Spatially separating language masker from target results in spatial and linguistic masking release

Navin Viswanathan, Kostas Kokkinakis, and Brittany T. Williams

Citation: The Journal of the Acoustical Society of America **140**, EL465 (2016);
View online: https://doi.org/10.1121/1.4968034
View Table of Contents: http://asa.scitation.org/toc/jas/140/6
Published by the Acoustical Society of America

Spatially separating language masker from target results in spatial and linguistic masking release

Navin Viswanathan,[a] Kostas Kokkinakis, and Brittany T. Williams
Speech-Language-Hearing: Sciences and Disorders, University of Kansas, 1000 Sunnyside Avenue, Lawrence, Kansas 66045, USA
navin@ku.edu, kokkinak@ku.edu, btw@ku.edu

Abstract: Several studies demonstrate that in complex auditory scenes, speech recognition is improved when the competing background and target speech differ linguistically. However, such studies typically utilize spatially co-located speech sources which may not fully capture typical listening conditions. Furthermore, co-located presentation may overestimate the observed benefit of linguistic dissimilarity. The current study examines the effect of spatial separation on linguistic release from masking. Results demonstrate that linguistic release from masking does extend to spatially separated sources. The overall magnitude of the observed effect, however, appears to be diminished relative to the co-located presentation conditions.

[RS]
Date Received: September 20, 2016 **Date Accepted:** November 4, 2016

1. Introduction

Human listeners are routinely faced with the task of recognizing speech in the presence of competing background speech. Speech perception under such conditions is affected primarily by a combination of energetic masking (the reduction in intelligibility due to the spectro-temporal overlap of the target and background masker) and informational masking (reduction in intelligibility that occurs due to a combination of factors that reduces intelligibility after the effect of energetic masking has been accounted for). Specifically, in speech-in-speech recognition tasks, listeners often experience a boost in performance when the background speech differs along specific linguistic dimensions from the target speech (e.g., Calandruccio *et al.*, 2010). While different terms such as (*same*) *language interference effect* (Cooke *et al.*, 2008; Mattys *et al.*, 2009) and *target-masker language mismatch effect* (e.g., Brouwer *et al.*, 2012) have been previously used to characterize this effect, we opt to use the term *linguistic release from masking* (LRM; similar to Calandruccio *et al.*, 2016) to describe the performance improvement resulting from the target-masker linguistic mismatch. There are two primary reasons LRM may occur. First, LRM may occur because target and background speech differ in their acoustic-phonetic characteristics as in the case of different languages or different accents. This is because, relative to same language maskers, both energetic masking effects (due to lower spectro-temporal overlap between the speech sources) and informational masking effects (due to ease of perceptual segregation) are reduced. Second, LRM may reflect decreased semantic interference due to the lack of intelligibility of the competing background and thus a reduction of informational masking.

A number of previous studies have described a robust LRM benefit across different target-masker language pairs (e.g., Brouwer *et al.*, 2012), although the detection of this effect may depend on specific task demands placed on the listener (Mattys *et al.*, 2009). Furthermore, other recent studies, have shown that the similarity in the linguistic properties of the target and background languages is an important factor toward determining the prevalence of LRM benefits (Calandruccio *et al.*, 2010; Brouwer *et al.*, 2012). Interestingly, even in cases, where the background masker consists of accented versions of the target language, listeners show a reliable LRM benefit which is modulated by the intelligibility of the background accented speech (Calandruccio *et al.*, 2010). The aforementioned studies collectively indicate that LRM is a robust phenomenon that is, in part, determined by the target-masker linguistic similarity. Despite this strong evidence, it is unclear whether under more ecologically typical scenarios, listeners would actually experience LRM. Most experimental studies assessing linguistic release have used spatially co-located (i.e., originating from the

[a]Author to whom correspondence should be addressed.

same spatial location) target-masker conditions [cf. Freyman *et al.* (2001) used simulated spatial separation] that are ecologically atypical and may overestimate the overall LRM benefit because they do not permit source segregation based on spatial information and forces the listener to rely exclusively on acoustic-phonetic differences.

The effects of spatial-separation of target and masker on speech recognition is well-documented. For instance, it is clear that actual physical differences, or even perceived differences, between the individual spatial locations of a target and a masker result in a substantial increase of speech intelligibility compared to situations where the target and masking speech are co-located. This benefit, known as *spatial release from masking* (SRM), describes the reduction in masking that occurs in the presence of an active masker due to the availability of spatial acoustic cues. This type of masking release is fairly robust in normal-hearing listeners (e.g., Freyman *et al.*, 2001) and can produce substantial differences in speech perception.

