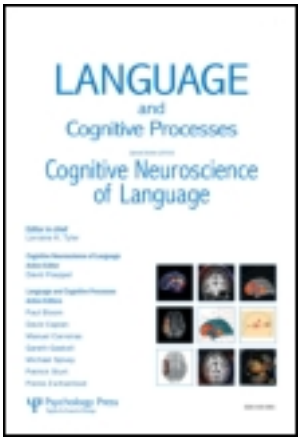

Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp20

The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues

Payam Ezzatian [a], Liang Li [b], M. Kathleen Pichora-Fuller [a c] & Bruce A. Schneider [a]

[a] Department of Psychology, University of Toronto, Mississauga, ON, Canada

[b] Department of Psychology, Peking University, Beijing, China

[c] Toronto Rehabilitation Institute, Mississauga, ON, Canada

Version of record first published: 25 Oct 2011.

To cite this article: Payam Ezzatian , Liang Li , M. Kathleen Pichora-Fuller & Bruce A. Schneider (2012): The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues, Language and Cognitive Processes, 27:7-8, 1056-1088

To link to this article: http://dx.doi.org/10.1080/01690965.2011.591934

LANGUAGE AND COGNITIVE PROCESSES
2012, 27 (7/8), 1056–1088

Psychology Press
Taylor & Francis Group

The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues

Payam Ezzatian[1], Liang Li[2], M. Kathleen Pichora-Fuller[1,3], and Bruce A. Schneider[1]

[1]Department of Psychology, University of Toronto, Mississauga, ON, Canada
[2]Department of Psychology, Peking University, Beijing, China
[3]Toronto Rehabilitation Institute, Mississauga, ON, Canada

To examine the effect of energetic and informational masking on the time-course of stream segregation, we presented listeners with semantically anomalous but syntactically correct target sentences (e.g., "A *house* should *dash* to the *bowl*") that were masked by a two-talker speech masker or steady-state noise masker. To determine the effect of each masker on the time-course of stream segregation, we measured performance as a function of keyword position (key words in italics). The results from Experiment 1 showed that performance improved as a function of keyword position under speech masking, but was relatively stable across keyword positions under noise masking. The results of subsequent experiments showed that the variation in performance across keywords under speech masking was primarily due to the vocal similarities between the competing talkers, and that interference from the semantic content of the masker played a secondary role in undermining performance. Taken together, these results indicate that stream segregation takes longer to build up when a speech target is masked by other speech in the absence of cues that aid stream segregation (e.g., spatial separation), but that it

Correspondence should be addressed to Bruce Schneider, Department of Psychology, University of Toronto, 3359 Mississauga Road North, Mississauga, ON, Canada L5L 1C6. E-mail: bruce.schneider@utoronto.ca

This research was supported by the Canadian Institutes of Health Research [MOP-15359] [STP-53875], Natural Sciences and Engineering Research Council of Canada [RGPIN 138472 and RGPIN 9956] and the National Natural Science Foundation of China [30711120563].

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2011.591934

takes little time to build up when a speech target is masked by a noise or when cues that aid stream segregation are available to listeners.

Keywords: Information masking; Speech masking; Streaming; Perceptual segregation.

Most everyday speech communication takes place in the presence of other sound sources. Therefore, before a listener can comprehend a spoken message, he or she must first detect the target speech signal and extract it from the mixture of competing sounds. Any competing source that is spectrally similar to a speech target can interfere with its processing at the auditory periphery such that there is energetic masking of the target speech. If energetic masking does not completely prevent the perception of the target speech signal, the listener must still perceptually segregate the target from the competing background sources before processing its linguistic content and comprehending the intended message. The ease with which a listener can segregate competing streams depends in part on the perceptual similarities between the target and competing streams. The more dissimilar the competing streams are, the easier it will be to perceptually separate them (Moore & Gockel, 2002). For example, although listening to a speech signal in the presence of construction or traffic noise can interfere with the detection of the target signal, it should not pose a great problem for stream segregation. Since construction or traffic noises are qualitatively dissimilar to speech sounds, it is highly unlikely that a listener will confuse the sound of a construction drill or a car engine with that of a human voice. This is not the case, however, when the interfering sounds are also speech. Competing speech sounds, in addition to energetically masking the target signal, can easily be confused with the target speech and interfere with its segregation from the background. This is especially true when the voices of the competing talkers are similar, such as when the talkers have the same gender, age, accent, and so on. In such cases, the shared acoustical characteristics of the concurrent speech streams can make it difficult for the listener to separate the target signal from the irrelevant competitors and keep the streams separate as the target sentences unfold. Failure to achieve or maintain stream segregation may also allow the semantic content of the irrelevant streams to interfere with the processing and comprehension of the target stream. The negative impact of speech maskers on spoken language comprehension above and beyond that due to energetic masking is commonly referred to as informational masking of speech by speech (e.g., Carhart, Tillman, & Greetis, 1969; Freyman, Helfer, McCall, & Clifton, 1999; Li, Daneman, Qi, & Schneider, 2004; Mattys, Brooks, & Cooke, 2009; Schneider, Li, & Daneman, 2007; Watson, 2005).

Differences in release from masking for speech and noise maskers

It is likely that the detrimental effect of acoustically cluttered environments on spoken language comprehension is in part related to how such environments undermine stream segregation. Evidence in support of this claim stems primarily from studies of cues that alleviate the interfering effects of background competitors on spoken language comprehension. In the majority of such studies, the cues that improve understanding are those that either introduce an actual spatial separation between the target and competing streams (real or perceived), or those that exaggerate the qualitative differences between the competing streams using manipulations that are likely to make stream segregation easier to achieve, e.g., by using talkers of different genders (Brungart, Simpson, Ericson, & Scott, 2001; Darwin, Brungart, & Simpson, 2003; Humes, Lee, & Coughlin, 2006) or languages (Calandruccio, Dhar, & Bradlow, 2010; Freyman, Balakrishnan, & Helfer, 2001; Garcia Lecumberi & Cooke, 2006; Van Engen & Bradlow, 2007). More importantly, the improvement in performance that results from the availability of the aforementioned cues has been shown to be significantly larger in situations in which target and masking streams are more likely to be confused, such as when the maskers contain speech, as opposed to situations in which target and masking streams are less likely to be confused, such as when the maskers are noise. For example, a number of studies examining the release from masking due to real or perceived spatial separation between target speech and background competitors have shown that the release from masking as a result of spatial separation is much larger for maskers consisting of other talkers than for noise maskers (Arbogast, Mason, & Kidd, 2002; Brungart & Simpson, 2002; Brungart et al., 2001; Ezzatian, Avivi, & Schneider, 2010; Freyman, Balakrishnan, & Helfer, 2001, 2004; Freyman, Helfer, & Balakrishnan, 2007; Freyman et al., 1999; Helfer and Freyman, 2005, 2008; Kidd, Arbogast, Mason, & Gallun, 2005; Li et al., 2004; Singh, Pichora-Fuller, & Schneider, 2008). The greater release of a speech target from speech masking when there is spatial separation of the sources likely reflects the improved segregation of the target stream from the background competitors. In contrast, since it is easier to distinguish a speech target from a noise masker than from a speech masker, spatial separation provides significantly less release from a noise masker.

The effect of speech masking on the time-course of stream segregation

There is some evidence to suggest that the effectiveness of speech maskers in disrupting spoken language comprehension is in part due to their effects on

stream segregation. What remains unclear, however, is exactly how speech maskers interfere with stream segregation. After all, while it is true that the presence of other talkers in the background can make it difficult to understand a target talker, it certainly does not prevent comprehension from occurring altogether. A common finding is that speech maskers are more detrimental to comprehension when they share acoustic and linguistic similarities with the target speech. In such situations, the high degree of similarity between the target and background streams may make it initially difficult for the auditory system to differentiate the competing streams. Hence, the auditory system may require more time to differentiate competing streams when they are similar, such as when segregating a speech target from a speech competitor, as opposed to segregating a speech target from a nonspeech competitor. Therefore, a potential way in which speech maskers may interfere with stream segregation is by lengthening the time required for stream segregation to occur.

