Research Article

Perceptual Organization of Interrupted Speech and Text

Valeriy Shafiro,^a Daniel Fogerty,^b Kimberly Smith,^c and Stanley Sheft^a

Purpose: Visual recognition of interrupted text may predict speech intelligibility under adverse listening conditions. This study investigated the nature of the linguistic information and perceptual processes underlying this relationship.

Method: To directly compare the perceptual organization of interrupted speech and text, we examined the recognition of spoken and printed sentences interrupted at different rates in 14 adults with normal hearing. The interruption method approximated deletion and retention of rate-specific linguistic information (0.5–64 Hz) in speech by substituting either white space or silent intervals for text or speech in the original sentences.

Results: A similar U-shaped pattern of cross-rate variation in performance was observed in both modalities, with

minima at 2 Hz. However, at the highest and lowest interruption rates, recognition accuracy was greater for text than speech, whereas the reverse was observed at middle rates. An analysis of word duration and the frequency of word sampling across interruption rates suggested that the location of the function minima was influenced by perceptual reconstruction of whole words. Overall, the findings indicate a high degree of similarity in the perceptual organization of interrupted speech and text.

Conclusion: The observed rate-specific variation in the perception of speech and text may potentially affect the degree to which recognition accuracy in one modality is predictive of the other.

Perception of spoken and written language by adults is highly robust and often succeeds despite severe distortions in the input signal or presentation in adverse environments. Findings from early experiments demonstrated that accurate recognition of speech and text can be achieved with less than 50% of the original input preserved (Miller & Friedman, 1957; Miller & Licklider, 1950). More recent research further indicates an association between the perception of distorted speech and text (George et al., 2007; Humes, Kidd, & Lentz, 2013; Smith & Fogerty, 2015), which is at least in part mediated by individual cognitive abilities involved in accessing and manipulating perceptual input (Besser, Koelewijn, Zekveld, Kramer, & Festen, 2013). However, beyond suggestions of “linguistic closure” that seem to link the perception of distorted speech

and text (Besser et al., 2013), the roles of perceptual organization and linguistic structure in recognition of distorted speech and text remain elusive. Diverse linguistic elements, such as words, syllables, or phonemes, combine in systematic ways into the overall linguistic structure of a language. Perceptual processes operating on various elements of linguistic structure to achieve recognition and, ultimately, comprehension of the intended meaning constitute the perceptual organization of speech and text. Distortion of speech and text through interruption may, in turn, interfere with specific perceptual processes, with the extent of interference depending on the rate of interruption. In the current study, periodic interruptions of speech and text were used to directly compare the processing of information in the two modalities.

Initial systematic investigations of the intelligibility of distorted speech and text often assumed a relationship based on the underlying modality-independent linguistic structure (Miller & Friedman, 1957; Müsch & Buus, 2001; cf. Stilp, 2011; Van Rooij & Plomp, 1991; Van Wijngaarden, Steeneken, & Houtgast, 2002). Conceived within the conceptual framework of information theory (C. E. Shannon, 1951), the early studies evaluated the relationship between input and output in terms of stimulus information transmitted and received, without directly addressing questions of the underlying perceptual organization or processing

^aDepartment of Communication Disorders & Sciences, Rush University Medical Center, Chicago, IL

^bDepartment of Communication Sciences & Disorders, University of South Carolina, Columbia

^cDepartment of Speech Pathology & Audiology, University of South Alabama, Mobile

Correspondence to Valeriy Shafiro: valeriy_shafiro@rush.edu

Editor-in-Chief: Frederick (Erick) Gallun

Editor: Lori J. Leibold

Received December 22, 2017

Revision received March 23, 2018

Accepted June 3, 2018

https://doi.org/10.1044/2018_JSLHR-H-17-0477

Disclosure: The authors have declared that no competing interests existed at the time of publication.

units involved. Instead, the primary research focus was on measurable input signal characteristics that affect perceptual outcome, such as proportion of speech or text heard or seen (Mattys & Samuel, 1997).

In later work, the role of perceptual organization in the perception of speech and text, along with possible involvement of phoneme, grapheme, syllable, and word as perceptual processing units, was addressed separately for each modality (Apel & Masterson, 2001; Healy, 1976; Plomp, 2001; Powers & Speaks, 1973). However, no general consensus emerged for either speech or text with respect to perceptual organization of distorted language input (Mattys & Samuel, 1997). For speech, some findings appeared to support the primary role of the syllable (Ghitza, 2013; Greenberg, Carvey, Hitchcock, & Chang, 2003), whereas others suggested words as the primary perceptual units (Cutler, 1994; Samuel 2001), along with the possibility that different elements of linguistic structure vary in dominance depending on the type and degree of distortion.

Recent investigations into the relationship between speech and text processing have focused on the perception of distorted text as a possible predictor of speech intelligibility under adverse listening conditions or in special populations, such as older listeners or listeners with hearing impairment. A significant relationship between the ability to recognize distorted speech and text has been found in some work (Besser, Zekveld, Kramer, Ronnberg, & Festen, 2012; George et al., 2007; Humes et al., 2013; Watson, Qiu, Chamberlain, & Li, 1996; Van Wijngaarden et al., 2002), but not all studies (Besser et al., 2013; Schoof & Rosen, 2014; Zekveld, Rudner, Johnsrude, & Rönnberg, 2013; Zekveld, Rudner, Kramer, Lyzenga, & Rönnberg, 2014). Several well-known factors, including sample size variation, listener age, cognitive and language abilities, and hearing status, have been suggested as potential contributors to the variability of findings across studies. However, one major factor, the perceptual organization of speech and text, somewhat surprisingly has not yet been investigated. If modality-independent elements of linguistic structure, such as words, syllables, or phonemes, are processed within a shared framework of perceptual-cognitive processes, then the strength of their relationship between speech and text may be influenced by the degree to which these elements of linguistic structure are affected by distortion. For example, when distortions obscure whole words and leave remaining words intact, as compared to when the distortions obscure information within phonemes or graphemes about equally for all words, perceptual processes involved in linguistic closure may be quite different.

