academy of Rehabilitative Audiology* 1993;26:39–50.
- Di Nocera F, Ferlazzo F, Borghi V. G theory and the reliability of psychophysical measures: A tutorial. *Psychophysiology* 2001;38:796–806. [PubMed: 11577903]
- Dixon, WJ., editor. *BMDP statistical software manual*. University of California Press; Berkeley: 1992.
- Giolas TG, Cooker HS, Duffy JR. The predictability of words in sentences. *Journal of Auditory Research* 1970;10:328–334.
- Giolas TG, Duffy JR. Equivalency of CID and revised CID sentence lists. *Journal of Speech and Hearing Research* 1973;16:549–555. [PubMed: 4783790]
- Grant KW, Walden BE. Spectral distribution of prosodic information. *Journal of Speech and Hearing Research* 1996;39:228–238. [PubMed: 8729913]
- Harris JD, Haines HL, Kelsey AP, Clack TD. The relation between speech intelligibility and electroacoustic characteristics of low fidelity circuitry. *Journal of Auditory Research* 1961;1:357–381.
- Healy EW. A minimum spectral contrast rule for speech recognition: Intelligibility based upon contrasting pairs of narrow-band amplitude patterns. *Dissertation Abstracts International* 1999;59(9-B):5123.
- Healy EW, Warren RM. The role of contrasting temporal amplitude patterns in the perception of speech. *Journal of the Acoustical Society of America* 2003;113:1676–1688. [PubMed: 12656400]
- Hinkle, RR. Unpublished doctoral dissertation. Purdue University; 1978. An investigation of list equivalency for auditory, visual, and auditory-visual performance using revised CID sentences.
- Hood RB, Dixon RF. Physical characteristics of speech rhythm of deaf and normal hearing speakers. *Journal of Communication Disorders* 1969;2:20–28.

- Kalikow DN, Stevens KN, Elliott LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America* 1977;61:1337–1351. [PubMed: 881487]
- Laures JS, Weismer G. The effects of a flattened fundamental frequency contour on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research* 1999;42:1148–1156.
- McDonald, RP. Test theory: A unified treatment. Erlbaum; Mahwah, NJ: 1999.
- McNemar, Q. Psychological statistics. 4th ed.. Wiley; New York: 1969.
- Miller GA, Nicely PE. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 1955;27:338–352.
- O' Brian N, O' Brian S, Packman A, Onslow M. Generalizability theory I: Assessing reliability of observational data in the communication sciences. *Journal of Speech, Language, and Hearing Research* 2003;46:711–717.
- O' Brian S, Packman A, Onslow M, O' Brian N. Generalizability theory II: Application to perceptual scaling of speech naturalness in adults who stutter. *Journal of Speech, Language, and Hearing Research* 2003;46:718–723.
- Plomp R, Mimpfen AM. Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 1979;18:43–52. [PubMed: 760724]
- Rippy JV, Dancer J, Pittenger JB. List equivalency of the CID everyday sentences (Harris revision) under three signal-to-noise ratios. *Ear and Hearing* 1983;4:251–254. [PubMed: 6628850]
- Scarsellone JM. Analysis of observational data in speech and language research using generalizability theory. *Journal of Speech, Language, and Hearing Research* 1998;41:1341–1347.
- Shavelson, RJ.; Webb, NM. Generalizability theory: A primer. Sage; Newbury Park, CA: 1991.
- VanLeeuwen DM, Barnes MD, Pase M. Generalizability theory: A unified approach to assessing the dependability (reliability) of measurements in the health sciences. *Journal of Outcome Measurement* 1998;2:302–325. [PubMed: 9803717]
- Webster JC. Interlist equivalencies for a numeral and a vowel/consonant multiple-choice monosyllabic test for severely/profoundly deaf young adults. *Journal of Auditory Research* 1984;24:17–33. [PubMed: 6543863]
- Wilson S, Dancer J, Stamper J. Visual equivalency of Harris' revised CID everyday sentence lists. *Volta Review* 1984;86:267–273.
- Winer, BJ.; Brown, DR.; Michels, KM. Statistical principles in experimental design. 3rd ed.. McGraw-Hill; New York: 1991.

Table 1

Expected mean squares for the generalizability model used in Experiments 1 and 2.

Source of variability	No. levels	σ^2	Expected value of mean square ^a
COND	a (= 2, 3, or 4)	σ_C^2	$\sigma_e^2 + bc\sigma_C^2 + b\sigma_L^2 + c\sigma_{CS}^2 + \sigma_{SL}^2$
SENT	b (= 100)	σ_S^2	$\sigma_e^2 + ac\sigma_S^2 + \sigma_{SL}^2$
LISN (COND)	c (= 20)	σ_L^2	$\sigma_e^2 + b\sigma_L^2 + \sigma_{SL}^2$
COND $\times$ SENT		σ_{CS}^2	$\sigma_e^2 + c\sigma_{CS}^2 + \sigma_{SL}^2$
SENT $\times$ LISN (COND)		σ_{SL}^2	$\sigma_e^2 + \sigma_{SL}^2$

Note. The model involves the Listening Condition (COND), Sentence (SENT), Listener (LISN), and interactions. Analysis is based on two, three, or four listening conditions, 20 listeners per condition, and 100 Central Institute for the Deaf sentences. The SENT $\times$ LISN term serves as the error term for all factors except COND, which requires the calculation of a quasi F ratio, after Winer, Brown, and Michels (1991, pp. 374–377). Also see Demorest and Bernstein (1992, p. 886). Parentheses indicate nesting of listeners under condition.