Delay in the buildup of stream segregation has been well documented in "sequential streaming" studies, in which the perception of a sequence of two alternating tones can switch between a single coherent stream and two distinct streams, depending on the temporal and spectral proximity of the two tones (Bregman, 1990; Bregman & Campbell, 1971; Carlyon, Cusack, Foxton, & Robertson, 2001; Carlyon, Plack, Fantini, & Cusack, 2003; Cusack, Deeks, Aikman, & Carlyon, 2004; Miller & Heise, 1950; Snyder, Alain, & Picton, 2006; Sussman, Horvath, Winkler, & Orr, 2007; Van Noorden, 1975). Using a sequential streaming paradigm, Bregman (1978) showed that listeners perceptually segregate a sequence of tones into two distinct streams only after some time has passed during which the auditory system presumably accumulates evidence from the auditory input for the presence of multiple interleaved streams as opposed to just a single stream. The higher the degree of similarity between the alternating tones, the longer it takes for the auditory system to achieve the perception of two distinct streams (Anstis & Saida, 1985; Beauvios & Meddis, 1997; Bregman, 1978; Rogers & Bregman, 1993).

It is likely that a similar process takes place when a listener is faced with the task of parsing a complex auditory scene, such as a noisy or multi-talker environment. That is, when listening to a target sound in the presence of simultaneous competing sound sources, the amount of time it takes for the auditory system to segregate the target sound from the background competitors is likely to depend on the degree of similarity between the competing sources and the target sound. Therefore, when listening to speech in a noisy environment, it should take longer for stream segregation to build up when the competing sources are speech than when they are noise.

A delay in the buildup of stream segregation should act independently of energetic masking to interfere with speech processing. As a result, the

information at the beginning of a sentence could be incompletely processed in comparison to information later in a sentence even though energetic masking is equivalent over the entire sentence. Furthermore, recent evidence from electrophysiological studies of sequential streaming suggests that the streaming mechanism can reset itself after brief periods of silence or even quick switches in attention, requiring the perception of separate streams to build up again (Carlyon et al., 2001; Carlyon et al., 2003; Cusack et al., 2004). If these principles apply to concurrent stream segregation, then multi-talker environments should be especially detrimental to spoken language comprehension because in such situations stream segregation should take longer to build up, and periods of silence in the target stream or other distractions could easily cause the streaming mechanism to reset, resulting in the perception of the target stream being lost.

The present study

Although the buildup of stream segregation has received a fair amount of attention when different stimulus streams are interleaved, we know of no previous research that has examined this process when speech is presented concurrently with a masker. The purpose of the current series of experiments was to evaluate whether stream segregation takes longer to build up when target speech is masked by other speech as opposed to when it is masked by noise. We were also interested in understanding how acoustic and semantic differences between target and masking streams interact to influence the time-course of stream segregation. This was accomplished by collecting new data and by reanalysing existing data from previous experiments. Data presented in Experiments 1 and 3 of the current study were obtained from Ezzatian, Li, Pichora-Fuller, and Schneider (2011), from the younger participants in Experiments 3 and 4 of that study (no prime conditions only). Data presented in Experiment 2 were obtained from Ezzatian et al. (2010, from the native English-speaking participants in the precedence-effect conditions only). Data presented in Experiments 4 and 5 are novel and were obtained expressly for the purposes of the current study.

EXPERIMENT 1: THE EFFECT OF SPEECH AND NOISE MASKING ON THE TIME-COURSE OF STREAM SEGREGATION

In Experiment 1, we reanalysed data collected by Ezzatian et al. (2011) to examine the effect of speech and noise masking on the time-course of stream segregation. The target stimuli were short, semantically anomalous but syntactically correct English sentences, and the maskers were a steady-state noise masker and a two-talker speech masker that was also composed of

semantically anomalous sentences. To evaluate the differential effect of the two maskers on the time-course of stream segregation, we examined performance as a function of keyword position in the target sentences. We hypothesised that if the speech of the masking talkers does indeed delay the buildup of stream segregation more than the masking noise, then recognition performance for the words that appear earlier in the target sentences should be worse under the speech-masking condition than the noise-masking condition.

Methods

Participants

The 16 participants[1] (mean age = 20.67, $SD = 2.09$ years) were those from Experiment 3 of Ezzatian and colleagues. These participants were students at the University of Toronto–Mississauga who spoke English as a first language and had normal hearing in both ears (see Ezzatian et al., 2011, for details).

Materials and apparatus

Two hundred and eight anomalous sentences, each containing three keywords (e.g., "A *frog* will *arrest* the *pit*.") were used as target sentences (key words in italics). These were the same sentences originally developed and recorded by Helfer (1997), Freyman et al. (1999), and were spoken by a female talker.

Two maskers were used: a speech masker and a speech-spectrum noise masker. The speech masker, which was the same as that used in Freyman, Balakrishnan, and Helfer (2001), consisted of the speech of two females (different individuals than the target talker) uttering anomalous sentences not used as target sentences. The long-term average spectrum of this two-talker speech was approximately the same as the long-term average spectrum of speech of the target talker. This masker, which consisted of a 40-second long string of 35 anomalous sentences, was continuously looped throughout a session.[2] The noise masker was a continuous speech-spectrum noise recorded from an Interacoustic AC5 audiometer (Assens, Denmark). The

[1] Previous studies have indicated that robust effects using these stimuli in this informational masking paradigm can be obtained with as few as 12–16 participants (e.g., Li, Daneman, Qi, & Schneider, 2004, 12 subjects; Freyman, Balakrishnan, & Helfer, 2004, 16 subjects per condition in the priming conditions). Hence, we were confident that 16 participants would be sufficient to explore the effects of word position in these experiments.

[2] The repetitious nature of the speech masker may have allowed participants to become familiar with this masker over the course of the experiment. Consequently, it is possible that performance might improve over time due to the participants becoming familiar with the masker.

noise masker was approximately 360 seconds in duration and was also played in a loop. The long-term average spectrum of the target-talker's sentences had relatively more energy in the high-frequency region (>5 kHz) than did the speech-spectrum noise.[3] All stimuli were digitised at 20 kHz using custom software and a 16-bit Tucker Davis Technologies (TDT, Gainesville, FL) System II. The stimuli were converted to analog using the TDT under the control of an Optiplex GX1 Dell computer. All stimuli were then low-pass filtered at 10 kHz to remove high-frequency artifacts produced by the digitisation process, amplified by a Harmon Kardon amplifier (HK 3370), and presented over a single 40-watt loudspeaker (Electro-Medical Instruments Co., Mississauga, Ontario, Canada) in an Industrial Acoustic Company (Bronx, NY, USA) double-walled sound-attenuating booth. To minimise spatial cues that would aid stream segregation, both the target and masking stimuli were presented from the same loudspeaker. Participants faced the loudspeaker located at a distance of 3.4 feet at 0° azimuth.

Four lists of 17 sentences each (4 practice and 13 target sentences) were presented under each masking condition. Target sentences were always presented at 60 dBA, but masker levels were adjusted to create four signal-to-noise ratios (SNRs), –12, –8, –4, and 0 dB SNR. The SNR levels were computed in the same way as by Li et al. (2004); the RMS value was calculated after pauses in the speech signal longer than 100 ms were removed. The SNRs remained constant throughout the presentation of a single sentence list, but varied randomly across lists. The order in which the sentences were presented was counter-balanced across the masker and SNR combinations.

In each trial, the onset of the masker preceded the onset of the target sentence by exactly 1 second, but continued until the end of the target stimulus. Participants were required to repeat back the entire target sentence after each presentation. The responses of the participants were audio-recorded. The experimenter scored which keywords in each target sentence were repeated correctly. To check on the accuracy of the experimenter's scoring, an independent rater, who was not as familiar with the stimulus materials as was the experimenter, also scored the performance of two randomly chosen participants. The two raters agreed on 97% of the items. The data presented here are based on the experimenter's coding of the responses because of his greater experience with the stimulus materials.

[3] This spectral difference between the two types of maskers indicates that the amount of energetic masking produced by the two types of maskers may have differed.

Results

As an extension of the original analysis by Ezzatian et al. (2011), which was conducted only for the utterance-final keyword, in the present analysis the percentage of correctly repeated words at each keyword position was computed for each participant under each SNR by Masker combination. Each participant's average correct performance was used to compute the dB SNR corresponding to 50%-correct performance (threshold) for each keyword position (see Figure 1). These thresholds were computed by fitting logistic psychometric functions of the form:

$$y = \frac{1}{1 + e^{-\sigma(x-\mu)}}$$

of the data.[4] Psychometric functions were computed by minimising Chi-Square (see Yang et al., 2007 for a more detailed description).

