Original Article

Trends in Hearing
2017, Volume 21: 1–14

DOI: 10.1177/2331216517709385
journals.sagepub.com/home/tia
SAGE

Effectiveness of Two-Talker Maskers That Differ in Talker Congruity and Perceptual Similarity to the Target Speech

Lauren Calandruccio[1], Emily Buss[2], and Kristina Bowdrie[1]

Abstract

Previous work has shown that masked-sentence recognition is particularly poor when the masker is composed of two competing talkers, a finding that is attributed to informational masking. Informational masking tends to be largest when the target and masker talkers are perceptually similar. Reductions in masking have been observed for a wide range of target and masker differences, including language: Performance is better when the target and masker talkers speak in different languages, compared with the same language. The present study evaluated normal-hearing adults' sentence recognition in a two-talker masker as a function of the perceptual similarity between the target and *each* of the two masker streams. The target was English, and the maskers were composed of English, time-reversed English, or Dutch. These three masker types are known to vary in the informational masking they exert. The two talkers within the two-talker maskers were either congruent (e.g., both English) or incongruent (e.g., one English, one Dutch). As predicted, mean performance was worse for the congruent English masker than the congruent time-reversed English or congruent Dutch maskers. Incongruent two-talker maskers, with just one English masker stream, were only modestly less effective than the congruent English masker. This result indicates that two-talker masker effectiveness was determined predominantly by the one masker stream that was most perceptually similar to the target. Speech recognition in a single-talker masker differed only marginally between the English, Dutch, and time-reversed English masker types, suggesting that perceptual similarity may be more critical in a two-talker than a one-talker masker.

Keywords

informational masking, two-talker maskers, sentence recognition, incongruent maskers, masker effectiveness, speech recognition, perceptual similarity

Date received: 28 August 2016; revised: 11 April 2017; accepted: 12 April 2017

Introduction

Speech recognition in the context of one or more competing speech streams has been a topic of study for some time (e.g., Miller, 1947). Performance on tasks of speech-on-speech recognition is thought to reflect the difficulties many listeners have when listening in noisy environments such as a "cocktail party" (Cherry, 1953). Often, the masking observed in speech-on-speech experiments exceeds that predicted based on the temporal and spectral overlap between the target and competing signal(s) within the auditory periphery alone. Further, the intelligibility measured in speech-on-speech masking experiments is not only often poorer but also tends to be more variable between listeners than that measured for speech-in-noise masking experiments (Festen & Plomp, 1990). This "additional" masking has been described as perceptual masking (e.g., Carhart, Tillman, & Greetis, 1969; Elliott, Connors, Kills, & Levin, 1979) but now is more commonly referred to in the literature as "informational"

[1]Department of Psychological Sciences, Program in Communication Sciences and Disorders, Case Western Reserve University, Cleveland, OH, USA
[2]Department of Otolaryngology/Head and Neck Surgery, University of North Carolina School of Medicine, Chapel Hill, NC, USA

Corresponding author:
Lauren Calandruccio, Department of Psychological Sciences, Program in Communication Sciences, Case Western Reserve University, 11635 Euclid Avenue Room 337, Cleveland, OH 44106, USA.
Email: lauren.calandruccio@case.edu

masking (Brungart, 2001; Brungart, Simpson, Ericson, & Scott, 2001; Freyman, Helfer, McCall, & Clifton, 1999). The phrase ''informational masking'' is also used in the psychophysical literature to describe the masking associated with perceptual similarity and stimulus uncertainty observed for nonspeech auditory stimuli (Kidd, Mason, Deliwala, Woods, & Colburn, 1994; Kidd, Mason, Rohtla, & Deliwala, 1998; Watson, Kelly, & Wroton, 1976).

Typically, a small number of competing speech streams produces more informational masking than a multitalker babble composed of many different speech streams. For open-set sentence recognition, *two* competing talkers have been observed to cause significant amounts of informational masking (Freyman, Balakrishnan, & Helfer, 2004; Rosen, Souza, Ekelund, & Majeed, 2013). In 2004, Freyman et al. assessed sentence recognition for a number of different speech-based maskers, with and without a perceived spatial separation between sources based on the precedence effect. The rationale was that this spatial cue would reduce informational masking by promoting segregation but would have no beneficial effect with respect to energetic masking. In this experiment, the target was always played through a loudspeaker directly in front of the listener. In the baseline, no separation condition, the masker also played from the front. In the perceived-spatial separation condition, the masker was played from two speakers, one directly in front and other off to the side, with the stimulus from the side of the listener leading by 4 ms. This asynchrony made the masker sound as if it was coming from the side, while the target was perceived as coming from the front. Freyman et al. measured performance for maskers composed of 2, 3, 4, 6, and 10 talkers. The largest benefit of perceived target and masker spatial separation was observed for the two-talker masker condition, indicating that much of the difficulty listeners had with the two-talker masker in the baseline condition was due to an inability to separate the target from the competing speech streams.

Other experimental methods have also provided evidence that much of a two-talker masker's effectiveness is due to informational rather than energetic masking contributions. In 2001, Brungart tested listeners using the coordinated response measure (CRM; Moore, 1981), a highly structured speech test in which keywords are selected from among a limited list of alternatives (e.g., ''Ready Baron go to blue two now,'' where only the underlined ''call sign,'' ''color,'' and ''number'' vary from trial to trial). One benefit of using the CRM is that the closed-set format makes it easy to determine whether incorrect listener responses are based on intrusions from the competing masker stream. Brungart provided compelling evidence that significant decreases in performance occurred for two-talker masker conditions beyond what could be accounted for by energetic masking. Further, the distribution of the error patterns indicated that many listener responses were intrusions from one of the masker streams. This result was interpreted as indicating that the listeners had difficulty segregating the target from the masker speech; intrusions occurred for both positive and negative signal-to-noise ratios (SNRs).

Speech-on-speech recognition for two-talker maskers consisting of different languages than the target speech has also shown large reductions in informational masking (Freyman, Balakrishnan, & Helfer, 2001; Van Engen & Bradlow, 2007). That is, if a listener is asked to recognize English speech in the presence of an English masker, recognition is typically poorer than that observed for the same task in the presence of a non-English masker (e.g., Mandarin, Spanish, or Greek; Calandruccio, Brouwer, Van Engen, Dhar, & Bradlow, 2013). The mismatch in target and masker languages is beneficial even when the listener can understand, with high proficiency, both languages (Calandruccio & Zhou, 2014). These data imply that the lack of linguistic meaning from the mismatched language is not what is causing the release but rather the differences between the two languages provides listeners with a segregation cue that allows them to improve their overall speech recognition.

Failure to segregate the target and masker speech in a two-talker masker is unlike what is traditionally observed during energetic masking tasks in which the target speech becomes less audible with increases in masking noise level. It is unclear what makes a two-talker masker so effective with respect to informational masking; however, the effectiveness of different two-talker maskers is likely dictated by (a) stimulus uncertainty and (b) target and masker similarity.