As noted earlier, whether LRM occurs when the target and the masker are spatially separated remains to be established. In fact, the findings of Freyman *et al.* (2001) offer a reason to doubt whether LRM persists under spatial separation. In experiment 3, the authors investigated whether SRM depended on the intelligibility of background language. They tested monolingual English listeners using speech backgrounds produced by a bilingual Dutch-English speaker that were either produced in Dutch or Dutch-accented English. Critically, these backgrounds were presented either with or without simulated spatial separation. Their results indicated clear SRM effects in both language conditions. However, because they were not focused on LRM, they did not evaluate whether Dutch maskers produced better performance than English or the effect of spatial separation on the resulting LRM. A visual inspection of their results (see Freyman *et al.*, 2001, Fig. 9, p. 2120) suggests a clear LRM for higher signal-to-noise ratios (SNRs) when the speech sources were co-located. However, there appears to be no LRM for the spatially displaced condition.

Motivated by these findings and by the general observation that in typical listening situations multiple masking sources often originate from different locations around a listener, we argue that a specific examination of whether LRM effects extend to spatially displaced speech sources is warranted. In the present study, similar to Freyman *et al.* (2001), we used Dutch language backgrounds to examine whether listeners can benefit from LRM under three spatial configurations. Specifically, we maximized the spatial separation by using spatially displaced conditions to account for any asymmetry in addition to a spatially co-located condition. Our rationale is that, by using relatively similar languages (thereby minimizing LRM by minimizing acoustic-phonetic differences; see Calandruccio *et al.*, 2013) under conditions of maximal spatial separation (thereby maximizing SRM), we provide the strongest test of whether LRM effects persist under spatially separated listening conditions.

2. Methods

2.1 Participants

Forty-two undergraduate students from the University of Kansas with American English as their first language and no exposure to Dutch participated in this study for course credit. All subjects reported no history of speech or hearing disorders. Median age of subjects was 19 years (ages from 18 to 21 years).

2.2 Materials

Three female, native English speakers produced the English stimuli. The target stimuli consisted of English sentences taken from the revised Bamford-Kowal-Bench (BKB-R) Standard Sentence Test (Bench *et al.*, 1979). Each sentence contained three or four keywords (e.g., *The CLOWN had a FUNNY FACE*). Following Brouwer *et al.* (2012), the English background stimuli consisted of syntactically well-formed, semantically simple sentences that were produced by the two speakers who did not record the target sentences. Translated versions were used to elicit the Dutch background stimuli that were recorded by a pair of female, Dutch, native speakers. All stimuli were recorded in a sound-attenuating booth at a 44.1 kHz sampling rate and a 16-bit resolution.

Two-talker babble maskers were created by combining waveforms of each pair of speakers for each language. Following Brouwer *et al.* (2012), we reduced unequal amounts of energetic masking between the two language conditions by minimizing differences in the long-term average speech spectrum (LTASS) of the two background speech tracks. Pilot testing ensured that the overall amount of spectral manipulation

was minimal and that the original and normalized files were not easily distinguishable. Random portions of babbles were excised from this babble to serve as background stimuli for each trial. The background stimuli always preceded the target onset by 0.5 s and persisted 0.5 s after the target offset. The root-mean-square (RMS) levels of the target sentences were all equalized to 66 dB sound pressure level (SPL) and that of the background maskers were equalized to 71 dB SPL and 76 dB SPL to produce SNRs of −5 dB and −10 dB, respectively. An SNR equal to −5 dB results in a fairly challenging listening task and produces conditions with strong informational masking (e.g., see Calandruccio *et al.*, 2010). The −10 dB SNR condition was chosen to increase task difficulty while simultaneously ensuring that there were no floor effects in task performance. Finally, for each target-masker configuration, three different spatial listening conditions were created. In all conditions, the target was placed in front of the listener (0°), similar to a typical conversational scenario, and the masker originated from either the front of the listener (0°), the right of the listener (+90°) or the left of the listener (−90°). To place the target speech in front of the listener and the masker in different virtual locations in the horizontal plane, we used publicly available pre-recorded head-related transfer functions (HRTFs). These HRTFs were recorded inside an anechoic chamber at the University of Oldenburg (Kayser *et al.*, 2009) using behind-the-ear hearing aid dummy shells placed on the pinnae of an artificial head-and-torso simulator. The left and right ear (binaural) stimuli containing the combined auditory target and masker streams were generated for each spatial location in MATLAB by convolving each target and masker stimulus with each pair of the direction-dependent HRTFs (−90°, 0°, +90°). Subjects in pilot testing were successful in externalizing the auditory image thus confirming that most perceptually significant acoustic cues were retained in the HRTFs.