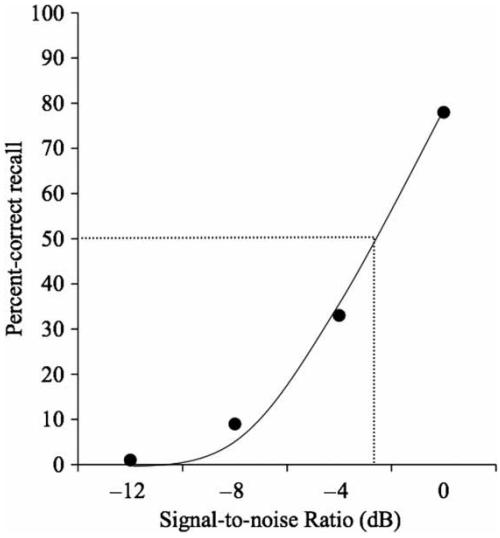

Figure 1. An illustration of how the 50% threshold was determined for the first keyword when the background was a steady-state noise. First the percentage of times that the first keyword was correctly recognised was plotted as a function of SNR. Second, a logistic psychometric function was fit to the data (see text). Third, the SNR corresponding to 50% correct was determined from this psychometric function.

[4] y represents the probability of correctly identifying a keyword, x is the SNR in dB, μ represents the dB SNR level corresponding to 50%-correct performance, and σ determines the slope of the fitted function.

Thresholds computed from the individual psychometric functions are plotted in Figure 2 as a function of masker and keyword position. As can be seen in Figure 2, average performance is better when the target sentences are masked by steady-state noise than when they are masked by two-talker speech. Moreover, Figure 2 shows that when the background masker is noise, average performance is better at the first keyword position, but drops off slightly by positions 2 and 3. However, when the background consists of two-talker speech, average performance is poorer at the first keyword position, but almost equals performance for the noise masker by the third keyword.

To confirm this pattern of results, individual thresholds were entered into a repeated-measures Analysis of Variance (ANOVA) with Masker (Noise, Speech) and Keyword Position (1, 2, 3) as within-subject factors. The ANOVA revealed a significant effect of Masker on thresholds; $F(1, 15) = 7.49$, $MSE = 5.34$, $p = .015$. On average, thresholds were 1.29 dB SNR lower (better) when the masker was speech-spectrum noise than when it was two-talker speech.

The main effect of Keyword Position did not reach statistical significance, $F(2, 30) = 2.69$, $MSE = 3.83$, $p = .084$, but there was a

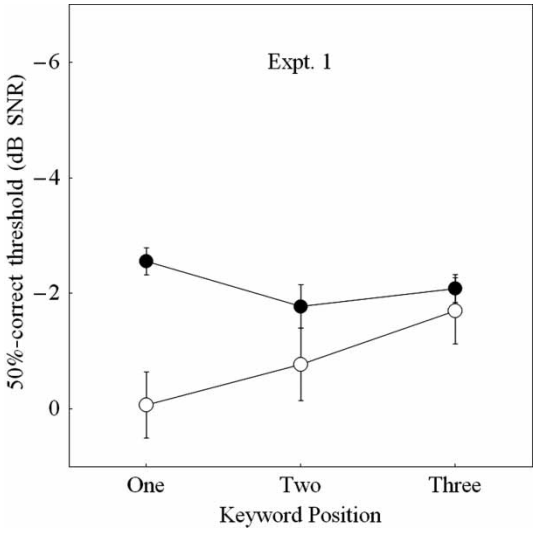

Figure 2. Average thresholds in dB SNR as a function of Keyword Position for the participants in Experiment 1 (data from Ezzatian, Li, Pichora-Fuller, & Schneider, 2011). Open circles represent the condition in which target sentences were masked by a two-talker speech masker. Solid circles represent the condition in which target sentences were masked by a speech-spectrum noise masker. Error bars represent the standard error of the means.

significant Keyword Position by Masker interaction, $F(2, 30) = 5.81$, $MSE = 9.33$, $p = .007$. This interaction was analysed using a Student Neuman Kuels (SNK) test of multiple comparisons, which showed that while thresholds were statistically equivalent at each keyword position when the masker was speech-spectrum noise, the average threshold for the third keyword was significantly lower (better) than that for the first ($p < .01$) and second keywords ($p < .05$) when the masker was two-talker speech. Thresholds for the first and second keywords were statistically equivalent under the speech-masking condition.

Discussion

Word recognition performance for keywords in short anomalous sentences was examined according to the position of the word in the utterance to determine whether there is any evidence that masking a speech stream with other similar speech streams will result in a delay in stream segregation, whereas masking a speech stream with a noise will not. Performance was poorer at the first and second keyword positions when the masker was speech, but improved to the level of the noise masker by the third and final keyword. When the masker was noise, there was no change in performance as a function of keyword position.

If the variation in performance as a function of keyword position in the speech-masking condition is in fact related to differences in the buildup of stream segregation, then the findings suggest that stream segregation takes longer to build up when competing streams share strong similarities, such as when they are all made up of speech streams spoken by female talkers, but occurs without much delay when the competing streams share few similarities such as when they are made up of a speech stream and a noise stream. Nevertheless, it is necessary to rule out other factors before concluding that the differential influence of background masker on keyword position in Experiment 1 reflects a delay in the time-course of stream segregation when the masker consists of speech. It is possible that differential performance as a function of keyword position has little to do with the time-course of stream segregation, but is related to differences in the acoustic characteristics of the target keywords themselves. It is also possible that the word frequencies of the keywords appearing later in the sentences were higher than those of the earlier keywords. Howes (1957) tested the relationship between word frequency (occurrence in written text) and word intelligibility in noise and found a 4.5 dB SNR drop per logarithmic unit of word frequency. Therefore, to investigate potential differences in word frequency, the textbook frequency of occurrence of each keyword in the target sentences was retrieved from two different online databases, the first based on the frequencies found by Thorndike

and Lorge (1944), and the second on the subtitle (SUBTL) word frequencies compiled by Brysbaert and New (2009) who found these frequencies to be a better predictor of word processing time than those based on the Kučera and Francis (cited in Brysbaert & New, 2009) and CELEX norms. Word frequencies computed using either database did not vary significantly over word position, $F(2, 621) = 2.409$, $MSE = 114809898.7$, $p = .091$ for Thorndike-Lorge and $F(2, 621) = 1.001$, $MSE = 1710989.8$, $p = .368$ for SUBTL.

It is also possible, although unlikely, that the keywords at the end of the target sentences have higher intensities, which could have led to their being perceived more easily in the noisy background than the sentence-initial words. To evaluate this possibility, a sample of 208 sentences from the total available pool of anomalous sentences used in the current study was analysed using Praat software to extract the intensity values for each of the three keywords in each sentence. These intensity values were entered into a one-way ANOVA with word position as the factor. This ANOVA did reveal a significant effect of keyword position on intensity, $F(2, 621) = 44.457$, $MSE = 13.301$, $p < .001$. The SNK analysis on word intensity revealed that on average, words in the final keyword position were of lower intensity than words in the first two keyword positions ($p < .05$), but that the words in the first two positions did not significantly differ from each other with respect to intensity. The results of this analysis show that neither increasing word intensity nor word frequency can account for the findings of Experiment 1. It is worth noting that differences in the characteristics of the target keywords themselves are an unlikely culprit for the results because such differences would need to affect performance differently for the two masking sounds to explain the pattern of results. Hence, variations in stimulus characteristics seem to be an unlikely explanation for the findings in this experiment.

If the findings are related to a delay in the ability of the auditory system to separate the co-located target and masking streams when they are similar (i.e., when they both consist of speech), then this delay might be much less pronounced or even disappear when the target and speech-masking streams originate from different spatial locations (i.e., when spatial separation provides an additional cue to stream segregation). On the contrary, if the findings are not due to a delay in the buildup of stream segregation, then introducing a spatial separation between the target and masking streams might produce the same pattern of results as when the target and masking streams were presented from the same spatial location. These hypotheses are examined in Experiment 2, with a reanalysis of data obtained from Ezzatian et al. (2010).

EXPERIMENT 2: THE EFFECT OF SPEECH AND NOISE MASKING ON THE TIME-COURSE OF STREAM SEGREGATION USING THE PRECEDENCE EFFECT

In Experiment 2, we reanalysed data collected by Ezzatian et al. (2010), in which the same procedures and materials were used as in Experiment 1, but with the target and masking stimuli coming from different apparent spatial locations. We hypothesised that if the effects observed in Experiment 1 indeed reflect a delay in the ability to segregate concurrent speech streams that are co-located, then this delay should be reduced when the target and speech-masking stimuli appear to emanate from different spatial locations, since the perceived spatial separation between the target and masking stimuli should make it much easier to achieve stream segregation. In contrast, under noise masking, since the noise and speech streams are already easy to separate from each other, introducing a perceived spatial separation should not have a significant effect on how long it takes for the auditory system to separate these competing streams.