For speech-on-speech recognition tasks, decreases in informational masking are observed when uncertainty is reduced by cuing the listener to what or who to listen for (Freyman et al., 2004). Decreases in target and masker similarity have been shown to reduce informational masking for a wide range of features, including the talker identity and gender (Brungart, 2001), language (Freyman et al., 2001; Garcia Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007), accent (Calandruccio, Dhar, & Bradlow, 2010), semantic content of the speech (Brouwer, Van Engen, Calandruccio, & Bradlow, 2012), syntactic content of the speech, and meaningfulness (Rhebergen, Versfeld, & Dreschler, 2005). While the bulk of data show effects of target and masker similarity, there are counter examples in the literature, where stimulus differences did not reduce masking. For example, Calandruccio et al. (2010) and Dirks and Bower (1969) saw no reduction in masking with a foreign

language masker. Nevertheless, in all of these reports, target and masker similarity was manipulated congruently for the two masking talkers; that is, both masking talkers were either similar or dissimilar to the target. As a consequence, it is unclear how each masker stream contributes to the informational masking exerted by the combined two-talker masker for sentence recognition.

In contrast to the open-set sentence recognition tasks reviewed earlier, Iyer, Brungart, and Simpson (2010) addressed the relative contributions of each masker stream for a closed-set sentence identification test with a two-talker masker that began synchronously with the target. The target was a CRM sentence, and the two-talker masker always included at least one other CRM sentence; in both cases, the CRM sentence was in English. The second masker stream was either similar to the target—in this case an additional CRM sentence—or dissimilar. Dissimilar speech maskers were an English sentence from a corpus other than the CRM, a sentence spoken in a language other than English, a time-reversed sentence spoken in English, or a time-reversed sentence spoken in a language other than English. The amount of informational masking was relatively consistent across the different two-talker maskers. On the basis of these results, the authors concluded that the perceptual similarity between the target and the second masker talker was not an important factor determining the amount of informational masking. That is, if one of the two talkers is perceptually similar to the target, then features of the other masker stream do not matter.

Three experiments were conducted to explore different parameters of speech-on-speech masking. Experiment 1 assessed the effect of similarity between the target and two-talker masker speech on informational masking and overall speech recognition. An open-set sentence recognition task was performed with each of five two-talker maskers, which differed in similarity to the target speech and differed in the congruency between the two talkers within the masker speech. Experiment 2 evaluated the importance of similarity between the target and the individual speech streams used to create the two-talker masker speech used in Experiment 1. This experiment allowed for an investigation of the cost associated with increasing the number of masker talkers from one to two, an effect described as the multimasker penalty (Durlach, 2006). Finally, Experiment 3 investigated the effect of keyword position on speech recognition in a two-talker masker. This experiment was conducted to follow-up on results observed in Experiment 1, which indicated worsening performance with increasing keyword position, which was not in agreement with previous literature (Ezzatian, Li, Pichora-Fuller, & Schneider, 2012).

Experiment 1: Two-Talker Maskers Differing in Congruency and Similarity to the Target

Rationale

The purpose of the first experiment was to assess the importance of perceptual similarity between the two competing talkers and the target speech with respect to informational masking for an open-set sentence recognition task, with onset asynchrony between the masker and subsequent target. In the following experiments, English, time-reversed English, and foreign speech were utilized so that the perceptual similarity between the target and masker speech could be varied, while maintaining spectral and temporal features of an auditory environment characteristic of the *cocktail party* situation. One motivation for using open-set sentences was that natural communication is typically less semantically constrained than responses for a closed-set sentence task. Another feature of the present experiment was asynchronous gating of the target and masker. In natural speech, background talkers rarely stop and start synchronously with the target of interest. Asynchronously gating the target and masker would provide the listener with more opportunities to become familiar with the features of that masker, which in turn might facilitate stream formation and segregation from the target.

Methods

Participants. Twenty adult listeners (age range: 18 to 51 years; mean age = 24 years; $SD = 8.4$ years) were recruited from Case Western Reserve University and the Cleveland, Ohio area. Otoscopic evaluations were performed on all participants to ensure that participants' ear canals were clear prior to ear-tip insertion. Audiometric thresholds were tested using standard clinical procedures (American Speech-Language-Hearing Association, 2005) and were confirmed to be <25 dB HL at octave frequencies between 250 and 8000 Hz, bilaterally, for all listeners. All listeners completed a linguistic questionnaire prior to experimental testing and were confirmed to be native speakers of American English.

Five lists of basic English lexicon (BEL) sentences (Calandruccio & Smiljanić, 2012) were used as target stimuli. The BEL sentences include 20 lists of 25 sentences. Each BEL sentence is five to seven words in length and contains four keywords to be used for scoring, resulting in 100 keywords per list. An example BEL sentence (with keywords capitalized) is, "The FRUIT and SALAD TASTE FRESH." Lists 1, 2, 5, 12, and 16 were used for testing.

The target talker was a 26-year-old female, native speaker of American English. For the sentences used,

her mean fundamental frequency (F0) was 202.9 Hz and her mean speaking rate was 4.0 syllables/second. Target sentences were digitally recorded at a 44.1 kHz sampling rate with 16-bit resolution in a sound-treated room using a Shure SM81 cardioid condenser microphone with pop filter attached and placed 12 in. from the talker's lips (as described in Calandruccio & Smiljanić, 2012). The average duration of the BEL sentences used in this study was 2.1 s ($SD = 0.22$ s). All target sentences were root-mean-square equalized using Praat.

Five different two-talker maskers were used for testing, each consisting of two streams of speech approximately 60 s in length (56 to 63 s). Streams of English speech consisted of concatenated Harvard sentences (IEEE Subcommittee on Subjective Measurements, 1969). English sentences were recorded by a 27-year-old female, native speaker of American English (Talker "E"), using the same instrumentation described earlier for the Target talker. Streams of Dutch were the same IEEE sentences, translated into Dutch (Brouwer, et al., 2012). Dutch sentences were recorded by a 25-year-old female, native speaker of Dutch (Talker "D"). Her recordings were made in a sound-attenuated room at the Max Planck Institute for Psycholinguistics in the Netherlands, with a 22.05-kHz sampling rate and 24-bit accuracy (see Brouwer et al., 2012). Recordings from both masker talkers were root-mean-square normalized.

Dutch was chosen as the second language because it is acoustically and phonetically very similar to English. Both English and Dutch are from the same linguistic family (Indo-European and West Germanic). They are also from the same rhythmic class (stress-timed), have a similar number of vocalic phonemes (14 and 13 for English and Dutch, respectively), and have a similar number of consonantal phonemes (24 and 26 for English and Dutch, respectively; Booij, 1999). Talker E and Talker D produced 100 sentences in English and Dutch, respectively. The only English and Dutch sentences spoken by Talkers E and D that were included in the maskers were those with similar speaking rates (syllables/second) and average F0 between the two different talkers. This resulted in 24 *pairs* of sentences. For example, one of the 24 *English/Dutch sentence pairs* included an English sentence with an average F0 = 179.3 Hz and a speaking rate of 3.9 syllables/second and a Dutch sentence with an average F0 = 180.0 and a speaking rate of 3.9 syllables/second. Based on the 24 English and Dutch sentences included in the maskers, Talker E had a mean F0 of 177.2 Hz and a speaking rate of 3.7 syllables/second, while Talker D had a mean F0 of 177.5 Hz and a speaking rate of 4.1 syllables/second. Two separate *t*-tests indicated no significant difference between average F0 or speaking rate between Talkers E and D—$t(46) = .10$, $p = .919$ and $t(46) = 1.61$, $p = .115$, respectively.