Both speech and text convey linguistic information, which comprises shared knowledge of lexical items, semantic associations, and syntactic rules. However, the procedural representation of that information can differ between the two modalities and may result in different ways of organizing, integrating, and processing the perceptual input. As an ontologically older means of communication, speech perception is learned without formal instruction. In contrast, text reading has to be explicitly taught in order

to be acquired. Speech is a product of physically lawful (co)articulatory activity captured in the temporally unfolding acoustic signal. It is rich with indexical information about the speaker and environment. In contrast, text is a secondary “invented” form of language representation, consisting, in English, of discrete nonoverlapping characters. It is a spatial signal whose visual forms and their relationships to underlying sound structures are arbitrary across languages. Because of these differences, some have argued that speech and text processing are fundamentally different, with speech processing based on a detailed episodic auditory code, in contrast to the abstract representation of text (Port, 2010).

A number of different methods have been used to study the perception of distorted text. Some stemmed from the letter guessing task (Müsch & Buus, 2001; C. E. Shannon, 1951; Stilp, 2011; Van Rooij & Plomp, 1991; Van Wijngaarden et al., 2002). Others implemented text distortion methods that did not target specific linguistic elements (Humes, Burk, Coughlin, Busey, & Strauser, 2007; Krull, Humes, & Kidd, 2013; Watson et al., 1996), manipulating the overall proportion of text shown. One popular method in the latter category is the Text Reception Threshold (TRT) task (George et al., 2007; Zekveld, George, Kramer, Goverts, & Houtgast, 2007), which has been used in the majority of the studies investigating the relationship between distorted speech and text. Typically in this task, individual words or sentences are shown on the screen with regularly spaced vertical bars masking portions of the displayed text. The main parameter in the TRT task is the proportion of unmasked text, which remains visible to the reader. This is controlled by systematically changing the width of the masker bars to reveal more or less of the target text (see Figure 1). One advantage of this method is that it avoids any a priori assumptions about the nature of linguistic elements that may be affected by this experimental manipulation.

Applications of the TRT task to sentences have varied in a number of important parameters, such as whether the sentences are presented word by word or as a whole, the color of text and masker bars, the timing of text presentation, and the number of masker bars per sentence (Krull et al., 2013; Smith & Fogerty, 2016; Zekveld et al., 2007). Whereas some of these parameters, such as color or display size, are text specific, others have ready analogues to parameters from auditory experiments of speech perception in which portions of the speech signal are periodically interrupted by masking noise or replaced with silence. The duration of missing or retained speech fragments is controlled by the interruption rate, usually given in hertz, whereas the proportion of audible speech in the input signal is typically controlled by the duty cycle, which corresponds to the proportion of speech retained within each interruption period.

The large variation in speech recognition accuracy as a function of interruption rate is a consistent finding from research on interrupted speech. When half of the original speech is interrupted by either silence or masking noise, recognition accuracy by adults with normal hearing is

Figure 1. An example of sentences processed across several interruption rates for text and speech stimuli. The presentation of this sample sentence lasted 3.1 s for both text and speech conditions.

generally around 50% correct for low interruption rates (< 1 Hz) and reaches 90% and above for high interruption rates (> 8 Hz). In the middle rates (1–4 Hz), recognition accuracy often declines relative to both lower and higher interruption rates, giving the cross-rate accuracy function a characteristic U-shape. These differences in cross-rate recognition can be further accentuated by variation in duty cycle (Miller & Licklider, 1950; Shafiro, Sheft, & Risley, 2011). Across interruption rates, individual fragments of retained speech following interruption differ in terms of their linguistic content. At low rates, whole words or large word fragments may be captured, whereas syllable-sized fragments may be available at the middle rates, and at faster rates, each phoneme is generally sampled one or more times.

Depending on the nature of the missing and available linguistic information, different perceptual processes may operate on the different types of linguistic input available at different interruption rates (Shafiro et al., 2011). More cognitively driven top-down processes related to semantic and linguistic information of the signal may operate at lower rates, whereas low-level grouping processes involved in temporal smoothing of the signal may be more engaged at higher interruption rates. It has been suggested that the dip in accuracy that gives the function its characteristic U-shape may represent a discontinuity in the perceptual organization of the speech input—such as a transition between perception based on word-sized fragments with large word-sized gaps to that based on (sub)phoneme-sized fragments that sample every word (Miller & Licklider, 1950).

Similar to interrupted speech, the frequency of spacing of the masking bars in relationship to text in the TRT task may produce text fragments of different length, affecting the nature of available linguistic input (e.g., whole words, syllables, graphemes/phonemes). Indeed, there is some variation in previous studies in terms of the average number of masking bars applied to sentence stimuli. For instance, about 14–16 bars per sentence appear to be used as maskers in Zekveld et al. (2007), whereas almost three times as many are used in Smith and Fogerty (2015). Although the proportion of text in both studies can be manipulated in comparable amounts, the length of the text fragments available for the same proportion of text seems to differ.

A comparison of the 52% condition illustrated in Figure 1 of Zekveld et al. (2007) and the 50% condition illustrated in Figure 2 of Smith and Fogerty (2015) demonstrates that each letter is sampled once or twice in the latter but not in the former case where some letters are completely obscured by the masking bars. Thus, despite both conditions having a comparable total proportion of text, the number of letters sampled is different. These differences in how linguistic input is sampled may potentially play an important role in measures of overall accuracy and affect the strength of the relationship between speech and text across different procedures.

This study investigated the perception of interrupted speech and text in order to determine potential similarities and differences in perceptual organization of the underlying linguistic structure. Specifically, recognition accuracy for spoken and printed sentences was compared directly in terms of the size of preserved and deleted text and speech fragments. If the perception of both interrupted speech and interrupted text depends on the same modality-independent elements of linguistic structure, systematic distortions of these linguistic elements will have a similar effect on recognition accuracy for both speech and text. Alternatively, differences in recognition accuracy between speech and text matched on the amount of preserved and deleted signal information across interruption conditions can indicate modality-specific aspects of their perceptual organization.