Methods

Participants

The data in Experiment 2 were collected from 16 younger adults from the University of Toronto at Mississauga (mean age = 21.69, $SD = 1.74$ years) who had no previous exposure to the materials used in this experiment. As in Experiment 1, these participants spoke English as a first language and had normal hearing in both ears. For further details see Ezzatian et al. (2010).

Materials, apparatus, and procedure

All materials, equipment, and procedures used in Experiment 2 were the same as those used in Experiment 1. The only difference between the two experiments was in the perceived location of the target and masking stimuli. In Experiment 1, the target and speech masker were played over a single loudspeaker and were therefore perceived as originating from the same spatial location. However, in Experiment 2, a perceived spatial separation was introduced between the target and masking stimuli using the precedence effect (Broadbent, 1954; Freyman et al., 1999; Li et al., 2004; Wallach, Newman, & Rosenzweig, 1949).

Following the procedures outlined by Li et al. (2004), we presented both the target and masking stimuli from two loudspeakers separated 45° to the right and left of the listener. In both masking conditions, the presentation of the target sentences from the loudspeaker on the right led their presentation from the loudspeaker on the left by 3 ms, whereas the presentation of the

masking sounds from the loudspeaker on the left led their presentation from the loudspeaker on the right by 3 ms. This manipulation created the illusion that the target sounds originate from the loudspeaker on the right and that the masking sounds originate from the left loudspeaker, despite the fact that target and masking sounds were played over both loudspeakers at all times. This perceived spatial separation between the target and speech stimuli is achieved without affecting the SNR at either ear to any significant extent (see Li et al., 2004). The effect is quite robust and persists even when the listener is made aware of the illusion created by the precedence effect.

Results

The dB SNR corresponding to 50%-correct recall for each keyword was estimated from individual psychometric functions in the same manner as in Experiment 1. The average of these thresholds is displayed in Figure 3 (upper left) as a function of keyword position and the type of masker. Two things are immediately evident from this plot. First, thresholds are lower (better) when target sentences are masked by two-talker speech than when they are masked by speech-spectrum noise. Second, there does not seem to be much variation in word recognition performance as a function of keyword position under either type of masking. Figure 3 (upper left) shows that when the masker is noise, the pattern of word recognition performance does not seem to differ much from Experiment 1. However, when the masker is speech, performance on earlier keywords is much closer to the final keyword than was the case in Experiment 1.

To evaluate the findings displayed in Figure 3 (upper left), individual thresholds were entered into a Masker by Keyword Position repeated-measures ANOVA. Whenever the assumption of sphericity was violated in this analysis, values from the Greenhouse-Geisser corrected tests were used. The ANOVA revealed a significant main effect of Masker on thresholds, $F(1, 15) = 17.89$, $MSE = 64.75$, $p = .001$. On average, thresholds improved by 1.6 dB when the masker was two-talker speech compared to when it was speech-spectrum noise.

The main effect of Keyword Position was marginally significant, $F(2, 30) = 3.25$, $MSE = 1.22$, $p = .053$. This effect was driven by the average difference in thresholds between the first and second keywords, with the average thresholds being slightly better for the first keyword regardless of the type of masker (mean difference = 0.34 dB, $p = .058$). The interaction between Masker and Keyword Position did not reach statistical significance, $F(1.326, 30) = 1.96$, $MSE = 2.70$, $p = .176$.

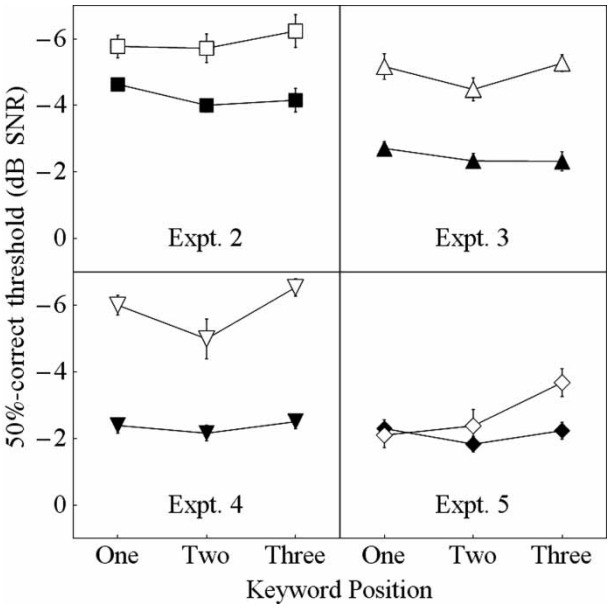

Figure 3. Average thresholds in dB SNR as a function of Keyword Position for the participants in Experiments 2, 3, 4, and 5. The filled symbols in each panel represent the condition in which the target sentences were masked by a speech-spectrum noise masker. In Experiment 2 (data from Ezzatian, Avivi, & Schneider, 2010) there was a perceived spatial separation between target and masker. In the remaining three experiments, the target sentences and maskers were co-located. The open symbols in Experiments 2, 3, 4, and 5 represent the respective conditions in which masker was two-talker speech (open squares, Experiment 2), 3-band noise-vocoded two-talker speech masker (open upward pointing triangles, Experiment 3, data from Ezzatian, Li, Pichora-Fuller, & Schneider, 2011), 16-band noise-vocoded two-talker speech masker (open downward pointing triangles, Experiment 4), and reversed two-talker speech (open diamonds, Experiment 5). Error bars represent the standard error of the means.

Discussion

In Experiment 1 we found that masking a speech stream by other similar speech streams results in poorer performance on the earlier keywords of the target speech, whereas masking a speech stream with a nonspeech noise does not. We hypothesised that if this finding is related to a delay in the buildup of stream segregation, then introducing a perceived spatial separation between the target and masking stimuli should reduce this effect by making stream segregation easier to achieve. In Experiment 2, we confirmed this hypothesis by examining word recognition performance as a function of word position when there was a perceived spatial separation between the same target and masking sounds as had been tested in Experiment 1. Indeed, when there was a perceived separation between the

target sentences and the speech masker, the variation in performance as a function of keyword position that was observed in Experiment 1 was no longer present. In contrast, the pattern of performance across keywords was virtually identical across the two experiments when the masker was noise, indicating that when the target and masking sounds appeared to originate from different spatial locations, the time-course of stream segregation was the same regardless of whether the masking sounds were the speech of other talkers or a continuous noise.

EXPERIMENT 3: THE EFFECT OF PROSODY ON THE TIME-COURSE OF STREAM SEGREGATION

In Experiment 3 we investigated whether or not the envelope fluctuations of the masker sentences, in the absence of voicing cues or semantic content, were sufficient to produce the word-position effect observed in Experiment 1. In Experiment 3, we reanalysed data collected by Ezzatian et al. (2011) in an experiment in which the speech masker was noise-vocoded using three frequency bands (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). In noise vocoding, the amplitude envelope of the signal is determined for certain frequency bands and then it is used to modulate a noise within the same bands. This procedure removes the fine structure of the speech signal, hence making it qualitatively unlike human speech. However, depending on the number of bands used in vocoding, the content of the speech signal can sometimes be easily understood. In a pilot experiment conducted by Ezzatian et al. (2011), it was estimated that using three bands to vocode the target sentences (see Table 1) yields correct identification of about 13% of the keywords in the target sentences when these sentences are presented to listeners without any background interference. In Experiment 3, the speech masker was vocoded using three bands to provide a reasonable, albeit conservative, compromise between minimising linguistic content while preserving as much of the amplitude envelope information in the signal as possible. Specifically, we expected that if the word-position effect in Experiment 1 depended on either the semantic content or the voice properties conveyed by the fine structure of the speech masker, then the effect would disappear when the three-band vocoded masker was used because these cues had been significantly reduced. However, if the word-position effect depended solely on the gross envelope properties of the speech masker, then we expected the effect to remain even after the speech masker was noise-vocoded using three bands.