2.3 Procedure

Listeners were tested individually, seated in front of a computer in a sound-isolated room. Stimuli were presented binaurally through Sennheisser HD-558 headphones (Sennheiser, Hanover, Germany). Prior to testing, listeners were asked to transcribe ten sentences presented in quiet conditions to train them to identify the target speaker. During each experimental trial, one target sentence was presented in conjunction with the masker. The subjects were instructed to listen to sentences spoken by a female native English speaker in the presence of two-talker background speech. In addition, they were informed that the background speech would be perceived from various azimuth locations. They were asked to type what they heard from the target speaker and report individual words if they were unable to identify the whole sentence. They were only allowed to listen to the sentence once. Each participant listened to 20 sentences from each of the three spatial locations in two background languages resulting in a total of 120 sentences. The presentation order of these sentences were completely randomized for each subject. The SNR of the target-background combinations, which determines the difficulty of listening conditions, was manipulated between subjects. All other factors were manipulated within subjects.

3. Results

Data from all subjects (N = 20 in SNR = −5 dB and N = 22 in SNR = −10 dB conditions) were included in the final analyses. Figure 1 depicts intelligibility scores (percent key words, per the BKB database, correctly reported) by location and masker language. The left and right panels depict performance in the −5 and −10 dB SNR conditions, respectively. As evident from Fig. 1, the general pattern of results was qualitatively similar across both SNRs. As expected, overall performance was lower in the more difficult, −10 dB SNR condition. Regardless of masker language, performance was better when the masker was spatially separated from the target rather than when co-located suggesting a clear SRM effect. Across spatial conditions, it also appears that listener performance was better when presented with the Dutch masker language than with English. This suggests that listeners reliably benefitted from LRM.

To evaluate these trends statistically, inferential analyses were performed. The percent correct scores were transformed into respective logit values in order to normalize the error variance. In four instances, proportion scores of 1 were replaced by 0.99 to avoid singularities in the transformed data. Data were submitted to a 2 (masker language) × 3 (location) factorial repeated measures analysis of variance (ANOVA) for each SNR (which was manipulated between subjects). Table 1 summarizes the results of this analysis. In both SNRs, the main effects of masker language ($p < 0.001$) and location ($p < 0.001$) were significantly different indicating that listeners experienced a

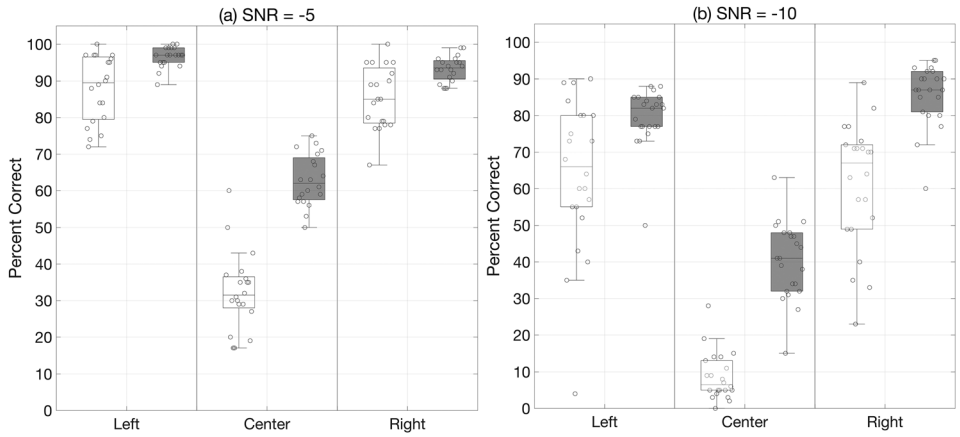

Fig. 1. Sentence recognition performance (percent correct) for 42 listeners in the presence of two-talker English (white) and two-talker Dutch (gray) maskers in the −5 dB (left) and −10 dB (right) SNR conditions. Each open circle represents the score of a single participant. Boxplots in the two masking conditions (English and Dutch) are plotted for each spatial location condition (left, center, right). The boxes depict the values between the 25th and 75th percentiles and the whiskers denote the ±1.5 interquartile range. Medians are shown as horizontal lines.

clear advantage due to both LRM and SRM. There was also a significant interaction between masker language and location ($p < 0.01$) suggesting that the amount of LRM depended on spatial location in both SNRs.

