Contextual variability during speech-in-speech recognition

Susanne Brouwer[a] **and Ann R. Bradlow**
Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, Illinois 60208
s.m.brouwer@uu.nl, abradlow@northwestern.edu

Abstract: This study examined the influence of background language variation on speech recognition. English listeners performed an English sentence recognition task in either "pure" background conditions in which all trials had either English or Dutch background babble or in mixed background conditions in which the background language varied across trials (i.e., a mix of English and Dutch or one of these background languages mixed with quiet trials). This design allowed the authors to compare performance on identical trials across pure and mixed conditions. The data reveal that speech-in-speech recognition is sensitive to contextual variation in terms of the target-background language (mis)match depending on the relative ease/difficulty of the test trials in relation to the surrounding trials.

PACS numbers: 43.71.Es, 43.71.Sy, 43.72.Dv [SGS]
Date Received: February 22, 2014 **Date Accepted:** May 19, 2014

1. Introduction

Speech communication in everyday life often takes place in multi-talker environments in which speech may be highly degraded relative to optimal communicative settings. An important characteristic of multi-talker situations is that they can be highly unpredictable. For example, at an international congress, background speech can vary in the number, the gender, and the language of the talkers. The aim of the current study is to examine the influence of background language variation on speech-in-speech recognition.

Previous work has only shown small or no effects of cross-trial masker variation for speech-in-speech recognition in which talkers, semantic content, or location of the background speech signal was held constant or varied across trials (Brungart and Simpson, 2004; Freyman *et al.*, 2007; Jones and Litovsky, 2008), suggesting that speech-in-speech recognition accuracy may be dominated by within-trial target and masker characteristics (i.e., energetic factors) with a severely limited role for cross-trial, contextual (i.e., informational) factors. These studies all used multi-talker background babble that matched the language spoken in the target (e.g., English-in-English). In contrast, several studies have demonstrated a recognition advantage when the target and background babble are not spoken in the same language (e.g., Garcia Lecumberri and Cooke, 2006; Van Engen and Bradlow, 2007). This work provides accumulating evidence that target-background language similarity, i.e., a match vs mismatch between target and background language, plays a significant role in speech-from-speech segregation (e.g., Brouwer *et al.*, 2012). However, since matched target and masker languages are likely to share more short- and long-term acoustic properties than mismatched target and masker languages, manipulation of the target-background language relationship likely involves changes in both energetic and informational masking properties, and therefore the target-background language mismatch benefit cannot be solely attributed to linguistic/informational rather than to purely energetic factors.

[a]Author to whom correspondence should be addressed.

In the present study we investigate whether variation in the target-background language relationship across trials within a test session (i.e., the introduction of informational masking due to masker uncertainty), influences speech-in-speech recognition for a consistent set of test trials (i.e., under controlled energetic masking conditions). Whereas previous work focused on contextual variation (i.e., masker uncertainty) in terms of talkers, semantic content or location of the background speech signal (Brungart and Simpson, 2004; Freyman *et al.*, 2007; Jones and Litovsky, 2008), we focus here on background language variation, a dimension of variation that is increasingly common in an era of globalization and that, as mentioned above, previous work has suggested (but not yet indisputably verified) may introduce a source of informational, in addition to energetic masking. Moreover, in an effort to more closely match real-world communicative settings, we use meaningful sentence stimuli with an open set response format, rather than formulaic sentences with a closed-set response format or nonsense sentences. Finally, rather than examining the effect of contextual variability (masker uncertainty) on overall performance across entire test sessions with varying degrees of masker uncertainty, we examine the influence of background language variation on a consistent subset of trials (test trials) when presented in the context of other trials (surrounding trials) in the same session that involve either more or less masking than the test trials. In so doing, we can gain a more precise picture of the effect of masker uncertainty on speech-in-speech recognition.

How might masker uncertainty influence performance? On the one hand, listeners' performance on a speech-in-speech recognition task might decrease whenever the language of the background speech changes unpredictably from trial to trial, regardless of whether the test trials are subject to more or less masking (due to matched versus mismatched target and background languages) than the surrounding trials. This overall increase in informational masking due to masker uncertainty may result from interruption of adaptation to the masker (i.e., a decrease in learning to tune out a consistent masker). On the other hand, while masker uncertainty may prevent learning to tune out the masker, a mixed condition that includes both matched and mismatched language trials (i.e., maskers of varying effectiveness) may also influence the process of adaptation to the target (i.e., learning to tune into the target). That is, any cost of masker uncertainty for test trials with highly effective maskers (relatively difficult, matched language trials) may be offset by a benefit of target familiarity (i.e., familiarity with the target talker and language) gained from surrounding trials with less effective maskers (relatively easy, mismatched language trials). In addition, for relatively easy test trials with mismatched language maskers, the cost of masker uncertainty may "conspire" with the reduced target recognition accuracy of more difficult matched language surrounding trials. In this case, the cost of masker uncertainty for adaptation to the masker (tuning out) may be added to the cost of greater masking in the surrounding trials for adaptation to the target (tuning in).