TABLE 1
Boundary frequencies (Hz) for vocoding a speech signal using 2- to 16-bands (frequencies shown for even numbered bands)

2-band	300	1528	6000														
4-band	300	722	1528	6000													
6-band	300	494	814	1528	2210	3642	6000										
8-band	300	477	722	1061	1528	2174	3066	4298	6000								
10-band	300	405	546	737	994	1528	1810	2443	3296	4447	6000						
12-band	300	385	494	634	814	1045	1528	1722	2210	2837	3642	4674	6000				
14-band	300	372	460	570	706	875	1083	1342	1528	2058	2549	3158	3911	4844	6000		
16-band	300	382	477	590	722	878	1061	1276	1528	1825	2174	2584	3066	3632	4298	5080	6000

Methods

Participants

The participants in Ezzatian et al. (2011) were 16 college-aged younger adults (mean age = 21.7, $SD = 2.64$ years) with no previous exposure to the materials or to vocoded speech. As in Experiments 1 and 2, these participants spoke English as a first language, and had normal hearing in both ears. For further details see Ezzatian et al. (2011, Experiment 4).

Materials, apparatus, and procedure

All materials, equipment, and procedures used in Experiment 3 were the same as those used in Experiment 1. Similar to Experiment 1, the target and masking stimuli were presented from a single loud speaker located in front of the participants. However, in Experiment 3, the two-talker speech masker was noise-vocoded using three bands. As before, listeners were required to repeat the target anomalous sentences immediately after their presentation and were scored on the keywords in these sentences.

Results

The average dB SNR corresponding to 50%-correct identification as a function of Masker and Keyword position is plotted in Figure 3 (upper right). As can be seen in this figure, average performance is better with the noise-vocoded masker in the background than the steady-state noise masker. Figure 3 also shows that performance is fairly stable across the three keyword positions when the masker is noise. When the masker is 3-band noise-vocoded speech, performance appears to be slightly worse on the second keyword but seems to be equivalent between the first and final keyword position.

For a statistical evaluation of the results, individual thresholds were entered into a repeated-measures ANOVA with Keyword Position and Masker as within-subject variables. This ANOVA revealed a significant effect of Masker on thresholds, $F(1, 15) = 150.97$, $MSE = 152.19$, $p < .001$. On average, thresholds were 2.52 dB lower with the 3-band noise-vocoded speech masker in the background than with the steady-state speech-spectrum noise in the background. The main effect of Keyword Position was also significant, $F(2, 30) = 3.74$, $MSE = 2.42$, $p = .036$. As revealed by an SNK test, thresholds for words that occurred in the first keyword position were an average 0.53 dB lower than thresholds for words that occurred in the second keyword position ($p < .05$). However, thresholds for words occurring in the first and third keyword positions or those occurring in the second and third keyword positions were not significantly different. More importantly, the

interaction between Masker and Keyword Position was not statistically significant, $F(2, 30) = 1.97$, $MSE = 1.30$, $p = .157$.

Discussion

The results of Experiment 3 showed that performance was significantly better when the masker was noise-vocoded speech than when it was speech-spectrum noise. Although there was a slight variation in performance as a function of keyword position, the steady-state noise masker and the vocoded speech masker were equivalent in their effect on performance across the keywords. Hence, when the two-talker speech masker was noise-vocoded to retain some of the information in its amplitude envelope, while eliminating its fine structure cues and semantic content, there was no longer a progressive improvement in performance across keyword position. The fact that performance did not vary across keyword position for both the steady-state noise masker and the vocoded speech masker indicates that the gross fluctuations in the amplitude envelope of the speech masker were insufficient to produce the delay in the time-course of stream segregation observed in Experiment 1 when the masker was intact two-talker speech. These results also suggest that either the semantic content and/or the fine structure cues that were eliminated in Experiment 3 were necessary to interfere with stream segregation.

EXPERIMENT 4: THE EFFECT OF SEMANTIC CONTENT ON THE TIME-COURSE OF STREAM SEGREGATION

Since the speech masker in Experiment 1 had semantic content whereas the noise masker did not, it is possible that the word-position effect in Experiment 1 resulted from the obligatory and automatic processing of the linguistic content of the two-talker masker. The irrelevant contents of the speech masker may have initially interfered with the processing of the keywords in the target sentences, resulting in poorer performance on the earlier keywords in the sentences. Experiment 3 showed that when both voicing and semantic content were eliminated by vocoding with 3 bands, the word position effect disappeared. Hence in Experiment 4, we investigated the separate roles played by the linguistic content and the fine structure of the speech masker by vocoding the two-talker speech masker using 16 bands (see Table 1). Previous research has shown that the identification of single words in sentences can reach near perfect levels with as few as 4 bands (Shannon et al., 1995), and a pilot experiment conducted by Ezzatian et al. (2011) showed that recognition of the words in these target sentences was better than 67% when they were presented to younger adults in quiet using 8

bands of vocoding. Since the sentences in the two-talker speech masker are semantically and syntactically similar to those used in the target sentences, we presumed that using 16 bands to vocode the speech masker would allow a significant amount of its linguistic content to be distinguishable to listeners. As before, since noise vocoding removes the fine structure cues in the masker, the vocal similarities of the masking talkers to the target talker were eliminated by the vocoding procedure, thereby removing this factor as a confound in the experiment.

If the word-position effect encountered in Experiment 1 persists after the two-talker masker is vocoded with 16 bands, it would indicate that the delay in the time-course of stream segregation in Experiment 1 was due mainly to the interference of the linguistic content of the two-talker masker with stream segregation. However, if the word-position effect is no longer present after the speech masker is vocoded with 16 bands, it would indicate that semantic interference from the speech masker did not delay stream segregation in Experiment 1, leaving the contribution of fine structure cues as a possible explanation for the pattern of results.

Methods

Participants

A new set of 16 younger adults (mean age = 20.44, $SD = 1.67$ years) who had not been exposed to the experimental stimuli participated in this experiment. These new participants met the same criteria as those who had participated in Ezzatian et al. (2010, 2011) as described earlier for Experiments 1, 2, and 3 of the current study.

Materials, apparatus, and procedure

All materials, equipment, and procedures used in Experiment 4 were the same as those used in Experiment 3. There were two conditions in Experiment 4. In one condition, the target anomalous sentences were presented with the 16-band vocoded speech masker in the background and in the other condition the target sentences were presented with the steady-state noise masker in the background. However, because there were only two conditions in this experiment (there were four conditions in Experiments 1–3), 8 of the 16 lists (instead of only 4) were used in each condition.

As before, participants were required to repeat the target anomalous sentences immediately after their presentation in the noise masker or the 16-band masker, and they were scored on the number of keywords recalled correctly.

Results

Figure 3 (lower left) plots average thresholds on the three keyword positions as a function of the type of masker. As can be seen in this figure, overall performance is much better when the masker is 16-band noise-vocoded speech than when it is steady-state noise. The pattern of performance across the keyword positions seems to be relatively stable across all keywords when the masker is noise, a finding that is consistent with that of the analyses conducted in Experiments 1–3. When the masker is 16-band vocoded speech, average performance seems to be worse for the second keyword position as compared to that for the first and third keyword positions. However, the first and third keyword positions seem to have equivalent average thresholds.

Consistent with the pattern of results displayed in Figure 3 (lower left), an ANOVA with Masker and Keyword Position as within-subject factors revealed a significant effect of Masker on thresholds, $F(1, 15) = 183.06$, $MSE = 292.88$, $p < .001$. On average, thresholds were better by 3.5 dB when the masker was 16-band noise-vocoded speech than when it was speech-spectrum noise. The effect of Keyword Position on thresholds was also significant, $F(2, 30) = 6.804$, $MSE = 7.402$, $p = .004$. The SNK post hoc tests revealed a significant difference in average thresholds between the first and second keywords, mean difference = 0.621 dB $t(30) = 3.368$, $p < .05$, and the second and third keyword, mean difference = 0.946 dB $t(30) = 5.130$, $p < .05$. The difference between the first and third keywords was not statistically significant, $t(30) = 1.763$, $p > .05$. More importantly, the interaction between Masker and Keyword Position was not statistically significant, $F(1, 30) = 2.11$, $MSE = 2.98$, $p = .139$.

Discussion

The results of Experiment 4 indicate that the pattern of performance as a function of keyword position did not depend on the type of maskers. This finding suggests that when the two-talker speech masker from Experiment 1 is noise-vocoded using 16 bands to preserve its linguistic content, its effect on the time-course of stream segregation is minimal and statistically no different than the effect of a steady-state noise masker. Thus, it is unlikely that the apparent delay in the buildup of stream segregation that was observed with the two-talker masker in Experiment 1 was due solely to the linguistic content of this masker.