To further tease apart these effects, we explored two specific questions. First, does LRM persist when the speech sources are spatially separated? To answer this question, we performed a planned analysis of simple main effects for both SNRs. Again, in each SNR, the LRM remained reliable for both spatially separated conditions (Bonferroni corrected $p < 0.001$) confirming the persistence of LRM effects under spatial separation. Second, we asked how the relationship between LRM and spatial separation changes between the two SNR conditions. To answer this, for each subject, and for each masker language, we calculated an average intelligibility score in the spatially separated conditions by averaging the intelligibility scores in the −90° and +90° location. Following this, we calculated LRM scores by subtracting the intelligibility in English from the Dutch conditions in both spatial conditions for each subject. We submitted these LRM scores to a 2 (spatial location: co-located and separated) × 2 (SNR: −5 dB and −10 dB) omnibus mixed ANOVA. Figure 2 depicts the average LRM scores observed for the co-located and the separated spatial locations for −5 dB SNR (left panel) and −10 dB SNR (right panel). We found a reliable effect for location [$F(1, 39) = 89.88$, $p < 0.001$, $\eta_p^2 = 0.70$] confirming that the LRM was greater in the co-located condition. Finally, the significant interaction between location and SNR [$F(1, 39) = 14.92$, $p < 0.001$, $\eta_p^2 = 0.28$] confirms that in more difficult listening conditions (Fig. 2, right panel), LRM benefits increase especially under conditions of spatial separation. The effect for SNR [$F(1, 39) = 4.25$, $p = 0.046$, $\eta_p^2 = 0.01$] was significant indicating that in the harder SNR there is slightly more LRM benefit.

4. Discussion

Previous studies that have reported a benefit for speech-in-speech recognition tasks with language backgrounds have used target-masker pairs that were spatially co-located. In typical listening situations, listeners often derive a substantial benefit due to spatial separation in addition to linguistic differences between the sources. In this study, we investigated whether LRM effects persist even under conditions in which the target and masker speech are spatially separated. We did so by manipulating both the

Table 1. Summary of analyses of variance results for both SNR conditions. All compared effects were significant at the $p < 0.001$ level.

Source	SNR = −5 dB	SNR = −10 dB
Language	$F(1, 19) = 77.38$, $p < 0.001$, $\eta_p^2 = 0.80$	$F(1, 21) = 203.27$, $p < 0.001$, $\eta_p^2 = 0.91$
Location	$F(2, 38) = 274.50$, $p < 0.001$, $\eta_p^2 = 0.94$	$F(2, 42) = 369.40$, $p < 0.001$, $\eta_p^2 = 0.95$
Language × Location	$F(2, 38) = 5.29$, $p = 0.009$, $\eta_p^2 = 0.22$	$F(2,42) = 36.04$, $p < 0.001$, $\eta_p^2 = 0.63$