Two unique single-talker streams were created for each masker Talker (E and D) using different sentence concatenation orders, such that each of the 24 sentences fell at different time points in the two streams. Streams of reversed English speech were generated digitally. There were five two-talker masker conditions in total. Three of the maskers included *congruent streams* (English speech [Eng/Eng], time-reversed English speech [RevEng/RevEng], and Dutch speech [Dutch/Dutch]), while two of the maskers included *incongruent streams* (English *plus* time-reversed English speech [Eng/RevEng] and English *plus* Dutch speech [Eng/Dutch]). Recall that all English masker recordings were made by one talker (Talker E), and all Dutch masker recordings were made by another talker (Talker D). The long-term average speech spectra of the five two-talker maskers were normalized to the grand average long-term average speech spectra of all five maskers. Informal listening tests did not indicate an audible difference between the original and long-term average speech spectra-normalized files, this manipulation controlled for spectral differences across the five masker conditions. The normalization procedure was completed using MATLAB with a 2048-point sampling window.

Procedure. Testing took place in a double-walled, sound-attenuated sound suite (Acoustic Systems) with listeners seated in a comfortable chair facing the observation window. Listeners were instructed to listen for a female talker who would be speaking sentences, while two other female talkers spoke in the background. It was explained that the target voice would begin at a louder level relative to the competing talkers, and that this difference in level should be used to identify the target voice. It was also explained that during the familiarization phase, the target talker voice would become quieter over time relative to the competing talkers, but that the same target voice would be used throughout testing. Stimuli were presented binaurally via ER1 insert headphones (Etymotic; Elk Grove Village, IL). The experimental program was run on a desktop computer using custom software developed using Max software (Cycling '74; Walnut, CA) and a soundcard (M-Audio, Fast Track Pro; Cumberland, RI). The masker was gated on and off with a 500 ms lead and lag time surrounding the presentation of each target sentence. The target speech was presented at a fixed level of 65 dB SPL throughout the experiment, and SNR was manipulated by adjusting the masker level. The SNR was defined relative to the overall level of the masker, such that at 0 dB SNR the target talker was 3 dB higher than either of the individual masker talkers.

During the familiarization phase, listeners heard 20 sentences in each masker, with 5 sentences at each of the following SNRs: +5, 0, −3, and −5 dB. Four of the five two-talker maskers were used during the familiarization phase. The Dutch masker was excluded, as we

anticipated and observed during pilot testing that this masker would be the easiest for our listeners to segregate from the target talker voice. Familiarization was completed before the testing phase began.

All experimental testing was conducted at −5 dB SNR. The five masker conditions (Eng/Eng, Eng/RevEng, Eng/Dutch, RevEng/RevEng, and Dutch/Dutch) were completed in random order, with one BEL list of sentences presented in each condition. Per instructions, listeners repeated back what they heard after each sentence. The experimenter recorded each keyword as either *correct* or *incorrect*, only giving a *correct* score if the word was repeated exactly as it was presented in the target sentence. Changes in morphological endings (e.g., -s plurals, tense changes, etc.) were scored as *incorrect*. The last 22 sentences of each 25-item BEL list were used to calculate performance scores (88 keywords in total) for each masker condition, allowing three sentence trials for the listener to adjust to a change in the masker (Brungart & Simpson, 2007). The average test time for experimental conditions was approximately 20 min.

Participant responses were recorded using a digital audio recorder. A second independent examiner scored the recorded responses. Reliability was not assessed for one of the 20 listeners, as the audio was not recorded during testing due to a failure in the recording device. Average interrater reliability was 94.9%. All disagreements were reevaluated by a third independent examiner who was blinded to the experimental hypothesis; that examiner made a final determination for all keyword scores.

Sentence recognition scores were transformed to rationalized arcsine units (Studebaker, 1985) to stabilize error variance prior to statistical analyses. This transformation was completed due to some performance scores that were below 20% correct. All statistical analyses were conducted using rationalized arcsine units scores; however, percent correct data are also presented later to describe means and standard deviations.

Results

Congruent maskers. A regression analysis with subject as a random factor was conducted to evaluate the main effect of masker condition for all *congruent* maskers (Eng/Eng, RevEng/RevEng, and Dutch/Dutch). The main effect of masker condition was significant, $F(2, 38) = 144.79$, $p < .001$. A post hoc honestly significant difference test revealed that the Dutch/Dutch masker (mean = 73.7%, $SD = 10.7$) was significantly less effective than the RevEng/RevEng masker (mean = 66.4%, $SD = 8.1$), which was significantly less effective than the Eng/Eng masker (mean = 31.6%, $SD = 14.1$; as indicated by letter groupings in Figure 1).

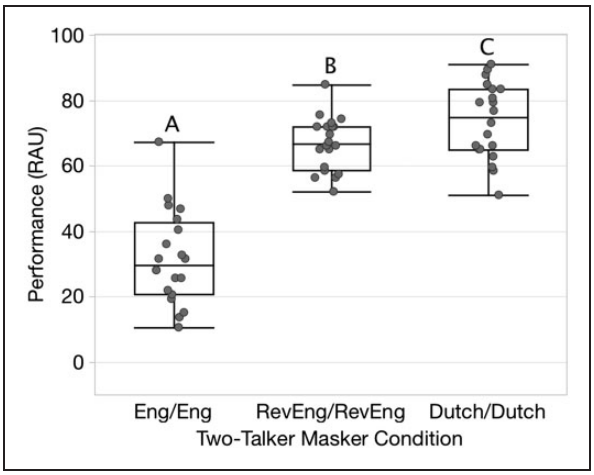

Figure 1. Sentence recognition performance (RAU) for normal-hearing listeners in three congruent two-talker masker conditions (Eng = English, RevEng = time-reversed English). Box plots indicate the median value of the data (indicated by the vertical line within box) and the interquartile range (indicated by the box length). Whiskers have a maximum length of 1.5 times the interquartile range, with the caveat that whiskers never extend past the full range of data. Gray circles indicate individual data points. Boxes under different letter groupings are statistically different from each other.
RAU = rationalized arcsine units.

Incongruent maskers. To determine the effect of masker talker congruency within a two-talker masker, a regression analysis with subject as a random factor was conducted to evaluate masker effectiveness for the two-talker congruent English masker (Eng/Eng) and incongruent two-talker maskers (Eng/RevEng and Eng/Dutch). The main effect of masker condition was significant, $F(2, 38) = 8.95$, $p < .001$. A post hoc Tukey honestly significant difference test revealed that results in the Eng/Eng masker and the Eng/RevEng masker were not significantly different (Eng/RevEng mean = 36.9%, $SD = 14.2$). The Eng/Dutch was significantly a less effective masker (mean = 43.9%, $SD = 13.9$) compared with the Eng/Eng masker, but it was not significantly different from the Eng/RevEng masker. Tukey groupings are depicted in Figure 2. Based on comparisons between listener performance in the congruent and incongruent maskers, a high degree of similarity between the target and at least one of the two masker streams appears to play a dominant role in the masker's overall effectiveness.