The results of Experiment 4 show a 3.5 dB improvement in performance in the vocoded masking condition relative to the continuous noise masking condition. Despite the fact that the words in the 16-band speech masker were far more discernible than those in the 3-band speech masker, which could have resulted in more interference from the linguistic content of the masker

sentences, thresholds in Experiment 4 were about 1 dB lower than those observed in Experiment 3, where the two-talker masker was vocoded using only 3 bands (2.52 dB). Since using more bands to noise vocode the speech masker in Experiment 4 allowed more of its amplitude envelope information to be preserved, it is likely that the difference in performance between the 3-band and 16-band masking conditions is related to the fuller range of fluctuations that were available to listeners with the 16-band masker. This masker may have allowed listeners to gain more glimpses of the target sentences than was possible with the 3-band speech masker, thereby leading to better performance with the 16-band than the 3-band noise-vocoded speech masker. Moreover, the fact that thresholds were lower for the 16-band than for the 3-band noise-vocoded speech maskers suggests that the semantic content of the two-talker masker appears to have had little or no effect on performance when the voice cues were eliminated by noise vocoding. That is, the linguistic content of the masker seems to be of little importance to performance when its semantic content is largely retained but voice cues are lost.

EXPERIMENT 5: THE EFFECT OF VOICE CUES ON THE TIME-COURSE OF STREAM SEGREGATION

The results of the previous experiments rule out that the delay in the buildup of stream segregation observed in Experiment 1 is due solely to the gross fluctuations in the amplitude envelope or the semantic contents of the masker sentences. The use of fine structure cues associated with the fundamental frequency and harmonic structure that differentiated the voices must be considered as a remaining factor. Given that the target and masking sentences were uttered by females of the same approximate age and background, it is possible that the delay in the time-course of stream segregation observed in Experiment 1 was caused by the difficulty of segregating three qualitatively similar voices and assigning one to the target talker. Indeed, as Brungart and colleagues have demonstrated, increasing the differences between the voices of the target and masking talkers (e.g., by manipulating gender) can result in a significant improvement in performance in speech-masking experiments (Brungart et al., 2001; Darwin et al., 2003; Humes et al., 2006). Hence, the confusability between the voices alone, independent of the linguistic contents of the masking speech streams, could have resulted in a delay in the buildup of stream segregation in Experiment 1.

If this assertion is true, then repeating Experiment 1 should result in the same pattern of performance even if the linguistic contents of the speech masker is made indistinguishable, as long as the voices of the target and

masking talkers remain similar. Experiment 5 was conducted to explore this hypothesis.

Methods

Participants

Sixteen younger adults (mean age = 20.38, $SD = 1.78$ years) without prior exposure to the stimuli participated in this experiment. These younger adults met the same criteria as those used to determine the eligibility of participants in the previous four experiments.

Materials, apparatus, and procedure

Experiment 5 was conducted using the same materials, equipment, and procedures as in the previous experiments. There were two conditions in Experiment 5, a noise-masking condition using the standard noise masker, and a speech-masking condition in which the speech masker from Experiment 1 was time-reversed to make its semantic content unrecognisable while still preserving the vocal characteristics of the talkers. However, because there were only two conditions in this experiment (there were four conditions in Experiments 1–3), 8 of the 16 lists (instead of only 4) were used in each condition.

Results

Figure 3 (lower right) plots average thresholds for each keyword position as a function of the type of masker. As can be seen in this figure, overall performance seems marginally better with the reversed speech masker than the steady-state noise masker in the background. Furthermore, whereas performance is relatively stable across keyword positions when the masker is noise, it improves with keyword positions when the masker is reversed speech.

An ANOVA on the thresholds with Masker and Keyword Position as within-subject factors showed that despite a 0.6 dB improvement when the masker was reversed speech, the effect of masker on thresholds was not statistically significant, $F(1, 15) = 3.45$, $MSE = 8.59$, $p = .083$. However, the effect of Keyword Position, and the interaction of Masker and Keyword Position on thresholds were significant, $F(2, 30) = 7.66$, $MSE = 6.98$, $p = .002$ and $F(2, 30) = 5.65$, $MSE = 5.40$, $p = .008$, respectively.

The SNK post hoc tests of this interaction revealed that when the masker was noise, average thresholds were statistically equivalent across keyword

positions;[5] however, when the masker was speech, the average threshold for the final keyword was significantly lower (better) than that for the first, mean difference = 1.58 dB, $t(30) = 9.124$, $p < .05$ and second keywords, mean difference = 1.30 dB, $t(30) = 7.510$, $p < .05$, whereas average thresholds on the first and second keywords were statistically equivalent, (mean difference = 0.279 dB, $t(30) = 1.614$, $p > .05$.

Discussion

The most important finding of Experiment 5 is the differential effect of the two maskers on performance across keyword positions. Consistent with previous findings, performance was equivalent across keyword positions when the masker was noise. However, when the masker was time-reversed speech, performance on the first two keyword-positions was significantly poorer than that on the final keyword position, a finding that is identical to that obtained in Experiment 1. These results suggest that the delay in the buildup of stream segregation in the speech-masking condition of Experiment 1 was most likely due to the confusability of the target and masking talkers' voices. Since the target and masking speech streams originated from the same spatial location in both experiments and were uttered by female talkers of the same age and background, it is likely that the similarity between the competing voices may have initially made it difficult for the listeners to segregate them, resulting in poorer performance on the earlier parts of the target sentences.

To examine whether the duration of exposure to the vocal characteristics of the speech target and reversed speech masker affected performance, we compared the performance of listeners on the first four lists to which they were exposed, when the reversed speech masker was present, to their performance on the second four lists. (Because all four SNRs were presented in different random orders during the first-four and second-four lists, we averaged across the four SNRs associated with each set of four lists to determine the average percentage of words correctly identified as a function of word number for the first and second halves of this condition.) Figure 4 shows that word identification increased from the first half to the second half of the session, but that increased exposure to the stimuli did not alter the word position effect. A two (first half vs. second half) by three (word position) within-subject ANOVA indicated a significant effect of duration of exposure, $F(1, 15) = 32.92$, $p < .001$, a significant effect of word position, $F(2, 30) = 21.326$, $p < .001$, but no interaction between the two factors, $F(2,$

[5] Keyword 1 vs. Keyword 2, mean difference = 0.469 dB, $t(30) = 2.713$, $p > .05$; Keyword 1 vs. Keyword 3, mean difference = 0.063 dB, $t(30) = 0.365$, $p > .05$; Keyword 2 vs. Keyword 3, mean difference = 0.406 dB, $t(30) = 2.350$, $p > .05$.

30) < 1. Hence, although duration of exposure to the stimuli improves word identification, it does so equally for all three word positions. If increased familiarity with the stimuli led to more rapid segregation of the speech target from the speech masker, we would expect the word position effect to be attenuated in the second half of the sessions. Apparently, increased familiarity with the stimuli (both target and reversed speech masker) results in a general improvement in performance without any indication that it increases the rapidity with which the speech target can be segregated from the speech masker.

The results of Experiment 5 also show that, similar to Experiment 1 but contrary to Experiments 2, 3, and 4, listeners did not seem to benefit from the fluctuations in the time-reversed speech masker. In the current experiment, despite a 0.6 dB improvement in thresholds relative to when the masker was steady-state noise, the listeners' performance was statistically equivalent across the two masking conditions. One possible reason why a listener might not have been as able to take as much advantage of the troughs in reversed speech as they could in normal speech could have to do with the predictability of the occurrence of the troughs. In normal speech listeners can anticipate when troughs (gaps) are likely to occur so that they could focus their attention at these points (e.g., Astheimer & Sanders, 2009). However, it is unlikely that listeners will be able to anticipate the occurrence of gaps in

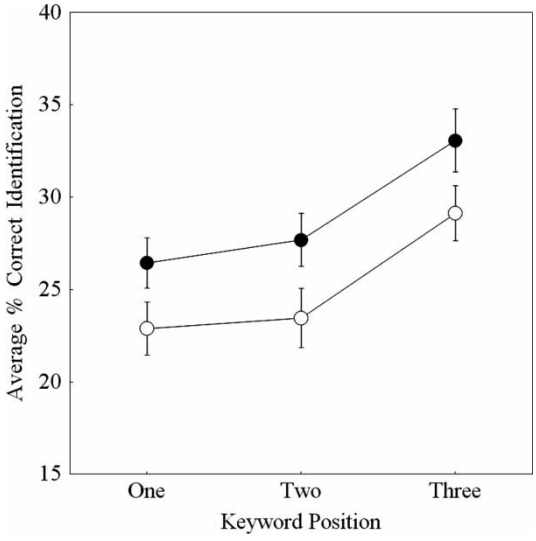

Figure 4. Average percent correct identification (averaged over SNR) of target words as a function of word position for the first (open circles) and second (filled circles) halves of the condition in which the masker was reversed speech. Error bars represent the standard error of the means.

reversed speech, thereby depriving them of the opportunity to focus attention on the signal during the gaps. It is also possible that the listeners in Experiment 5 were unable to take full advantage of the troughs in the amplitude envelope of the reversed speech masker due to an increase in the degree of forward masking introduced by time-reversing the speech masker. Rhebergen, Versfeld, and Dreschler (2005) found that Dutch listeners performed better when listening to Dutch sentences in the presence of a Swedish masker (which was incomprehensible to them), than when they listened to Dutch sentences masked by a reversed Swedish masker. They attribute this to a greater degree of forward masking for reversed speech than for normal speech.