Method

Participants

Fourteen young adults with normal hearing (mean age = 23 years, range: 19–29 years) who were enrolled at the University of South Carolina participated for course credit. All participants were native speakers of American English and had audiometric thresholds less than or equal to 20 dB HL at the octave frequencies between 250 and 8000 Hz and normal or corrected-to-normal vision.

Stimuli

Institute of Electrical and Electronics Engineers (IEEE)/Harvard sentences (IEEE, 1969) were used as stimuli for

both the text and speech conditions. The IEEE sentences, each containing five key words, have low contextual predictability and are arranged in phonetically balanced lists. Ninety sentences were used in the auditory conditions, and 80 different sentences were used in the text conditions. Testing in each modality consisted of presenting sentences interrupted by silence or white space, for speech and text, respectively. Audio recordings were spoken by a single male talker, who was a native speaker of American English (Loizou, 2007). Figure 1 displays example stimuli in each modality.

Speech stimuli in the auditory modality were created using periodic interruption at 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz. The duty cycle was fixed at 50% (i.e., half of each period was replaced by silence). The starting phase of the interruption cycle was randomized for each sentence. Interruption windows used a 2-ms raised cosine on/off ramp to minimize the introduction of transients. Following interruption, each auditory stimulus was low-pass filtered at 2000 Hz in MATLAB using an 80-order, linear-phase finite impulse response filter. A control condition used low-pass filtered speech without interruption.¹

Text stimuli used for presentation in the visual modality were created to be analogous to speech stimuli. To establish a correspondence between text and speech in terms of time taken in milliseconds to space taken in pixels, the total horizontal size of the 1,600 × 200 pixel visual window was taken to correspond to 4 s. A spoken sentence with the median sentence duration was then chosen as the calibration sentence. The duration of this spoken sentence was expressed in proportion to 4 s. This proportion was then used to select the text font size that resulted in the same proportion of pixels in the horizontal dimension (re: total visual window) “taken up” by the text of the sentence, including white space between characters. Use of the Arial font with a red font color on a white background was selected to approximate Zekveld et al. (2007). The font size was 22.5 points, which calibrated the pixel display. These font settings were used on the remaining sentences.

To examine the conversion from speech to text, the 90 sentences used in the auditory condition were compared in both modalities. The average ratio of sentence duration in milliseconds to the 4-s reference time frame was 0.63 ($SD = 0.07$). For these same sentences converted to text, the average ratio of sentence length in pixels to the total horizontal display area was 0.62 ($SD = 0.07$). A similar analysis of the 80 sentences used in the text condition revealed nearly identical values ($M = 0.63$, $SD = 0.07$). Such close correspondence in ratios suggests that the current

procedure of converting spoken sentence duration from temporal to spatial parameters represents the interrupted sentences comparably across the two modalities. However, it is important to note that, for English, the correspondence between the duration of a phoneme and its orthographic representation in pixels is not linear. Furthermore, due to necessary approximation used in the conversion process and because different sentences were used for the visual and auditory conditions, the millisecond-to-pixel correspondence was less precise at the level of individual sentences. This is illustrated in Figure 2, which displays a scatter plot of individual sentence lengths in milliseconds and in pixels for the 90 sentences used in the auditory conditions. The scatter indicates that sentence lengths in pixels did either underestimate or overestimate the auditory duration of many individual sentences. Nevertheless, the median deviation of these sentences when measured in pixels from the values predicted using the conversion formula remained quite small, that is, 3 pixels, suggesting that the current conversion method captures the overall characteristics of sentence length quite well.

To be consistent with the interruption of speech by silence in the auditory modality, interruption of the text stimuli was created by using white masking bars. The on and off times in milliseconds were scaled to pixels by the factor of 0.4 (1,600 pixels/4,000 ms) based on the same reference window duration of 4 s. The text sentences were presented in their entirety, rather than one word at a time, to be consistent with typical reading processes. Interruption rates were calibrated between auditory and visual materials such that the text image was interrupted the same number of times across the image as an auditory version of the sentence would be interrupted given the duration of the spoken sentence. The text of each sentence was also presented on the screen for the duration of the auditory recording of the sentence followed by replacement of the text with a fixation point.

Figure 2. Scatter plot of the lengths of the 90 individual sentences used in the auditory modality when represented in pixels (abscissa) and milliseconds (ordinate). The solid line is a linear regression of the data with an R^2 of .25.

¹Low-pass filtering was used to decrease the information redundancy of speech in comparison to text. The specific low-pass cutoff frequency was selected based on our pilot research and previous work (Sheft et al., 2010) indicating that, in young listeners with normal hearing, the shape of the cross-rate intelligibility function for interrupted speech tends to change from continuously rising to a U-shape following low-pass filtering.

Design and Procedure

A total of 16 experimental conditions were completed, consisting of eight interruption rates (i.e., 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) for each of the two modalities (i.e., auditory and visual). A final auditory control condition was also tested that presented uninterrupted sentences low-pass filtered at 2000 Hz. Stimuli were blocked by condition with 10 unique sentences per block. Sentences were not used more than once during an experimental run. The presentation order of modality was counterbalanced among participants, and the test order of the different interruption rates within each modality was randomized for each participant. Immediately prior to each experimental condition, participants completed a short training that consisted of five Hearing in Noise Test (Nilsson, Soli, & Sullivan, 1994) sentences processed according to the current experimental condition. This training, without feedback, was immediately followed by the experimental testing, which consisted of a single test block of 10 sentences.