Individual differences. Figures 1 and 2 indicate substantial individual differences, with variability across listeners, on the order of 70 percentage points in several cases. Individual differences across the five masker conditions were evaluated by computing the correlation between performance in the Eng/Eng condition and the less effective masker conditions. Pearson correlation

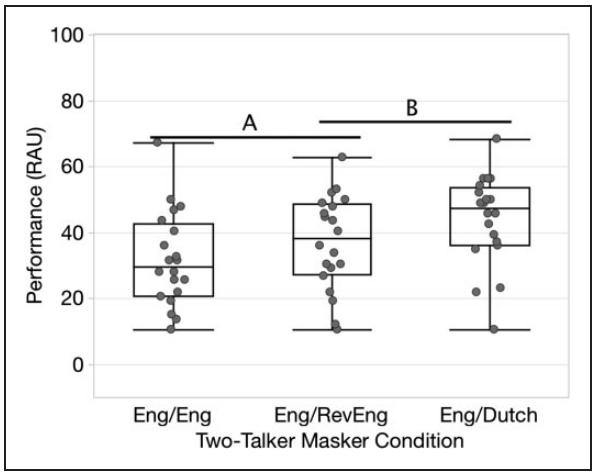

Figure 2. Sentence recognition performance (RAU) for normal-hearing listeners in three two-talker masker conditions that vary in talker congruency (Eng = English, RevEng = time-reversed English). Plotting conventions follow those of Figure 1.
RAU = rationalized arcsine units.

coefficients were $r = .81$ (Eng/Eng vs. Dutch/Dutch), $r = .44$ (Eng/Eng vs. RevEng/RevEng), $r = .57$ (Eng/Eng vs. Eng/Dutch), and $r = .60$ (Eng/Eng vs. Eng/RevEng). The finding of relatively consistent results across conditions within a listener supports the idea that individual differences reflect somewhat consistent differences in the listener's ability to recognize speech in a speech masker, as opposed to stimulus variability or measurement noise.

Keyword level results. Ezzatian et al. (2012) reported that performance as a function of keyword position differed depending on the masker. In that study, performance tended to improve between the first and final keyword for sentences presented in a two-talker masker (whether played time forward or time reversed), but performance worsened between the first and final keyword for sentences presented in a spectrally shaped noise masker. This result was interpreted as showing that the length of time needed to form an auditory stream is longer for a speech-on-speech recognition task using a two-talker masker compared with a speech-in-noise recognition task. On the basis of these results, we were interested in knowing whether a similar trend was evident in any of the conditions of the present experiment. It was predicted, based on the data by Ezzatian et al., that performance would improve between the first and final keywords for the most effective masker, the Eng/Eng masker, which was likely most difficult to segregate from the target speech.

Four performance scores were calculated for each masker condition for every participant, based on keyword position (see Table 1). Since listeners were tested in every masker condition over 22 sentences (88 keywords), these new scores were based on 22 keywords or condition. A regression analysis with subject as a random factor was used to evaluate the main effects of keyword position and masker condition, as well as the interaction of keyword position and masker condition. The main effects of keyword position and masker condition were significant, $F(3, 361) = 41.10$, $p < .001$ and $F(4, 361) = 227.45$, $p < .001$, respectively, while the interaction was not, $F(12, 361) = 0.278$, $p = .992$. Post hoc Tukey analyses indicated that the best performance was observed for Keyword Position 1 (mean = 59.5%, $SD = 20.4$) and the worst performance was observed for Keyword Positions 3 (mean = 46.6%, $SD = 22.9$) and 4 (mean = 44.2%, $SD = 23.0$) for all masker conditions.

Table 1. Percent correct performance for keyword position within a target sentence.

Two-talker masker	Keyword 1	Keyword 2	Keyword 3	Keyword 4
Eng/Eng	40.2 (14.2)	33.9 (13.7)	27.3 (16.4)	25.2 (17.4)
Eng/RevEng	47.5 (14.7)	39.3 (16.5)	31.4 (16.8)	29.5 (14.9)
Eng/Dutch	52.3 (11.6)	43.9 (13.5)	40.5 (18.0)	39.1 (18.0)
RevEng/RevEng	75.5 (12.6)	70.7 (11.9)	63.4 (8.6)	59.8 (11.8)
Dutch/Dutch	82.0 (7.9)	75.0 (10.6)	70.5 (14.0)	67.3 (17.7)

Note. Means are reported, with *SD*s indicated in parentheses.

To determine how local SNR affected performance across keyword position, an analysis was conducted using Praat to determine the average level of the first and fourth keywords for every target sentence used in the experiment. A paired *t*-test indicated that the level of the fourth keyword was significantly lower than the first keyword, $t(124) = 14.53$, $p < .001$. The first keyword had an average level of 66.4 dB SPL ($SD = 1.1$ dB), while the fourth keyword had an average level of 63.5 dB SPL ($SD = 1.5$ dB). This reduction in level across keywords could explain the failure to replicate the improvement in performance as a function of keyword position for a speech-on-speech recognition task previously observed by Ezzatian and coworkers (Ezzatian, Li, Pichora-Fuller, & Schneider, 2015; Ezzatian et al., 2012).

Discussion

The purpose of this experiment was to determine, for an open-set sentence recognition task, the effect of masker-talker congruency on two-talker masker effectiveness. The data reported earlier suggest that if one of the two talkers in a two-talker masker is perceptually similar to the target speech, the masker will be highly effective. Compared with the congruent English masker (Eng/Eng), performance improved by 42.1 and 34.8

percentage points, respectively, with the congruent Dutch (Dutch/Dutch) and congruent time-reversed English (RevEng/RevEng) maskers. Despite this difference between congruent maskers, incongruent maskers with one English talker were relatively similar in their masker effectiveness compared with the congruent Eng/Eng masker. Performance improved by only 12.3 and 5.3 percentage points with the incongruent Eng/Dutch and Eng/RevEng maskers compared with the Eng/Eng masker, respectively; this improvement was significant for Eng/Dutch but not for Eng/RevEng. While masker type had a robust effect for congruent maskers, it had a much smaller effect for incongruent maskers that included one English talker.