^aBased on Cornfield and Tukey's (1956) algorithm as implemented by BMDP 8V (Dixon, 1992).

Table 2

Generalizability analysis of a model involving Listening Condition, Sentence, Listener, and interactions.

Estimates of variance components			
Component in model	Variance	Variance adjusted for fixed effect [*]	Percentage variance accounted for
COND	742.60	556.95 [*]	31.22
SENT	335.96	335.96	18.83
LISN (COND)	53.82	53.82	3.02
COND × SENT	168.56	126.42 [*]	7.09
SENT × LISN (COND)	710.52	710.52	39.83
Total		1,783.67	99.99

Note. The listening conditions produced four mean intelligibilities (approx. 20%, 40%, 60%, and 80%), based on 20 listeners per condition and 100 CID sentences. The dependent variable is the mean percentage of key words correct per sentence per listener.

^{*} adjustment = variance (1 – 1/a).

Table 3

Generalizability analysis of a model involving Listening Condition, Sentence, Listener, and interactions.

Estimates of variance components			
Component in model	Variance	Variance adjusted for fixed effect*	Percentage variance accounted for
COND	411.54	274.50*	16.83
SENT	412.43	412.43	25.28
LISN (COND)	63.51	63.51	3.89
COND × SENT	160.81	107.26*	6.58
SENT × LISN (COND)	773.51	773.51	47.42
Total		1,631.21	100.00

Note. This analysis is based on conditions producing three mean intelligibilities (approximately 20%, 40%, and 60%), 20 listeners per condition, and 100 CID sentences. The dependent variable is the mean percentage of key words correct per sentence per listener.

Table 4

Generalizability analysis of a model involving Listening Condition, Listener, and interactions.

Estimates of variance components			
Component in model	Variance	Variance adjusted for fixed effect*	Percentage variance accounted for
COND	−5.97	(set to 0.0)*	0.00
SENT	448.71	448.71	27.50
LISN (COND)	88.99	88.99	5.45
COND × SENT	257.81	128.91*	7.90
SENT × LISN (COND)	965.36	965.36	59.15
Total		1,631.97	100.00

Note. This analysis is based on two moderately intelligible listening conditions, 20 listeners per condition, and 100 CID sentences. The dependent variable is the mean percentage of key words correct per sentence per listener.

Table 5

Generalizability analysis of a model involving Listening Condition, Listener, and interactions.

Estimates of variance components			
Component in model	Variance	Variance adjusted for fixed effect*	Percentage variance accounted for
COND	23.65	11.83*	1.12
SENT	292.48	292.48	27.66
LISN (COND)	33.67	33.67	3.18
COND $\times$ SENT	151.10	75.55*	7.14
SENT $\times$ LISN (COND)	643.99	643.99	60.90
Total		1,057.52	100.00

Note. This analysis is based on two high-intelligibility listening conditions, 20 listeners per condition, and 100 CID sentences. The dependent variable is the mean percentage of key words correct per sentence per listener.

Table 6

Ranges of mean individual-listener intelligibilities for four selected conditions.

Condition	Range (<i>n</i> = 20)	Good <i>M</i> (<i>n</i> = 5)	Good <i>SD</i>	Poor <i>M</i> (<i>n</i> = 5)	Poor <i>SD</i>
40PC	26.38–50.00	47.89	2.19	31.92	3.54
60PC	42.39–75.31	70.60	3.11	48.06	3.42
3-OCT	37.49–69.82	68.00	1.57	41.48	3.91
ADJ	35.68–72.71	68.57	3.36	46.18	6.79

Note. Also shown are means and standard deviations of the intelligibility scores for the best five (good) and lowest-performing five (poor) listeners. Data are percentage-correct key words across 100 CID sentences. 40PC = 40% intelligibility condition; 60PC = 60% intelligibility condition; 3-OCT = 3 octave separation condition; ADJ = adjacent band condition.

Table 7

Correlations among selected listening (AUD) and speech reading (VIS) conditions based on mean percentage-correct intelligibility scores for 100 CID sentences.

Mode	Condition	1	2	3	4	5	6
AUD	1. 40PC	—					
	2. 60PC	.67 ^a (.50, .79)	—				
	3. 3-OCT	.47 ^a (.24, .65)	.59 ^a (.39, .73)	—			
	4. ADJ	.84 ^a (.74, .90)	.75 ^a (.61, .84)	.51 ^a (.29, .68)	—		
VIS	5. D&B-M	.32 ^a (.07, .53)	.42 ^a (.18, .61)	.28 ^a (.03, .50)	.32 ^a (.07, .53)	—	
	6. Hinkle	.45 ^a (.22, .63)	.43 ^a (.20, .62)	.22 (−.04, .45)	.40 ^a (.16, .59)	.74 ^a (.60, .84)	—

Note. Numbers in parentheses are confidence intervals at p = .005 and .995. D&B-M = male speaker from Demorest and Bernstein (1992); Hinkle = female speaker from Hinkle (1978).

^aSignificantly > 0.0, α = .01.