Participants were seated in a sound-attenuating booth for all testing. Speech stimuli were presented monaurally to the right ear for all participants via a Sennheiser HD 280 Pro headphone. Speech stimuli were calibrated using a Larson Davis 800B sound level meter with linear weighting to a level of 70 dB SPL relative to the average level of the sentences prior to interruption. Participants responded using a MATLAB custom response interface displayed on a 17 in. ELO touchscreen display. The interface displayed the text (or, for auditory presentation, the word *listen*) during stimulus presentation followed by *respond*. A button also appeared following the presentation to allow participants to continue to the next test item when ready. For all conditions, participants were instructed to respond verbally by repeating the sentence they either heard or read. Participants were encouraged to guess. Responses were audio-recorded for offline scoring and analysis by trained raters. A strict scoring procedure required participants to repeat each key word exactly (i.e., no missing or extra suffixes). However, key words could be repeated in any order. Each test block consisted of a total of 50 key words.

Results

The performance functions for recognition of speech and text at each interruption rate are shown in Figure 3. Speech intelligibility for low-pass filtered sentences presented without any interruptions in the control condition was 96.8% correct ($SD = 0.78$), indicating that stimuli were highly intelligible for this group of listeners despite filtering. Prior to further data analyses, all percent correct scores were transformed to rationalized arcsine units to stabilize the error variance (Studebaker, 1985). An initial 2 (modality) $\times$ 8 (interruption rate) repeated-measures analysis of variance demonstrated a significant main effect of interruption rate, $F(7, 91) = 256.6$, $p < .001$, $\eta^2 = .95$. No significant main effect of modality was observed ($p > .05$), but modality did significantly interact with interruption

Figure 3. Proportion of words correct for recognition of 50% preserved speech and text stimuli. Error bars represent 1 SEM.

rate, $F(7, 91) = 21.4$, $p < .001$, $\eta^2 = .62$. Paired-samples comparisons between the two modalities demonstrated significant differences at the most extreme interruption rates (0.5 and 64 Hz, $p < .05$ with a Holm-Bonferroni correction) where visual text processing resulted in better performance. This trend reversed to favor auditory performance at the intermediary interruption rates of 2, 4, 8, and 16 Hz but reached significance only for 8 Hz (with a Holm-Bonferroni correction). Overall, there was remarkable similarity in the shape of the performance function across interruption rates for the two modalities. Particularly notable and consistent with the prior research on interrupted speech (Bhargava, Gaudrain, & Baskent, 2016; Huggins, 1975; Nelson & Jin, 2004; Powers & Speaks, 1973; Saija, Akyürek, Andringa, & Başkent, 2014; Shafiro et al., 2011; Shafiro, Sheft, & Risley, 2016) was the dip in performance for both modalities at 2 Hz.

The consistency in the shape of the rate-accuracy function for speech and text suggests that subjects tended to rely on similar kinds of information in both modalities. A further comparison of the length of on/off intervals across rates in relationship to word length was consistent with the notion that performance at each rate was strongly influenced by rate-specific sampling of individual words (Miller & Licklider, 1950). Specifically, the average word length for speech and text was 321 ms and 126 horizontal pixels, respectively. Both are smaller than 1000 ms/400 pixels for 0.5 Hz and 500 ms/200 pixels for 1 Hz for speech and text, respectively. Thus, at these low interruption rates, the stimulus-on intervals tended to capture word-sized fragments, which, on average, could provide sufficient information to identify individual words. On the other hand, the same-size, stimulus-off intervals were too large for the missing words to be “filled in,” resulting in responses that consisted mostly of the words conveyed by the stimulus on intervals, as described below. In contrast, with 2-Hz interruption, the stimulus-on intervals on average captured only single incomplete fragments of individual words, resulting

in a sequence of word fragments with few, if any, complete words. This may account for the prominent dip in the rate–accuracy functions for both speech and text. Finally, at faster interruption rates of 4 Hz and higher, the stimulus-on intervals of speech and text provided, on average, multiple samples of each word, resulting in a continuously rising intelligibility function.

The dependence of these two stimulus factors—average number of whole words per each stimulus-on interval and average number of samples per word—on interruption rate is illustrated in the left panel of Figure 4. The dashed–dotted function, which corresponds to the left ordinate axis of the figure, represents the proportion of the average word captured within a single on interval (i.e., the ratio of the average word duration to stimulus-on interval). The dashed function, which utilizes the right ordinate axis, represents the number of times the average word is sampled, which is the ratio of the average word duration to the whole interruption period (the on and off intervals together). The intersection of the two functions coincides with the rate at which an accuracy minimum is found in the subject data for both speech and text recognition.

The right panel of Figure 4 illustrates a simple approach to predicting intelligibility using these two stimulus factors that characterize interrupted speech or text. As argued above, the average number of words per ontime determines performance at low interruption rates, and information integration across the multiple interruptions per word sets recognition at higher rates. Consequently, to estimate the information conveyed by each stimulus factor in

isolation, the stimulus functions of the left panel were scaled in the right panel so that the words-per-ontime function (the dashed–dotted function) roughly coincided with recognition ability at the lowest interruption rates, whereas the interruptions-per-word function (the dashed function) approximated higher rate performance. An exponential factor was added to the interruptions-per-word function to give the logistic function asymptote at high rates that characterizes subject data. Shown with a solid line, the final function of the right panel combines through probability summation the predictions of the two stimulus–factor functions. This two-factor prediction of recognition accuracy follows the basic characteristics of the subject data of Figure 3. The open circles in the right panel of Figure 4 show the mean subject performance averaged across speech and text conditions at each interruption rate. The goodness of fit, as estimated by R^2 , between the probability summation function and the mean subject data was 0.98, with a root-mean-square deviation of 0.03. With adjustment of model scaling factors, an R^2 of .98 was also obtained when separately fitting either the text or speech data. Recognition accuracy predicted by the two stimulus–factor functions thus illustrates that the ratio of the length of individual word fragments captured in the stimulus-on interval to the average word duration, along with the frequency of whole word sampling, can provide a basis for the overall shape of the rate–accuracy function for both speech and text.