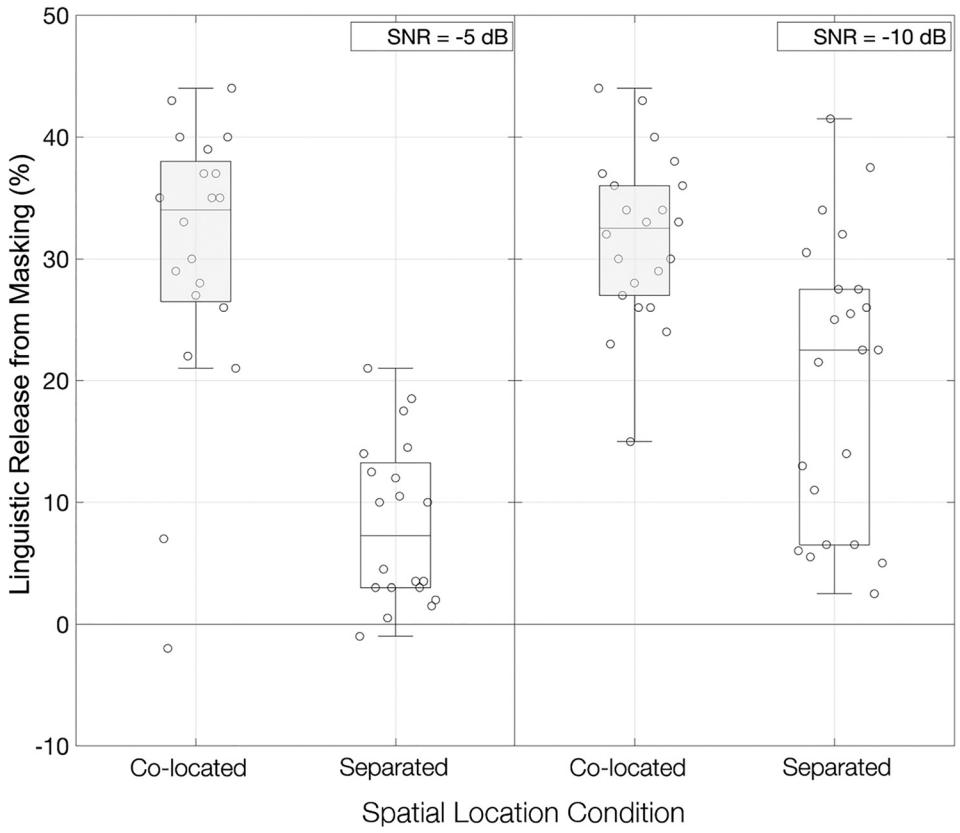

Fig. 2. The individual benefit due to linguistic release from masking plotted in percentage points for the 42 listeners tested in different spatial locations (co-located vs separated) in the −5 dB (left) and −10 dB (right) SNR conditions. Each open circle represents the benefit estimated for every individual. The boxes depict the values between the 25th and 75th percentiles and the whiskers denote the ±1.5 interquartile range. Medians are shown as horizontal lines.

language and the spatial location of the background masker relative to the target in two SNR conditions. Our results, across both SNR conditions, indicated that listeners demonstrated a reliable increase in performance when the locations of the target and background speech differed—a clear indication of SRM. Similarly, the listeners' performance improved when the language of the target and the background was mismatched—a clear indication of LRM. Critically, the benefit due to LRM was reliable across all different spatial locations tested, confirming that this benefit will occur under typical listening conditions. Finally, follow-up analyses indicated that the size of this effect is significantly reduced when the sources are spatially displaced. We offer a possible explanation for this outcome. Co-located presentation produces more challenging listening conditions, under which the segregability of the language masker is especially beneficial. This advantage is less useful under spatially separated conditions that already provide listeners with additional information for segregation. A corollary suggestion is that, in general, spatially displacing the competitors results in a substantially improved intelligibility of the target in the English masker conditions (see Fig. 2). This benefit, due to SRM, limits the overall advantage that listeners can experience due to LRM in less challenging SNRs. Taken together the claim that LRM is modulated by the overall difficulty of the listening conditions is supported by our finding of more LRM under spatial separation in the harder SNR (−10 dB) condition (also see Van Engen and Bradlow, 2007). To conclude, our results clearly demonstrate that listeners are better able to disregard background speech in a foreign language even when it is spatially separated from the target.

Acknowledgment

This research was supported by NSF grant BCS-1431105 to N.V.

References and links

Bench, J., Kowal, Å., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially–hearing children," Br. J. Audiol. **13**(3), 108–112.

Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (**2012**). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," J. Acoust. Soc. Am. **131**(2), 1449–1464.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., and Bradlow, A. R. (**2013**). "Masking release due to linguistic and phonetic dissimilarity between the target and masker speech," Am. J. Audiol. **22**(1), 157–164.

Calandruccio, L., Dhar, S., and Bradlow, A. R. (**2010**). "Speech-on-speech masking with variable access to the linguistic content of the masker speech," J. Acoust. Soc. Am. **128**(2), 860–869.

Calandruccio, L., Leibold, L. J., and Buss, E. (**2016**). "Linguistic masking release in school–age children and adults," Am. J. Audiol. **25**(1), 34–40.

Cooke, M., Lecumberri, M. G., and Barker, J. (**2008**). "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," J. Acoust. Soc. Am. **123**(1), 414–427.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**(5), 2112–2122.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. (**2009**). "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," EURASIP J. Adv. Signal Processing **2009**(1), 298605.

Mattys, S. L., Brooks, J., and Cooke, M. (**2009**). "Recognizing speech under a processing load: Dissociating energetic from informational factors," Cognit. Psychol. **59**(3), 203–243.

Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multitalker background noise)," J. Acoust. Soc. Am. **121**(1), 519–526.