Congruent versus incongruent. Iyer et al. (2010) systematically explored the effects of incongruent speech maskers for two-talker masker combinations for a CRM task. The main focus of their work was to explore the "multimasker penalty" (Durlach, 2006). The multimasker penalty has been described with respect to performance observed for speech-on-speech recognition where a second competing talker is added to the auditory scene, and performance decreases more than predicted based on the additional energetic masking contributions of the second talker (Carhart et al., 1969; Durlach, 2006; Iyer et al., 2010; Kidd, Mason, Best, & Marrone, 2010). It is difficult to make direct comparisons between the data of the present study with those reported in Iyer et al. due to the differences between the tasks and the stimuli. In contrast to the present study, which used an open-set sentence recognition task, Iyer et al. used the closed-set CRM task. In that published study, maskers described as contextually relevant to the target were other CRM phrases, and maskers described as contextually irrelevant included randomly selected English readings, Dutch speech, Spanish speech, and time-reversed versions of all of the speech materials listed here. However, these procedural differences notwithstanding, the results are generally consistent across studies. Iyer et al. noted that the specific features of the second talker did not play a large role in determining masker effectiveness of a two-talker masker stream; if one of the two talkers were highly similar to the target speech, then that masker would be nearly as effective as a two-talker masker consisting of two CRM phrases.

Differences observed in masker effectiveness for the congruent maskers in the present study are consistent with previous literature. Each of the two streams of the congruent English masker was highly similar to the target speech, so it is no surprise that it was highly effective. It has been widely demonstrated that similarity between the target and masker speech increases confusion, and therefore masking (Moore & Gockel, 2012). The two congruent maskers composed of speech that was dissimilar to the target speech (RevEng/RevEng and Dutch/Dutch) were approximately 35–40 percentage points less effective. This improvement in recognition is quite large, especially given the fact that vocal characteristics, including F0 and speaking rate were very similar across maskers, which would increase perceptual similarity. Both the Dutch and time-reversed English congruent maskers were unintelligible to the listeners. However, there is a growing body of evidence indicating that intelligibility is not of critical importance for speech-on-speech masking. For example, marked reductions in masking have been observed when the masker speech differs from the target with respect to the talker's accent—despite high levels of intelligibility (Calandruccio et al., 2010), or when the target and masker are spoken in different languages that are both intelligible to the listener (Calandruccio & Zhou, 2014). These results suggest that acoustic differences between the target and masker speech may be the critical variable responsible for reduced masking rather than masker intelligibility.

It is not clear why the congruent RevEng/RevEng masker is more effective than the congruent Dutch/Dutch masker, a difference that is also observed for the associated incongruent maskers (Eng/RevEng and Eng/Dutch). Both the congruent time-reversed English and congruent Dutch maskers sound very perceptually different than the congruent English masker. However, it is not uncommon to see reports of time-reversed speech causing greater masking than would be expected due to the dissimilarity of time-reversed speech to time-forward target speech (Brungart & Simpson, 2007; Kidd, Mason, Swaminathan, Roverud, Clayton, & Best, 2016; Summers & Molis, 2004). Even when masking release is observed with a time reversed compared with a time-forward masker, it has been suggested that the reversed speech can cause increased forward masking (Rhebergen et al., 2005). Rhebergen et al. (2005) provided a very nice example of increased masking with time-reversed speech by presenting listeners with a foreign language masker played both time forward and time reversed. For their experiment, the listeners were native speakers of Dutch. Competing Swedish speech, played time forward and time reversed, served as the maskers. Regardless of the direction in time the competing speech was played, it was unintelligible to their listeners. Rhebergen et al. observed that the time-reversed Swedish speech caused just over a 2 dB more masking then forward Swedish. Rhebergen et al. attributed this difference to greater forward masking with the time-reversed masker, due to changes in the temporal envelope. This interpretation is consistent with the present data, where the time-reversed English masker was more effective than the Dutch masker.

Auditory stream segregation. Informational masking due to target and masker similarity is thought to reflect difficulties segregating the target and masker streams (Bregman, 1990). There is some evidence for speech-on-speech recognition tasks indicating that segregation of the target and masker speech may improve over time with stimulus exposure. For example, Ezzatian et al. (2012, 2015) reported that performance improved as a function of keyword position for a two-talker masker, whether that masker was played time forward or time reversed. In contrast, there was no effect of keyword position when the same sentences were played in noise, noise vocoded speech or a two-talker masker that was perceived to be spatially separated from the target. This differential effect of keyword position was interpreted as reflecting a delay in the buildup of time required to segregate the target from a perceptually similar masker, a task that becomes easier when the masker is noise based or when spatial segregation cues are provided. An analogous finding was reported by Richards, Shub, and Carreira (2011) for tone-burst stimuli. The task in that study was to detect a series of tone bursts presented synchronously with random-frequency masker bursts; performance was better when the train of masker bursts began before the signal. The benefit of a masker fringe could be larger in cases of signal uncertainty compared with cases where the signal is predictable (Wright & Dai, 1994). Based on these findings, it was hypothesized that improvements in performance over keyword position may serve as an indicator of a delayed buildup of stream segregation for target and masker combinations that are easily confused with each other.

In contrast to the prediction, an analysis of the present data indicated that performance *decreased* as a function of keyword position. Specifically, for all masker conditions, the first keyword position had significantly higher scores, followed by the second keyword position. The third and fourth keywords had significantly lower scores than the first two keywords. An analysis of target stimulus level across keyword position revealed that on average level of the fourth keyword was lower than the first keyword. Therefore, the poorer performance in the fourth keyword position relative to the first keyword position can be at least partially explained in terms of the *local* SNR. The target BEL sentences are all declarative sentences, and declarative sentences in English often fall in level (Cruttenden, 1997; Lieberman, 1967). The sentences used in Ezzatian et al. (2012) were also declarative and decreased in level as a function of keyword position. In the original paper that described the stimuli creation, a decrease in level as a function of keyword position was noted; however, the magnitude of that effect was not reported. Although worsening in the local SNR with keyword position is broadly consistent with the failure to find improvement in performance for later keywords in the present study, a more rigorous evaluation of this explanation requires additional information about the psychometric function associated with each keyword position. That analysis is presented as Experiment 3.

Experiment 2—Effectiveness of One-Talker Masker Streams

Rationale

To better understand the role of target and masker similarity, a follow-up experiment was conducted examining sentence-recognition performance in the presence of the three different one-talker maskers that were used to create the two-talker streams in Experiment 1. On the basis of the congruent masker data of Experiment 1, we predicted that the one-talker English masker would be more effective than the one-talker Dutch masker, and the one-talker-reversed-English masker would be intermediate. In addition to providing further evidence of masker effectiveness, the one-talker data are of interest to better understand the magnitude of the multimasker penalty.

Methods

Participants. Eighteen newly recruited adult native-English-speaking listeners (13 women and 5 men) with normal-hearing thresholds were recruited for participation in Experiment 2 (age 18–33 years old, mean age = 21 years, $SD = 3.7$ years).