To further explore the relationship between rate-specific speech and text processing, a correlation analysis was conducted on accuracy scores across different interruption

Figure 4. Illustration of the two aspects of word sampling following stimulus interruption. In the left panel, the number of words per ontime is indicated by the dashed–dotted function and left ordinate, whereas the number of interruptions per word is shown by the dashed function and right ordinate. The right panel shows intelligibility predicted by each of the stimulus factors alone, along with performance predicted when combining both sources of information. Additionally, mean subject performance averaged across text and speech conditions at each interruption rate is shown in the right panel with open circles.

rates between the two modalities. Contrary to expectations, the analysis revealed mostly weak correlation magnitudes and no discernable association patterns across rates. Only 11 out of 72 correlational pairs reached magnitudes above 0.4, with the highest value of 0.64. None of these correlations were significant at $p < .05$ after a Holm-Bonferroni correction. Given the overall close correspondence of cross-rate performance accuracy of speech and text, the lack of stronger associations among rates is surprising. However, this may be due to the procedural constraints of this study. A relatively small sample of 14 young subjects with normal hearing seems to have produced a compressed performance range in several interruption conditions (illustrated by the very small error bars in Figure 2), which could have confounded response magnitude and underpowered the analysis.

Alternatively, it is also possible that individual performance in interrupted speech and text conditions was affected by a set of modality-specific factors, thus reducing the correlation between speech and text performance within the same participants. If the effect of such modality-specific factors is sufficiently small compared with the effect of modality-general word-based perceptual processes, it can result in a similar mean performance shown by the overall cross-rate function shapes for speech and text, without producing associations in rate-specific accuracy scores for any specific condition. Although the reasons for the lack of correlations between speech and text are unclear, this finding indicates that the relationship between perceptual processing of speech and text is likely susceptible to a variety of both modality-general and modality-specific factors that can affect perceptual outcome.

Discussion

This study investigated recognition accuracy of interrupted speech and text to determine similarities and differences between modalities in the underlying perceptual organization. Speech and text conditions were matched to approximate the stimulus information transmitted at each rate in each modality. Between the auditory and visual modalities, results demonstrated a remarkable similarity in the overall recognition patterns across interruption rates, as well as several modality-specific differences. For both speech and text, a characteristic U-shaped function with a minimum around 2 Hz was observed, suggesting a fundamentally similar underlying perceptual organization. This was confirmed by the absence of a significant main effect of modality in statistical analysis. On the other hand, the significant interaction between rate and modality, as well as several rate-specific differences, indicated variation in how information was processed in each modality at different interruption rates.

The direction of the modality-specific effects in recognition accuracy for speech and text changed across interruption rates. Text recognition appeared to be significantly more robust than speech recognition at the lowest and the highest rates. This finding may not be surprising for the highest rate of 64 Hz in which periodic regular interruptions

effectively introduce an audible continuous distortion that may contribute to masking perceptually salient cues in the speech signal. This negative effect at 64 Hz for speech, however, would be absent for text masked by white bars, which have the same color as the background. On the other hand, the basis for the advantage of text over speech is less clear for the lowest rate of 0.5 Hz.

One possibility for the advantage of text over speech may arise from greater word preservation at a low interruption rate. However, this was not the case. Analysis of the text and speech stimuli with a 0.5-Hz interruption rate demonstrated that there were an equal number of 25 key words preserved for both text and speech. In this analysis, a key word was defined as *present* if 50% or more of the phonemes/graphemes were preserved and, *absent*, if fewer than 50% of the graphemes/phonemes were preserved. Partial graphemes and phonemes were treated as not preserved. The analysis thus indicated that the difference in accuracy was not due to more key words available in one modality.

A second possibility is that the instantaneous spatial presentation of all word-sized fragments of a full-text sentence in this study may have led to more effective use of contextual semantic and syntactic information for text than for spoken sentences. Although the total stimulus presentation time was the same for speech and text, for speech, stimulus fragments were unfolding in a temporally ordered sequence, with information about the last word not available until the end of stimulus presentation. This natural delay in the availability of all relevant fragments until the end of the stimulus in speech may impede the use of contextual information, leading to a relatively lower accuracy for speech than text.

To explore this possibility, we examined the ability of participants to use preserved context to fill in missing key words that were not preserved in the stimulus. For key words counted as present in speech, the average recognition accuracy was 60.6% ($SD = 7.0$), whereas for key words counted as absent, it was 3.7% ($SD = 4.3$). For key words counted as present in text, the average recognition accuracy was 82.9% ($SD = 15.3$), whereas for key words counted as absent, it was 1.4% ($SD = 0.7$). The low accuracy for absent key words for both speech and text suggests that participants were not using semantic or syntactic context to fill in words that were missing more than half of their phonemes or graphemes. Rather, accuracy in these conditions was based on the speech or text fragments for words that had at least half or more of their phonemes or graphemes preserved. It thus appears that, for text in the 0.5-Hz condition, participants were more effective in reconstructing whole words based on partial information than with speech. The instantaneous spatial presentation of all word-sized fragments of a full sentence in text may have granted additional time for top-down processes to operate on partial text. That is, participants had the entire sentence duration to process a few degraded key words. In addition to more processing time devoted to, for example, lexical search, the full sentence presentation may have also reduced

memory demands. For speech, however, key words were presented only as the sentence temporally unfolded, with half of the sentence duration consisting of the silent interruption intervals. Thus, in addition to a more limited processing time to resolve each key word, participants might also have had to deal with a greater memory load, which could have contributed to reduced performance in speech for the 0.5-Hz condition.

The superior accuracy of text over speech recognition at the lowest and the highest interruption rates is reversed at 8 Hz, where speech, even severely filtered, is recognized significantly more accurately than text. The speech advantage at this rate may be related to differences in the distribution of auditory and visual cues in speech and text, respectively. Acoustic cues for speech are continuous and overlapping in the stimulus due to coarticulation. Coarticulation can thus augment processes of perceptual integration and result in potentially more robust perceptual representations of phonemes, syllables, and words. Prosodic cues may also facilitate integration across interrupted speech segments. In contrast, printed text presents discrete non-overlapping visual cues that specify only a single character at a time. Such visual cues may be easily obscured at the intermediate interruption rates but are well preserved at very slow or fast interruption rates where individual letters are either presented without distortion or sampled multiple times, making them easier to resolve. Thus, although perception of whole words appears to be an important factor that influences the location of the minima of the rate-accuracy functions for speech and text, modality-specific differences in processing need to be taken into account when considering effect at a given interruption rate.