Experiment 1 compared English listeners' performance on a speech-in-speech recognition task with English targets in three different contextual conditions: Two different "pure" background language conditions (i.e., either 2-talker English or 2-talker Dutch babble for all trials) and one mixed background condition (i.e., a "mix" of trials with 2-talker English babble interleaved with trials with 2-talker Dutch babble). This setup allowed us to compare performance on English-in-English trials in the pure English-in-English condition versus the same trials (i.e., the same target sentences in the same background in the same serial positions within the experimental run) of the mixed condition (English-in-English mixed with English-in-Dutch). The crucial difference between the two conditions in each comparison is the surrounding trials which were matched language English-in-English trials for the pure condition and mismatched language English-in-Dutch trials for the mixed condition. Similarly, we compared recognition of English-in-Dutch trials in the pure English-in-Dutch condition versus in the mixed condition (English-in-Dutch mixed with English-in-English). These comparisons can thus give insight into contextual influences (i.e., the context surrounding the critical

test trials in one experimental run) on speech-in-speech recognition, taking into account the relative degrees of language masking of the test and surrounding trials. We selected English and Dutch as the languages for this study because our previous work has shown the target-background language mismatch benefit with this language pair, and the current study is part of this series of studies that have used this pair (Brouwer *et al.*, 2012; Calandruccio *et al.*, 2013). Furthermore, our previous work has shown a smaller masking release when a masker language is more linguistically close to the target speech than when it is distant (Calandruccio *et al.*, 2013). In the current study, we created a situation of potentially high confusion between target and masker signals by choosing Dutch as the mismatched language masker as this language is relatively linguistically close to English (the language of the targets).

2. Experiment 1

2.1 Method

Forty-eight native American-English listeners (27 females, age range 18 yrs to 27 yrs and 3 months) were tested. In a questionnaire they reported not having any hearing or speech impairments. Sixteen listeners participated in each condition, making this a between-subjects design.

Three native female American-English talkers and two native female Dutch talkers produced the target (English only, one talker) and babble stimuli (both English and Dutch, two talkers for each language). Eight lists of English sentences were selected from the revised Bamford-Kowal-Bench test (Bamford and Wilson, 1979) as targets. Each list contains 16 meaningful sentences with 3 or 4 keywords for a total of 50 keywords per list. For the background babble, 200 English meaningful sentences were taken from the Harvard/IEEE sentence lists (IEEE, 1969). These sentences were translated into Dutch for the Dutch babble. Two distinct 2-talker background babble tracks were created from these sentence recordings. From each of the 4 babble talkers' recordings (2 English and 2 Dutch talkers), 100 of the 200 sentences were pseudorandomly selected, resulting in 4 different 1-talker tracks (2 in English and 2 in Dutch). We chose to select only 100 sentences because we were not in need of using all the 200 sentences. The sentences were part of the Harvard/IEEE lists so we assumed that there were no big differences between the complexities of the sentences. Two-talker babble tracks were then created by mixing the talkers of the same language into one single audio file in Audacity©. Both tracks were equalized to the same root mean square level and the long term average speech spectra of the two tracks were normalized as a means of reducing unequal amounts of energetic masking between conditions. The English target sentences (by the third native English talker who was not one of the English babble talkers) were mixed online with the appropriate 2-talker background speech track for a given condition using Max/MSP©. On each trial, a random portion of the desired babble track was selected. The babble came on 500 ms before and continued for 500 ms after the target sentence. The level of the target sentences was fixed at 65 dB sound pressure level (SPL). Babble tracks were played at 68 dB SPL to produce a target-to-babble ratio of −3 dB. The combined speech and babble tracks were played out diotically over Behringer Pro XL headphones which were noise-cancelling.

In a sound-attenuating chamber, listeners were instructed to listen to English sentences spoken by a native American-English female speaker in the presence of background speech (babble). They were asked to repeat what they heard orally. Responses were digitally recorded and scored offline by a native American-English speaking experimenter. After a short practice session of eight trials [signal-to-noise ratio (SNR) of +5 and 0 dB, 4 trials each], the purpose of which was to familiarize the participants with the target talker's voice and with the speech-in-speech recognition task, participants were presented with a total of 128 experimental items. These sentences were presented in a fixed order in all conditions such that we could compare performance on the exact same test trials across the pure and the mixed conditions. Note that this

comparison therefore involves only approximately half of the trials of each condition. The test trials were presented in either (1) a consistent background of English babble (pure English condition), (2) a consistent background of Dutch babble (pure Dutch condition), or (3) in a mixed background with some trials involving Dutch and others involving English as the background language (mixed language condition). In the English and Dutch mixed condition (mixed language condition), 66 sentences were presented in English background babble, and 62 sentences were presented in Dutch background babble. The background language switched 30% of the time over one experimental run. Each test session was about 25 min.