Procedures. Procedures, test environment, software, and hardware were identical to those used in Experiment 1. Target stimuli were BEL sentences (lists 1, 3, 4, 8, 9, 10, 13, 14, 15, 17, 18, and 20). Masker stimuli consisted of the three, one-talker speech streams that were used to create the two-talker maskers in Experiment 1, which included English, reversed English, and Dutch. A traditional method of constant stimuli was employed with SNR blocked by BEL list. Stimuli were presented at −12, −15, −18, and −21 dB SNR, with the target speech fixed at 60 dB SPL. The target speech was presented at 60 dB SPL, and not 65 dB SPL as in Experiment 1 and 3, to ensure a comfortable listening level while employing a −21 dB SNR. Prior to experimental testing, listeners were familiarized with the task and the target talker with 25 total practice sentences presented at −5, −10, and −12 dB SNR. All three, one-talker maskers were presented, utilizing one masker per SNR. The average experimental test time was approximately 40 min. Average interrater reliability for all trials was 98.6%, with a third-independent examiner, naïve to the study hypotheses, determining the final scores for those trials with discrepancies between raters.

Results

Performance scores were calculated from the last 22 sentences of each BEL list. Data for each listener and masker condition were fitted with a logit function by minimizing the sum of squared error. Unlike Experiment 1, which looked at data in terms of percent correct tested at a fixed SNR, these data are analyzed in terms of 50% threshold performance (dB SNR). The distribution of thresholds in each of the three maskers is shown in Figure 3. As observed in Experiment 1, there were sizeable individual differences: Performance at −15 dB SNR spanned a range of approximately 50 percentage points across listeners, and thresholds spanned a range of approximately 8 dB. It should be noted that one listener's performance for the English one-talker masker was excluded from the analysis because the function fit was quite poor ($r^2 = -.12$). In contrast, all other fits were quite strong (median fit for the remainder of the data, $r^2 = .95$). Further, one other listener was excluded from the analysis due to an experimental error during testing.

A regression analysis with participant as a random effect indicated a significant effect of single-talker masker condition, $F(2, 31) = 6.63$, $p = .004$. Post hoc Tukey testing indicated that the Eng masker was most effective (mean threshold = −15.5 dB SNR) but was not significantly different than the RevEng stream (mean threshold = −16.4 dB SNR). The Dutch single-talker masker was least effective (mean threshold = −16.9 dB SNR) but was not significantly different than the RevEng. Relative to the English masker, performance with the Dutch masker reflected a small, significant (1.3 dB) linguistic-masking release. However, playing the English single-talker masker stream time forward or time reversed did not significantly affect performance (see Figure 3).

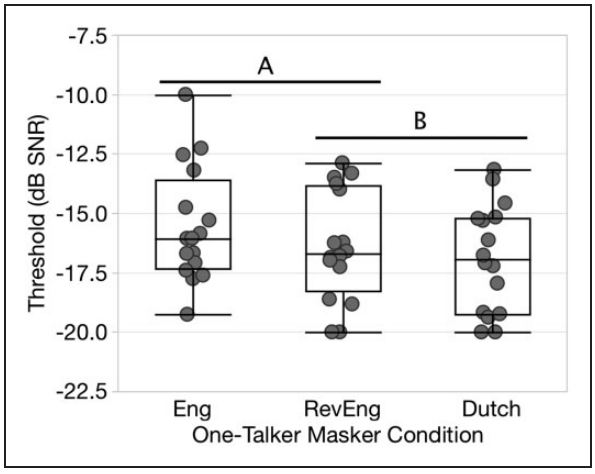

Figure 3. Speech-recognition thresholds for listeners with normal-hearing thresholds in three different one-talker masker conditions (English, time-reversed English, and Dutch). Plotting conventions follow those of Figure 1.

Discussion

The wide range in variability across listener thresholds is consistent with the literature for a one-talker masker (e.g., Miller, 1947), despite the fact that the task is significantly easier than a two-talker masker task. With respect to the multimasker penalty, the addition of the second masking talker in Experiment 1 caused the sentence recognition task to be much more difficult. This is illustrated in Figure 4, showing mean percent correct values from Experiment 1 and psychometric functions fitted to mean data in Experiment 2. The three two-talker maskers that include at least one stream of English are associated with a mean value of 36.8% correct at −5 dB SNR. Based on psychometric function fits, comparable performance in the one-talker English masker is associated with approximately −19 dB SNR; this 14-dB threshold difference (−19 minus −5) is one way of quantifying the multimasker penalty. In contrast to the large multimasker penalty observed for two-talker maskers that included at least one stream of English, the multimasker penalty for the Dutch/Dutch and RevEng/RevEng maskers was only 7 dB. This highlights the role of target and masker similarity in the multimasker penalty.

Experiment 3—Keyword Performance as a Function of Keyword Position in Relation to the Local SNR

Rationale

In Experiment 1, listeners were tested at a fixed SNR of −5 dB. Performance significantly decreased as a function

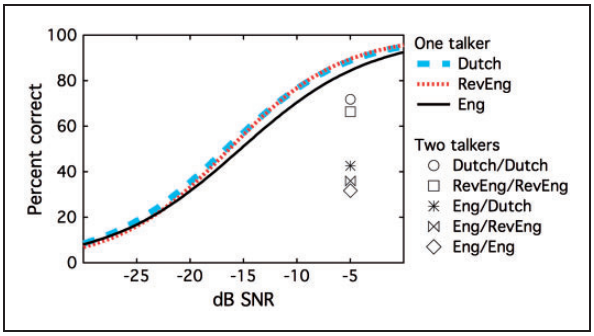

Figure 4. Sentence recognition (percent correct) as a function of SNR for the three one-talker maskers tested in Experiment 2. Mean data for Experiment 1 (two-talker maskers) are also shown at a fixed SNR of −5 dB.
SNR = signal-to-noise ratio.

of keyword position. However, keyword level also decreased as a function of keyword position. In Experiment 3, listeners were tested across a range of SNRs, so that a threshold (dB SNR) could be calculated for each keyword position. This allows a comparison of masking for the fourth keyword position relative to the first keyword position, taking into account the reduction in keyword level, which is typical of declarative target sentences. It was hypothesized that if there was a build-up in stream segregation across the length of each sentence trial, then an improvement in threshold as a function of keyword position would be observed once the reduction in stimulus level was accounted for.

Methods

Participants. The same cohort of 18 adults who completed Experiment 2 also completed Experiment 3.

Procedures. Test procedures, test environment, software, hardware, and instructions were identical to those used in Experiment 1. Target sentences included four lists of BEL sentences (lists 11, 12, 16, and 19); participants had not previously heard these sentences. The Eng/Eng masker condition from Experiment 1 was used for this experiment. This masker was chosen as it was the most effective of the two-talker maskers employed in Experiment 1, and as such likely caused the greatest amount of informational masking. The target speech was presented at a fixed level of 65 dB SPL throughout the experiment, and the presentation level of the masker stimuli varied based on the SNR. To familiarize the listeners with the task they were presented with 15 practice sentences, with 10 sentences available for additional practice if needed (e.g., if listeners had difficulty hearing out the target talker amongst the two competing talkers). During the familiarization period, listeners were presented with five sentences at the following SNRs in a descending order: +5, 0, and −1 dB. Immediately following the familiarization, the experiment began.

For data collection, sentences were presented for four SNRs: −1, −3, −5, and −7 dB. The four BEL sentence lists were presented in random order, with one list per SNR, and with SNRs tested in descending order (easiest to most challenging). A descending presentation order was used to improve performance (e.g., Brouwer et al., 2012); pilot data for this experiment indicated that performance at −7 dB SNR depended strongly on prior listening experience at more favorable SNRs.