Furthermore, given the bidirectional nature of modality-specific effects across interruption rates, it is possible that the strength of the relationship between speech and text was affected by the type and parameters of interruption applied. In other words, the processes involved in linguistic closure may be different depending on the characteristics of linguistic information preserved and missing in the input signal. For instance, linguistic closure for missing whole words based on semantic and syntactic context may involve a different set of processes than linguistic closure for missing phonemes or graphemes. Based on the current findings, it appears that rate-specific text effects should be considered in future studies of the predictive relationship between text and speech because different interruption rates may engage somewhat different perceptual processes. This conclusion does not minimize the role of cognitive abilities, such as working memory, but rather suggests other contributors that may affect the relationship between speech and text.

This study only investigated the effect of interruption rate on the perception of speech and text in the English language. It is unknown how factors such as average word duration, average number of times words are sampled, or fragment duration may influence the overall accuracy of interrupted speech and text in other languages. However, based on a preliminary evaluation of previous findings, it appears that word duration may play a similar role in other

alphabet-based languages, at least for speech. For instance, in a study by Saija et al. (2014), young and older adults with normal hearing heard Dutch sentences interrupted by silence or noise. The sentences were presented at the natural speaking rate, temporally compressed by 2.0 or temporally expanded by 0.5. For naturally spoken sentences interrupted with silence, a midrate dip was observed at 2.5 Hz, as could be seen in the top panel of Figure 3 of Saija et al. (2014). After the sentences were compressed to one half of the original duration, the interruption rate at which a dip was observed doubled to 5 Hz. In contrast, after the sentences were expanded to twice the original duration, the interruption rate for the dip was halved and observed at the rate of 1.25 Hz.

The word duration for the sentences used in Saija et al.'s study, reported by Clarke, Gaudrain, Chatterjee, and Baskent (2014), was 325 ms, which is quite close to the 321-ms word duration in the current work. It is notable that, based on word duration, both the location of the dip and the change in the dip location following temporal compression and expansion are consistent with the predictions of the current two-parameter model. That is, decreasing the average word duration by half would result in a greater proportion of whole words covered by ontime intervals and reduce the number of times whole words are sampled, effectively doubling the predicted rate for the function minimum. On the other hand, increasing the word duration would result in a smaller proportion of word duration covered by ontime intervals and increase the number of times whole words are sampled, effectively reducing the predicted interruption rate for the function minimum by half.

Such close correspondence between model predictions and independently obtained results is encouraging, but additional prospective research with speech and text under additional interruption conditions is necessary to test predictions of the current two-parameter model. It is possible that, for different languages or with different interruption parameters such as duty cycle, or use of noise rather than silence for speech interruptions, the roles played by other elements of the linguistic structure, such as syllables or phonemes, may differ. Furthermore, for logographic languages where a word or a phrase can be represented by a single character, a different cross-rate accuracy pattern may emerge. The current findings can provide initial predictions for the shape of the rate-accuracy function based on word length in either modality, which can be further refined with additional experimental data.

Overall, the similarity in the rate-accuracy function for speech and text suggests a similar underlying perceptual organization. However, it is important to recall that, after interruptions were applied to speech and text, the speech was also low-pass filtered at 2000 Hz, whereas no additional processing was applied to the text. Without filtering of the speech, a marked dip in the rate-accuracy function would be unlikely in the current study. Our pilot investigation for this study and our previous work (Shafiro et al., 2011; Sheft, Shafiro, & Kuvadia, 2010) indicates that, in young adults with normal hearing who are native

English speakers, a subject group similar to the current participant sample, the midrate dip in the rate–accuracy function for interrupted sentences does not typically appear without filtering or addition of distortions or configurations that reduce stimulus information. Prior research demonstrates a midrate dip with more challenging stimuli, such as multitalker sentence materials (Nelson & Jin, 2004), single words or unrelated word strings (Miller & Licklider, 1950; Shafiro et al., 2011), temporally compressed or vocoded speech (Nelson & Jin, 2004; Saija et al., 2014; Shafiro et al., 2016), nonnative (Smiljanic, Sheft, Chandrasekaran, & Shafiro, 2016), or listeners with hearing impairment (Shafiro, Sheft, Risley, & Gygi, 2015). The consistency of the dip in intelligibility around 2–4 Hz in speech across different types of stimulus distortions and listener populations suggests that similar underlying perceptual processes are engaged. On the other hand, the absence of the dip under more favorable stimulus conditions or when stimuli are presented to more skilled listening populations likely indicates the ability to compensate for reduced information in the midrate region, a view consistent with the present word-based two-parameter model.

It is remarkable that even with a large number of salient acoustic cues removed by filtering and interruptions, listeners were able to achieve nearly perfect accuracy when filtered speech was presented without interruptions and moderately high accuracy at high interruption rates of 32–64 Hz. This is consistent with previous reports of extremely robust speech perception with sparse spectral and temporal cues (Remez, Rubin, Pisoni, & Carrell, 1981; R. V. Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) and further confirms the conjecture that, as a linguistic communication signal, speech is generally considerably more robust than text (Plomp, 2001). In sum, for speech, limitations in the ability to perceptually integrate across temporal interruptions, particularly at midrates, only become apparent when speech redundancy is reduced, such as by the spectral filtering used in the current study. On the other hand, for text, the dip at midrates is present even without any additional distortion of the input.