SUMMARY, GENERAL DISCUSSION, AND CONCLUSIONS

Summary

The purpose of the current study was to examine how the masking of speech signals affects the time-course of stream segregation; that is, the amount of time required by the auditory system to identify and extract a target signal from a mixture of overlapping auditory streams. We proposed that masking a speech signal by acoustic streams that are similar to speech should result in a delay in the buildup of stream segregation, since increased similarity between concurrent streams will increase the likelihood that they are confused with each other. Conversely, masking a speech signal with auditory streams that are dissimilar to speech should result in little to no delay in the buildup of stream segregation, since it is less likely for auditory streams that share few similarities to be confused with each other. Consistent with these hypotheses, we found that listeners experienced a significant delay in the buildup of stream segregation when the utterances of a target talker were masked with the utterances of two similar-sounding talkers, but not when the target utterances were masked by a continuous noise. Furthermore, our results showed that the delay in segregating the target speech from the competing speech was primarily due to the fine structure cues associated with vocal similarities between the target and competing talkers and not to the gross fluctuations in the amplitude envelope or semantic content of the competing speech.

General discussion

The results of the reanalyses conducted in Experiments 1 and 2 indicate that the detrimental effect of background talkers on spoken language comprehension is in large part related to interference with auditory scene analysis; that is, when acoustic similarities exist between the speech streams of the

target and background talkers, the speech streams belonging to the background talkers interfere with the processing of the target stream primarily by increasing the amount of time it takes for the auditory system to segregate the target speech stream from the irrelevant background streams.

Furthermore, the results of this study, particularly those of the reanalysis conducted in Experiment 2, suggest that conditions in which a significant release from masking is achieved by the introduction of perceived spatial separation between the target and masking sounds (and possibly other manipulations such as using talkers of different genders), the resulting release from speech masking is related to the fact that these manipulations reduce the time it takes for the auditory system to segregate the target speech stream from the background competitors.

Speech masking

To allow a comparison between the speech maskers in the current study, the average threshold obtained for each of the speech maskers is plotted in Figure 5 (left) as a function of keyword position. An examination of this figure reveals that in terms of masking effectiveness, the unprocessed speech masker, which was presented without any spatial separation from the target sentences, was most effective in masking the target sentences. The second most effective masker was the time-reversed speech masker, which also resulted in a delay in the buildup of stream segregation. Note, however, that time-reversed speech is a less-effective masker than normal speech, suggesting that when stream segregation is incomplete (as it appears to be for both forward and reversed speech maskers), the semantic content of the masker interferes with the processing of the speech target. Finally, the remaining three maskers (the 3-band and 16-band noise-vocoded maskers and the intact speech masker with spatial separation) seem to be the least effective. To confirm these observations, mean thresholds were averaged across keyword positions and entered into a one-way ANOVA with Masker (Intact, Precedence Effect, 3-band vocoded, 16-band vocoded, time-reversed) as a between-subjects factor. This analysis revealed a significant effect of Masker on average thresholds, $F(4,\ 75) = 33.08,\ MSE = 2.63,\ p < .001$. The SNK analyses on the means revealed that performance on the intact two-talker masker was worse than all other masking conditions, followed by performance with the time-reversed speech masker, which was better than that with the intact masker, but worse than performance under the remaining three masking conditions. Finally, average thresholds in the conditions with target and masker were perceived to be spatially separated due to the precedence-effect, and the 3-band and 16-band vocoded maskers did not differ statistically from each other.

This pattern of results is interesting for several reasons. First, the only manipulation of the speech masker that resulted in nearly the same amount of masking as the intact speech masker was one where the fine structure voicing cues in the masking sentences were preserved, but their semantic content was, for the most part, eliminated. This suggests that in informational masking, interference from vocal similarities between competing talkers takes priority over semantic interference from the competing sentences. The semantic content of competing sentences result in additional interference only when vocal similarities between the overlapping speech streams have already caused an initial delay in stream segregation but not without this initial delay. Note that when this masker was noise-vocoded using 16 bands, presumably enough to preserve a significant amount of its linguistic content, it did not result in nearly as much masking as the two-talker masker in Experiment 1. In fact, noise vocoding the speech masker using 16 bands resulted in an approximately 5 dB release in masking relative to the original masker. In contrast, when the intact two-talker masker was time reversed in Experiment 5, thus preserving its vocal similarities with the target sentences but making its linguistic content far less discernible, it resulted in significantly more masking than the 16-band masker (mean difference = 3.12 dB). This is further demonstrated by comparing the speech maskers in the analyses in Experiments 1 and 2. These two maskers were identical to each other; the only difference between them was that the one in Experiment 1 originated from the same spatial location as the target sentences, whereas the one in Experiment 2 was perceived to originate from a different spatial location than the target sentences. However, as can be seen for the speech maskers in Figure 5 (left), while using the speech masker in Experiment 1 resulted in the most amount of interference with performance in the entire study, it resulted in minimal interference in Experiment 2, where performance was better by an average of 5 dB relative to Experiment 1. In fact, when there was a perceived spatial separation between the two-talker masker and the target sentences, it was no more effective in masking the target sentences than the 3-band vocoded masker, which shared neither the vocal nor linguistic similarities of the two-talker masker with the target sentences. This pattern of results highlights the importance of stream segregation in the masking of speech; in the absence of interference with stream segregation, the effectiveness of an otherwise formidable speech masker is reduced to that of a fluctuating noise.

Of further note is the difference between the precedence-effect speech masker, and the 3-band and 16-band maskers. A comparison of overall performance with these maskers indicates that all three caused only a minimal delay in the buildup of stream segregation, and therefore did not result in any notable amount of informational masking. However, since each

varied in the amount of fluctuations in its amplitude envelope, it would seem reasonable to assume that listeners would display differences in how they performed in each of these masking conditions. More specifically, since the precedence-effect masker had the fullest range of fluctuations in its amplitude envelope, one would expect performance to be better with this masker in the background, followed by the 16-band masker, and finally the 3-band masker. However, this was not the case; overall performance with these maskers was statistically equivalent. One way to interpret these results is that focusing attention to troughs in as few as three frequency bands provides the maximum benefit to listeners from gaining glimpses of the target signal, and that no further benefit is gained from increasing the amount of envelope information beyond three bands. It is also possible that the benefit to performance gained from having access to more fluctuations in the amplitude envelope of the speech maskers was offset by the accompanying increase in the interference from the linguistic contents of the maskers.

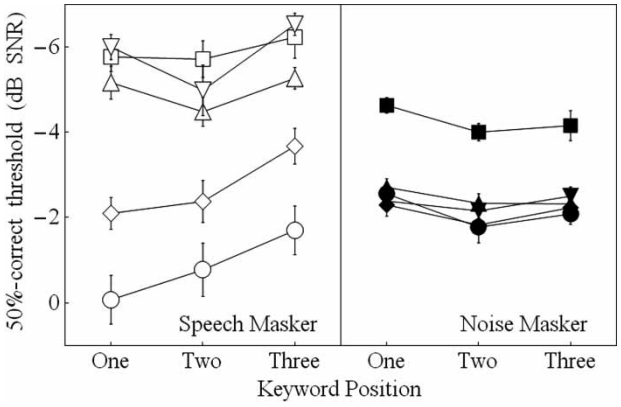

Figure 5. Left panel. Average thresholds in dB SNR as a function of Keyword Position in the speech-masking conditions of Experiments 1 to 5. Right panel. Average thresholds in dB SNR as a function of Keyword Position in the speech-spectrum, noise-masking conditions of Experiments 1 to 5. Open circles represent the condition in which target sentences were masked by a two-talker speech masker (Experiment 1), filled circles represent the noise masker in the same experiment. Open squares represent the condition in which target sentences were masked by a precedence-effect separated two-talker speech masker (Experiment 2), filled squares represent the noise masker in the same experiment. Open upward pointing triangles represent the condition in which target sentences were masked by the 3-band noise-vocoded, two-talker speech masker (Experiment 3), filled upward pointing triangles represent the noise masker in the same experiment. Open downward pointing triangles represent the condition in which target sentences were masked by the 16-band noise-vocoded two-talker speech masker (Experiment 4), filled downward pointing triangles represent the noise masker in the same experiment. Open diamonds represent the condition in which target sentences were masked by the time-reversed two-talker speech masker, filled diamonds represent the noise masker in the same experiment. Error bars represent the standard error of the means.