Testing took an average of 15 min. Listener responses were recorded and scored offline for reliability. Average interrater reliability for all trials was 99.1%, with an independent third examiner determining the final score for any discrepancies between the two raters (online and offline scoring).

Results

Data across the four SNRs were fitted with a logit function by minimizing the sum of squared error. Thresholds were estimated individually for each participant at each of the four keyword positions. Four of the 18 listeners recruited could not achieve at least 50% performance at any of the SNRs tested. Estimates of 50% threshold are less accurate without data points above and below the 50th percent point, and therefore, their data were excluded. As such, data for Experiment 3 are presented for 14 adult listeners (18–33 years old; mean age = 22 years old, $SD = 4.0$ years; 10 girls and 4 boys). The distributions of these thresholds are shown in Figure 5. Without taking the effect SNR for each keyword into

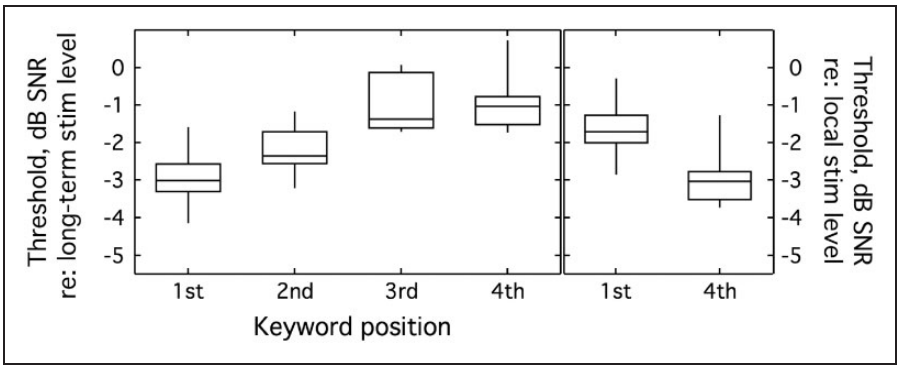

Figure 5. Speech-recognition thresholds (dB SNR) for 50% performance in the Eng/Eng two-talker masker condition as a function of keyword position. Thresholds on the left indicate increased thresholds (poorer performance) as a function of keyword position in dB SNR relative to the long-term average target presentation level of 65 dB SPL; however, data on the right depict speech-recognition thresholds once local SNR were taken into account, indicating an improvement in performance between the first and final keyword position.
SNR = signal-to-noise ratio.

account, the SNR (dB) required to achieve threshold performance of 50% correct significantly increased across keyword position, similar to the results observed in Experiment 1. This is shown in the left panel of Figure 5. A regression analysis with subject as a random effect indicated a significant effect of keyword position, $F(3, 39) = 25.86$, $p < .001$, with the poorest performance for Keyword Positions 3 and 4 (average thresholds = −1.0 and −.9 dB SNR, respectively). The best performance was observed for Keyword Position 1 (mean threshold = −2.9 dB SNR; see left panel of Figure 5). As in Experiment 1, analyses of the target stimuli revealed that the level of the fourth keyword was significantly lower than that of the first keyword—, $F(1, 176) = 405.55$, $p < .001$; mean level = 66.3 and 63.0 dB SPL for the first and fourth keyword, respectively. On average, the final keyword position was 3.4 dB less intense than the first keyword. Taking this level difference into account, thresholds were replotted in local SNR (see right panel of Figure 5). A paired t-test comparing thresholds in local SNR indicate that performance improved between the first and fourth keyword position, $t(13) = 15.64$, $p = .002$.

Discussion

The data from Experiment 3 confirm the prediction that performance improves across keywords for speech presented in the Eng/Eng masker once reductions in target level across keywords are taken into account. This change in performance is consistent with the results of Ezzatian et al. (2012, 2015), indicating that segregation of the target and masker speech improves over the time course of the trial; nevertheless, the overall magnitude of this improvement was likely less for this combination of target and masker stimuli than that observed by Ezzatian et al. since it was observed in their data without taking into account local SNR. An improvement in performance across keywords is consistent with the idea that auditory stream segregation builds up over time. However, one caveat is that the present experiment did not include a condition in which build-up of segregation was not expected (e.g., performance in a speech-shaped noise maker). More work is needed to understand the conditions under which performance improves over keywords.

General Discussion

In the first experiment, greater masking was observed when at least one of the two talkers within a two-talker masker was perceptually similar to the target. That is, whether the two talkers within the two-talker masker streams were congruent to each other or incongruent was minimally important; as long as one of the two talkers was perceptually similar to the target speech, an increase in masking was observed. Large decreases in masking were only observed for congruent two-talker masker streams that were perceptually dissimilar to the target speech (RevEng/RevEng and Dutch/Dutch).

Multimasker Penalty

Experiment 2 measured the effectiveness of the single-talker masker streams used to create the congruent and incongruent two-talker maskers used in Experiment 1. Recognizing speech in the presence of one competing talker is significantly easier compared with two competing talkers (Brungart et al., 2001; Carhart et al., 1969; Festen & Plomp, 1990; Miller, 1947). This decrease in performance between one and two competing talkers cannot be explained by energetic masking alone (Carhart et al., 1969; Iyer et al., 2010). As expected, performance in the three, one-talker maskers (Experiment 2) was better than performance in the two-talker masker (Experiment 1). Although the single-talker English masker stream was more effective than the single-talker Dutch stream, the difference was quite modest. Evaluating performance at −15 dB SNR, near the middle of the psychometric functions, performance differed by only 5.9 percentage points between English and Dutch conditions for the single-talker maskers; in contrast, performance in congruent Eng/Eng and Dutch/Dutch two-talker maskers differed by 40 percentage points.

For the masker stimuli used in these experiments, a multimasker penalty was observed for all maskers, regardless of masker talker congruency or target and masker similarity. Although this finding is not in full agreement with those reported in Iyer et al. (2010), which indicated that the multimasker penalty only occurred for negative SNRs *and* when at least one of the two competing talkers was perceptually similar (what they referred to as *contextually relevant*), it should be noted that the overall conclusions were quite similar between the two studies. Similar to the study by Iyer et al., a multimasker penalty was observed in the current experiments, which evaluated performance at negative SNRs. The task and the stimuli used in Iyer et al. were very different from those used in this study (e.g., closed- vs. open-set task; CRM vs. meaningful sentences). In the present study, all maskers—similar and dissimilar—consisted of concatenated recordings of isolated sentences, similar in structure to the target sentences. In the study by Iyer et al., irrelevant maskers included excerpts from connected discourse, which were likely perceptually *very* different from the target (CRM) stimuli on a number of dimensions (e.g., syntax, prosody, and semantics). It is possible that dissimilar target and masker stimuli in the current

experiment were nonetheless more similar to one another than those employed by Iyer et al. Therefore, it is possible that all of the stimuli used here—including the Dutch and reversed-English speech—were at least somewhat *contextually relevant*, and this perceptual similarity could have played a role in the multimasker penalty observed for all maskers.