An important factor that has not been addressed in the current study is the type of interruption used with speech and text. Previous research has demonstrated greater intelligibility of sentences interrupted by noise bursts than by silent gaps unless energetic masking prevents audibility of retained speech fragments (Nelson & Jin, 2004; Saija et al., 2014). Although the mechanism underlying the differences in intelligibility is not yet well understood for speech (Repp, 1992; Warren, 1999), the intelligibility advantage of noise over silence is robust (Nelson & Jin, 2004; Saija et al., 2014; Warren, 1999; Powers & Wilcox, 1977). Past research has further shown that the performance benefit of noise over speech is rate dependent with a greater magnitude in midrates, where minima in performance are typically found (Nelson & Jin, 2004; Saija et al., 2014). In the current study, speech was interrupted by silence, whereas text was interrupted by white vertical bars of the same color as the background. It is possible that using noise

and text-colored visual vertical bars may enhance compensatory perceptual processes and lead to a greater perceptual continuity and a greater fusion of linguistic elements into meaningful wholes (Saija et al., 2014; Warren, 1999). Future research can test whether performance across interruption rates for both speech and text will be affected similarly if noise and text-colored bars are used with the present test materials. Performance metrics other than intelligibility, such as perceived signal continuity, reaction time, eye tracking, and electrophysiological responses, may be utilized to determine perceptual processes engaged at different interruption rates and associated neurophysiological mechanisms.

Summary

Converging evidence indicates that recognition of interrupted text can predict recognition of speech under adverse listening conditions. This study investigated factors that may mediate this relationship. Specifically, recognition accuracy for the two different types of linguistic input, speech and text, was compared directly in terms of stimulus information across a range of interruption rates. Although numerous modality and presentation-specific differences may affect how information is encoded and processed, results indicated a close correspondence in the shape of the accuracy functions across interruption rate for the two modalities. Furthermore, the analysis of word durations in either milliseconds or pixels in relationship to on/off stimulus intervals across interruption rate indicates that cross-rate recognition accuracy for speech and text can be predicted based on the average duration of whole words. Therefore, the results suggest that words serve as key elements of linguistic structure within a fundamentally similar perceptual organization for both speech and text. Nevertheless, differences in performance with speech and text stimuli at several specific interruption rates also suggest the influence of modality-specific factors that affect how linguistic information is parsed and integrated. Further research is needed to elucidate the roles of additional general and modality-specific factors in perceptual organization of speech and text that underlies their relationship in languages other than English and under a broader set of interruption conditions and listener populations.

Acknowledgments

Support was provided in part by the National Institute on Deafness and Other Communication Disorders Grant R01-DC015465, awarded to Daniel Fogerty. Irraj Iftikhar provided research assistance supported in part by the South Carolina Honors College Science Undergraduate Research Funding Program. We are grateful to Deniz Başkent for the Dutch language information and references.

References

- Apel, K., & Masterson, J. J. (2001). Theory-guided spelling assessment and intervention: A case study. *Language, Speech, and Hearing Services in Schools*, 32(3), 182–195.

- Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2013). How linguistic closure and verbal working memory relate to speech recognition in noise—A review. *Trends in Amplification*, 17(2), 75–93. <https://doi.org/10.1177/1084713813495459>
- Besser, J., Zekveld, A. A., Kramer, S. E., Ronnberg, J., & Festen, J. M. (2012). New measures of masked text recognition in relation to speech-in-noise perception and their associations with age and cognitive abilities. *Journal of Speech, Language, and Hearing Research*, 55(1), 194–209. [https://doi.org/10.1044/1092-4388\(2011/11-0008\)](https://doi.org/10.1044/1092-4388(2011/11-0008))
- Bhargava, P., Gaudrain, E., & Başkent, D. (2016). The intelligibility of interrupted speech: Cochlear implant users and normal hearing listeners. *Journal of the Association for Research in Otolaryngology*, 17(5), 475–491. <https://doi.org/10.1007/s10162-016-0565-9>
- Clarke, J., Gaudrain, E., Chatterjee, M., & Başkent, D. (2014). T ain't the way you say it, it's what you say—Perceptual continuity of voice and top-down restoration of speech. *Hearing Research*, 315, 80–87.
- Cutler, A. (1994). The perception of rhythm in language. *Cognition*, 50, 79–81.
- George, E. L., Zekveld, A. A., Kramer, S. E., Goverts, S. T., Festen, J. M., & Houtgast, T. (2007). Auditory and nonauditory factors affecting speech reception in noise by older listeners. *The Journal of the Acoustical Society of America*, 121(4), 2362–2375. <https://doi.org/10.1121/1.2642072>
- Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, 4, 138. <https://doi.org/10.3389/fpsyg.2013.00138>
- Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—A syllable-centric perspective. *Journal of Phonetics*, 31(3–4), 465–485. <https://doi.org/10.1016/j.wocn.2003.09.005>
- Healy, A. F. (1976). Detection errors on the word the: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance*, 2(2), 235–242. <https://doi.org/10.1037/0096-1523.2.2.235>
- Huggins, A. W. F. (1975). Temporally segmented speech. *Perception Psychophysics*, 18(2), 149–157. <https://doi.org/10.3758/BF03204103>
- Humes, L. E., Burk, M. H., Coughlin, M. P., Busey, T. A., & Strauser, L. E. (2007). Auditory speech recognition and visual text recognition in younger and older adults: Similarities and differences between modalities and the effects of presentation rate. *Journal of Speech, Language, and Hearing Research*, 50(2), 283–303. [https://doi.org/10.1044/1092-4388\(2007/021\)](https://doi.org/10.1044/1092-4388(2007/021))
- Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience*, 7(55), 1–16. <https://doi.org/10.3389/fnsys.2013.00055>
- Institute of Electrical and Electronics Engineers. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3), 225–246.
- Krull, V., Humes, L. E., & Kidd, G. R. (2013). Reconstructing wholes from parts. *Ear and Hearing*, 34(2), e14–e23. <https://doi.org/10.1097/aud.0b013e31826d0c27>
- Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. Boca Raton, FL: CRC Press.
- Mattys, S. L., & Samuel, A. G. (1997). How lexical stress affects speech segmentation and interactivity: Evidence from the migration paradigm. *Journal of Memory and Language*, 36(1), 87–116. <https://doi.org/10.1006/jmla.1996.2472>
- Miller, G. A., & Friedman, E. A. (1957). The reconstruction of mutilated English texts. *Information and Control*, 1(1), 38–55. [https://doi.org/10.1016/s0019-9958\(57\)90061-x](https://doi.org/10.1016/s0019-9958(57)90061-x)
- Miller, G. A., & Licklider, J. C. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 22(2), 167–173. <https://doi.org/10.1121/1.1906584>
- Müsch, H., & Buus, S. (2001). Using statistical decision theory to predict speech intelligibility. I. Model structure. *The Journal of the Acoustical Society of America*, 109(6), 2896–2909. <https://doi.org/10.1121/1.1371971>
- Nelson, P. B., & Jin, S. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, 115(5), 2286–2294. <https://doi.org/10.1121/1.1703538>
- Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2), 1085–1099. <https://doi.org/10.1121/1.408469>
- Plomp, R. (2001). *The intelligent ear: On the nature of sound and perception*. Mahwah, NJ: Erlbaum.
- Port, R. F. (2010). Rich memory and distributed phonology. *Language Sciences*, 32(1), 43–55.
- Powers, G. L., & Wilcox, J. C. (1977). Intelligibility of temporally interrupted speech with and without intervening noise. *The Journal of the Acoustical Society of America*, 61, 195–199. <https://doi.org/10.1121/1.1978171>
- Powers, G. L., & Speaks, C. (1973). Intelligibility of temporally interrupted speech. *The Journal of the Acoustical Society of America*, 54, 661–667.
- Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science*, 212(4497), 947–949. <https://doi.org/10.1126/science.7233191>
- Repp, B. (1992). Perceptual restoration of a “missing” speech sound: Auditory induction or illusion? *Perception & Psychophysics*, 51(1), 14–32.
- Saija, J. D., Akyürek, E. G., Andringa, T. C., & Başkent, D. (2014). Perceptual restoration of degraded speech is preserved with advancing age. *Journal of the Association for Research in Otolaryngology: JARO*, 15, 139–148.
- Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12(4), 348–351.
- Schoof, T., & Rosen, S. (2014). The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners. *Frontiers in Aging Neuroscience*, 6(307), 1–16. <https://doi.org/10.3389/fnagi.2014.00307>
- Shafiro, V., Sheft, S., & Risley, R. (2011). Perception of interrupted speech: Effects of dual-rate gating on the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 130(4), 2076–2087. <https://doi.org/10.1121/1.3631629>
- Shafiro, V., Sheft, S., & Risley, R. (2016). The intelligibility of interrupted and temporally altered speech: Effects of context, age, and hearing loss. *The Journal of the Acoustical Society of America*, 139(1), 455–465. <https://doi.org/10.1121/1.4939891>
- Shafiro, V., Sheft, S., Risley, R., & Gygi, B. (2015). Effects of age and hearing loss on the intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 137, 745–756. <https://doi.org/10.1121/1.4906275>
- Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell Labs Technical Journal*, 30(1), 50–64. <https://doi.org/10.1002/j.1538-7305.1951.tb01366.x>

- Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234), 303–304. <https://doi.org/10.1126/science.270.5234.303>
- Sheft, S., Shafiro, V., Kuvadia, S., & Lorenzi, C. (2010). Intelligibility of temporally interrupted dual-band speech. *The Journal of the Acoustical Society of America*, 128(4), 2320. <https://doi.org/10.1121/1.3508179>
- Smiljanic, R., Sheft, S., Chandrasekaran, B., & Shafiro, V. (2016). Effect of language experience on the intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 139, 2189. <https://doi.org/10.1121/1.4950518>
- Smith, K. G., & Fogerty, D. (2015). Integration of partial information within and across modalities: Contributions to spoken and written sentence recognition. *Journal of Speech, Language, and Hearing Research*, 58(6), 1805–1817. https://doi.org/10.1044/2015_JSLHR-H-14-0272
- Smith, K. G., & Fogerty, D. (2016). Integration of partial information for spoken and written sentence recognition by older listeners. *The Journal of the Acoustical Society of America*, 139(6), EL240–EL245. <https://doi.org/10.1121/1.4954634>
- Stilp, C. E. (2011). The redundancy of phonemes in sentential context. *The Journal of the Acoustical Society of America*, 130(5), EL323–EL328. <https://doi.org/10.1121/1.3645966>
- Studebaker, G. A. (1985). A rationalized arcsine transform. *Journal of Speech and Hearing Research*, 28(3), 455–462. <https://doi.org/10.1044/jshr.2803.455>
- Van Rooij, J. C. G. M., & Plomp, R. (1991). The effect of linguistic entropy on speech perception in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, 90, 2985–2991. <https://doi.org/10.1121/1.401772>
- Van Wijngaarden, S. J., Steeneken, H. J., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America*, 111(4), 1906–1916. <https://doi.org/10.1121/1.1456928>
- Warren, R. M. (1999). *Auditory perception: A new analysis and synthesis* (pp. 134–154). Cambridge, United Kingdom: Cambridge University Press.
- Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *The Journal of the Acoustical Society of America*, 100(2), 1153–1162. <https://doi.org/10.1121/1.416300>
- Zekveld, A. A., George, E. L., Kramer, S. E., Goverts, S. T., & Houtgast, T. (2007). The development of the Text Reception Threshold test: A visual analogue of the Speech Reception Threshold test. *Journal of Speech, Language, and Hearing Research*, 50(3), 576–584. [https://doi.org/10.1044/1092-4388\(2007\)040](https://doi.org/10.1044/1092-4388(2007)040)
- Zekveld, A. A., Rudner, M., Johnsrude, I. S., & Rönnberg, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*, 134(3), 2225–2234. <https://doi.org/10.1121/1.4817926>
- Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience*, 8(88), 1–11. <https://doi.org/10.3389/fnins.2014.00088>