Unfortunately, the conditions of the current study do not allow for a direct assessment of these interpretations.

Noise masking

It is clear from the results of this study that a continuous noise masker results in a minimal delay in the buildup of stream segregation. However, it is possible that there were differences between the participant groups in the various analyses. Hence to make the comparison between the groups easier, average thresholds under all noise-masking conditions in this study are plotted in Figure 5 (right) as a function of keyword position. As can be seen in this figure, there seems to be little variation in performance as a function of keyword position when the masker is noise, and this pattern of performance is virtually identical across all experiments. Overall performance with the noise masker is also consistently similar across all experiments, with the exception of Experiment 2, where there was a perceived spatial separation between the noise masker and the target sentences. The mean thresholds in all the noise masking conditions were averaged across the three keyword-positions and entered into a one-way ANOVA with Masker as a between-subjects factor. This ANOVA revealed a significant effect of Masker on thresholds, $F(4, 75) = 25.09$, $MSE = 0.52$, $p < .001$. As expected, post hoc SNK analyses revealed that average performance with the precedence-effect noise masker was significantly better than that under all other noise-masking conditions (mean difference = 2.04 dB, $p < .05$), but that performance under the other noise-masking conditions was statistically equivalent.

Other researchers have also found a release from noise masking with a precedence-effect induced spatial separation. Indeed, Freyman et al. (1999) and Li et al. (2004) showed a similar release from noise masking using the same stimuli as the current study (1 dB, and 1.7 dB release, respectively). Those authors argued that the improvement in performance in their noise masking conditions was due to the interaural time differences between the target speech and noise masking streams, which led to a reduction in masking in the lower frequencies. Based on thresholds for 200-ms long, one-third octave noise bands masked with a steady-state speech-spectrum noise masker, Freyman et al. (1999) used the articulation index to predict that the benefit from the interaural time differences between the target and masking signals would approximately equal 2 dB when these signals are separated using the precedence effect relative to when they appear to emanate from the same spatial location. The benefit obtained in the current experiment is equivalent to this predicted benefit. Hence, it is likely that the improved

noise-masking thresholds in the precedence condition of the current study are due to the benefit provided by the interaural time differences created by the precedence effect and do not result from any differences between the participants.

Conclusions

The present study found that masking target speech with other, similar sounding speech results in a delay in the time required by the auditory system to extract the target speech from the background competitors. This delay in the time-course of stream segregation is primarily due to the vocal similarities between the target and masking talkers, but is exacerbated when the linguistic content of the background talkers is also discernible to listeners. However, providing cues that aid stream segregation, such as a perceived spatial separation, can for the most part alleviate the interference to stream segregation caused by background talkers.

When the background masker consists of noise, there is no evidence for a delay in the time-course of stream segregation, presumably, because a noise and a human voice are different enough to make their segregation by the auditory system relatively easy and immediate. Hence, providing cues such as perceived spatial separation between the target and masking streams results in only a minimal improvement in performance when target speech is masked by a noise.

Because speech maskers may lead to a delay in the buildup of stream segregation, it is possible that at least some of the comprehension difficulties experienced by listeners in multi-talker environments are related to the fact that more time is required to extract a speech target from concurrent speech competitors relative to when the competing streams consist of nonspeech sounds. If important information in the speech signal is lost during this delay in segregating the speech target from the irrelevant background streams, then the listener may need to expend additional cognitive resources to reconstruct the beginning of the signal or the listener may be unable to comprehend the message altogether if the information from the onset of the utterance cannot be restored. This is especially true in environments where cues such as spatial separation between the competing streams, or other qualitative differences between the target and background talkers are likely to be sparse, or among populations such as older adults or the hearing impaired who may be less able than younger adults to use these cues to segregate overlapping speech streams.

It is important to note, however, that our experiments were conducted with short anomalous sentences that placed minimal demands on the

memory and language processing centres of the brain and also eliminated possible benefits from the use of knowledge. The target and masking signals that are typically encountered in everyday life are likely to be far more complicated than the ones used in our experiments and, indeed, the use of longer, semantically coherent speech materials may lead to different results than those obtained here.

REFERENCES

Anstis, S., & Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(3), 257–271.

Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America*, *112*(5), 2086–2098.

Astheimer, L., & Sanders, L. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biological Psychology*, *80*(1), 23–34.

Beauvois, M. W., & Meddis, R. (1997). Time decay of auditory stream biasing. *Perception & Psychophysics*, *59*(1), 81–86.

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology-Human Perception and Performance*, *4*(3), 380–387.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sounds*. Cambridge, MA: MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*(2), 244–249.

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, *47*(3), 191–196.

Brungart, D. S., & Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *Journal of the Acoustical Society of America*, *112*(6), 2985–2995.

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, *110*(5), 2527–2538.

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

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *Journal of the Acoustical Society of America*, *128*(2), 860–869.

Carhart, R., Tillman, W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *Journal of the Acoustical Society of America*, *45*, 694–703.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 115–127.

Carlyon, R. P., Plack, C. J., Fantini, D. A., & Cusack, R. (2003). Cross-modal and non-sensory influences on auditory streaming. *Perception*, *32*(11), 1393–1402.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 643–656.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *Journal of the Acoustical Society of America*, *114*(5), 2913–2922.

Ezzatian, P., Avivi, M., & Schneider, B. (2010). Do non-native listeners benefit as much as native listeners from spatial cues that release speech from masking? *Speech Communication*, *52*(11–12), 919–929.

Ezzatian, P., Li, L., Pichora-Fuller, K., & Schneider, B. (2011). The effect of priming on release from informational masking is equivalent for younger and older adults. *Ear and Hearing*, *32*(1), 84–96.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America*, *109*, 2112–2122.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, *115*(5), 2246–2256.

Freyman, R. L., Helfer, K. S., & Balakrishnan, U. (2007). Variability and uncertainty in masking by competing speech. *Journal of the Acoustical Society of America*, *121*(2), 1040–1046.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, *106*(6), 3578–3588.

Garcia Lecumberi, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *Journal of the Acoustical Society of America*, *119*, 2445–2454.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language and Hearing Research*, *40*(2), 432–443.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America*, *117*(2), 842–849.

Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear and Hearing*, *29*, 87–98.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29*(2), 296–305.

Humes, L. E., Lee, J. H., & Coughlin, M. P. (2006). Auditory measures of selective and divided attention in young and older adults using single-talker competition. *Journal of the Acoustical Society of America*, *120*(5), 2926–2937.

Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, *118*(6), 3804–3815.

Li, L., Daneman, M., Qi, J. G., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance*, *30*(6), 1077–1091.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*, 203–243.

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *Journal of the Acoustical Society of America*, *22*(5), 637–638.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica United with Acustica*, *88*(3), 320–333.

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech (L). *Journal of the Acoustical Society of America*, *118*(3), 1274–1277.

Rogers, W. L., & Bregman, A. S. (1993). An experimental evaluation of 3 theories of auditory stream segregation. *Perception & Psychophysics*, *53*(2), 179–189.

Schneider, B. A., Li, L., & Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology*, *18*(7), 559–572.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Singh, G., Pichora-Fuller, M. K., & Schneider, B. A. (2008). The effect of age on auditory spatial attention in conditions of real and simulated spatial separation. *Journal of the Acoustical Society of America*, *124*(2), 1294–1305.

Snyder, J. S., Alain, C., & Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of Cognitive Neuroscience*, *18*(1), 1–13.

Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, *69*(1), 136–152.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Teachers College, Columbia University.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America*, *121*, 519–526.

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences* (Unpublished doctoral dissertation). Eindhoven University of Technology, The Netherlands.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). The precedence effect in sound localization. *American Journal of Psychology*, *62*, 315–336.

Watson, C. S. (2005). Some comments on informational masking. *Acta Acoustica*, *91*, 502–512.

Yang, Z. G., Chen, J., Wu, X. H., Wu, Y. H., Schneider, B. A., & Li, L. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Communication*, *49*, 892–904.