The ~7-dB multimasker penalty observed for the dissimilar congruent two-talker maskers employed in Experiment 1 was of similar magnitude to that reported by Carhart et al. (1969). The magnitude of the multimasker penalty for the congruent, similar (Eng/Eng) and incongruent (Eng/Dutch and Eng/RevEng) maskers used in this study was very large, nearing 15 dB SNR. It is not clear why the Eng/Eng masker was so effective, especially since the English single-talker masker was minimally different than the other two-talker streams.

Iyer et al. (2010) postulated that a limit in attentional resources might account for the finding that one masker stream within a two-talker masker that is similar to the target dominates the masker's effectiveness. In the case of the less effective congruent two-talker maskers, listeners are quickly cued in to the differences between the target and the masker speech, allowing for improved performance in sentence recognition. However, when the target is perceptually similar to masker speech, whether it would be one *or* two similar talkers, it may take additional cognitive resources and time to determine which speech is the target and which is the masker, resulting in poorer performance. Perhaps, the attentional resources required to suppress a dissimilar competing talker and segregate a similar competing talker are comparable to the attentional resources required to segregate two similar competing talkers. However, more research is needed to explore this idea further.

Stream Segregation in a Two-Talker Masker

Experiment 3 allowed us to investigate speech-on-speech recognition as a function of sentence keyword position while taking into account the local SNR associated with each keyword. Although we did observe that once the local SNR was accounted for, the mean threshold for Keyword 4 was significantly lower than threshold for the Keyword 1, the magnitude of the effect was still small (an average of 1.4 dB improvement in threshold). Ezzatian et al. (2012) reported an average improvement of approximately 2 dB between their first and final keyword positions, and they did not take into account reductions in local stimulus level at the end of their declarative, anomalous target sentences. More work is needed to understand the conditions under which performance improves over keywords for speech-on-speech recognition.

Conclusions

1. Congruency between two masking talkers minimally influenced the overall effectiveness of a two-talker masker for an open-set, sentence recognition task.
2. For a two-talker masker, masker effectiveness increased with greater perceptual similarity between the target and at least one of the masking talkers.
3. The multimasker penalty was observed for perceptually similar and perceptually dissimilar two-talker maskers, regardless of masker congruency. However, the multimasker penalty was much larger when at least one of the two talkers within a two-talker masker was perceptually similar to the target talker compared with congruent two-talker maskers that were perceptually dissimilar to the target speech.
4. Stream segregation appears to improve performance over the duration of a single declarative sentence when evaluated relative to the local SNR for a congruent two-talker masker that is perceptually similar to the target speech.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding support for this project is provided by the NIH-NIDCD (Grants Nos. R03DC015074 and R01DC007391).

References

American Speech-Language-Hearing Association. (2005). *Guidelines for manual pure-tone threshold audiometry*. Retrieved from www.asha.org/policy

Ben-David, B. M., Tse, V. Y. Y., & Schneider, B. A. (2012). Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker? *Hearing Research*, *290*(1), 55–63.

Booij, G. E. (1999). *The phonology of Dutch*. Oxford, England: Oxford University Press.

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

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.

Brungart, D. S., & Simpson, B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Perception & Psychophysics*, *69*(1), 79–91.

Brungart, D. S., Simpson, B. D., Ericson, M., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1102–1109.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2013). Masking release due to phonetic and linguistic dissimilarity between the target and masker speech. *American Journal of Audiology*, *22*, 157–164.

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

Calandruccio, L., & Smiljanić, R. (2012). New sentence recognition materials dveloped using a basic non-native English lexicon. *Journal of Speech, Language, and Hearing Research*, *55*(5), 1342–1355.

Calandruccio, L., & Zhou, H. (2014). Increase in speech recognition due to linguistic mismatch between target and masker speech: Monolingual and simultaneous bilingual performance. *Journal of Speech, Language, and Hearing Research*, *57*(3), 1089–1097.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America*, *45*(3), 694–703.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Cruttenden, A. (1997). *Intonation*. Cambridge, England: Cambridge University Press.

Dirks, D. D., & Bower, D. R. (1969). Masking effects of speech competing messages. *Journal of Speech and Hearing Research*, *12*(2), 229–245.

Durlach, N. (2006). Auditory masking: Need for improved conceptual structure. *The Journal of the Acoustical Society of America*, *120*, 1787–1790.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, *58*(9), 955–991.

Elliott, L. L., Connors, S., Kills, E., & Levin, S. (1979). Children's understanding of monosyllabic nouns in quiet and in noise. *Journal of the Acoustical Society of America*, *66*, 12–21.

Ezzatian, P., Li, L., Pichora-Fuller, K., & Schneider, B. A. (2015). Delayed stream segregation in older adults: More than just informational masking. *Ear and Hearing*, *36*(4), 482–484.

Ezzatian, P., Li, L., Pichora-Fuller, M. K., & Schneider, B. A. (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.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *109*(5), 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. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256.

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

Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, *119*(4), 2445–2454.

IEEE Subcommittee on Subjective Measurements (1969). IEEE recommended practice for speech quality measurements. *Standards Publication No. 297. IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246.

Iyer, N., Brungart, D. S., & Simpson, B. D. (2010). Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task. *The Journal of the Acoustical Society of America*, *128*(5), 2998–3010.

Kidd, G., Mason, C. R., Best, V., & Marrone, N. (2010). Stimulus factors influencing spatial release from speech-on-speech masking. *The Journal of the Acoustical Society of America*, *128*(4), 1965–1978.

Kidd, G., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *The Journal of the Acoustical Society of America*, *95*(6), 3475–3480.

Kidd, G., Mason, C. R., Rohtla, T. L., & Deliwala, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, *104*(1), 422–431.

Kidd, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, *140*(1), 132–144.

Lieberman, P. (1967). *Intonation, perception, language*. Cambridge, MA: MIT. Press.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105–129.

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 919–931.

Moore, T. J. (1981). Voice communication jamming research. In *AGARD Conference Proceedings 311: Aural Communication in Aviation* (pp. 2:1–2:6). Neuilly-Sur-Seine, France: AGARD.

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

Richards, V. M., Shub, D. E., & Carreira, E. M. (2011). The role of masker fringes for the detection of coherent tone pips. *The Journal of the Acoustical Society of America*, *130*(2), 883–892.

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, *133*(4), 2431–2443.

Studebaker, G. A. (1985). A ''rationalized'' arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462.

Summers, V., & Molis, M. R. (2004). Speech recognition in fluctuating and continuous maskers: Effects of hearing loss and presentation level. *Journal of Speech, Language, and Hearing Research*, *47*(2), 245–256.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, *121*(1), 519–526.

Watson, C. S., Kelly, W. J., & Wroton, H. W. (1976). Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty. *The Journal of the Acoustical Society of America*, *60*(5), 1176–1186.

Wright, B. A., & Dai, H. (1994). Detection of unexpected tones in gated and continuous maskers. *The Journal of the Acoustical Society of America*, *95*(2), 939–948.