Data were analyzed using a linear mixed-effects regression model (Baayen *et al.*, 2008) with keyword identification accuracy as the dichotomous dependent variable. A logistic linking function was used to deal with the categorical nature of the dependent variable. To address the issue of contextual variability during speech-in-speech recognition, we constructed a 2×2 model of recognition accuracy with condition as one contrast-coded fixed effect (pure vs mixed) and background language as the other (Dutch vs English), and the condition by language background interaction. Random intercepts were included for participants and items, along with a random slope for condition by items. Significance was assessed via likelihood ratio tests comparing the full model to a model lacking only the fixed effect (Barr *et al.*, 2013). In this model, a main effect of background language would be evidence for a replication of the mismatched language benefit (Brouwer *et al.*, 2012), and a main effect of condition would be evidence for an influence of contextual variability (masker uncertainty).

2.2 Results and discussion

Figure 1 (left panel) shows recognition accuracy scores for both English-in-English and English-in-Dutch test trials across the pure and the mixed condition. The dotted line connects performance on the subset of English-in-English trials ($n = 66$) of the pure English condition ($M = 61\%$) with the identical trials of the mixed language condition ($M = 65\%$). The solid line connects the subset of English-in-Dutch trials ($n = 62$) of the pure Dutch condition ($M = 84\%$) with the identical trials of the mixed language condition ($M = 75\%$). The analysis showed a main effect of background language [$\beta = 0.88$, standard error (s.e.) $= 0.22$, $\chi^2(1) = 14.96$, $p < 0.001$]. This indicates that our results replicate the finding that a mismatch between the background and target speech language facilitates recognition (Brouwer *et al.*, 2012). While there was no main effect of condition [$\beta = 0.14$, s.e. $= 0.18$, $\chi^2(1) < 1$, $p > 0.1$], the analysis revealed a background language by condition interaction [$\beta = 0.77$, s.e. $= 0.36$, $\chi^2(1) = 4.35$, $p < 0.05$]. Follow-up regressions revealed that this interaction reflected a significant effect of condition for the English-in-Dutch test trials [$\beta = 0.53$, s.e. $= 0.22$, $\chi^2(1) = 5.43$, $p < 0.05$], but not for the English-in-English test trials [$\chi^2(1) < 1$].

The results of Experiment 1 are partly consistent with the idea that speech-in-speech recognition decreases under conditions of contextual variability. The

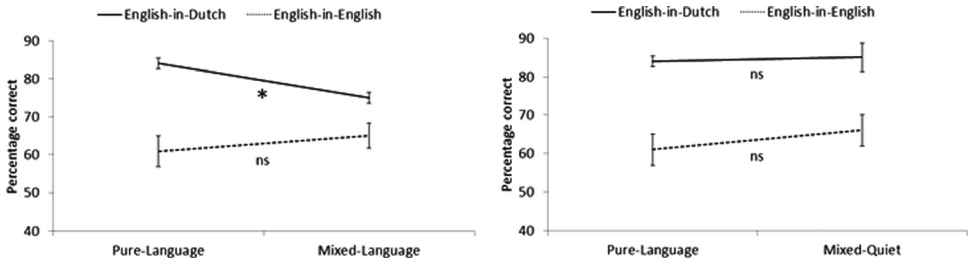

Fig. 1. Mean intelligibility scores in percentage correct keyword identifications for each condition in Experiment 1 (left) and Experiment 2 (right). Solid and dotted lines represent English-in-Dutch and English-in-English test trials, respectively. Error bars represent standard error.

recognition of English-in-Dutch test trials was lower when the surrounding trials were English-in-English (mixed condition) than when the surrounding trials were English-in-Dutch (pure Dutch condition). However, listeners performed equally well (or poorly) on the English-in-English test trials, irrespective of variation across trials in the background language. This inconsistent effect of contextual variability suggests that an effect of masker uncertainty may interact with other processes of target and/or background adaptation (i.e., tuning in and tuning out, respectively). In particular, variable versus stable recognition accuracy across pure and mixed blocks depends on the relative masking effectiveness of the test and surrounding trials. More specifically, the surrounding context in which test trials are presented matters (i.e., for English-in-Dutch trials) but this effect is also dependent on the type of test trials (i.e., no context effect for English-in-English test trials).

To investigate this more closely, we test in Experiment 2 whether the "spread" of masking from more difficult surrounding trials to easier test trials (i.e., the interruption of any adaptation, or tuning in to the target) can be reversed in a mixed block with easier surrounding trials. Specifically, we test recognition accuracy in a block that mixed English-in-Dutch trials with English-in-quiet trials (i.e., with no background babble). Thus, this mixed quiet condition has the same degree of masker uncertainty as the mixed-language condition of Experiment 1, but the critical comparison is across English-in-Dutch test trials surrounded by trials with either the same (pure block) or less (rather than more) masking. In view of the lack of effect of contextual consistency versus variability that we observed in Experiment 1 for the relatively difficult English-in-English test trials, we suspected that the predominant influence on recognition accuracy for these trials is within-trial, energetic masking factors. Nevertheless, Experiment 2 also tested whether extremely easy surrounding trials (i.e., trials with no background babble) can raise performance for the English-in-English trials.

3. Experiment 2

3.1 Method

Thirty-two American-English listeners (23 females, age range 18 yrs and 5 months to 22 yrs and 3 months) were tested. They reported not having any hearing or speech impairments in a questionnaire. Sixteen listeners participated in each condition, making this a between-subjects design.

The same materials were used as in Experiment 1, except that the English-in-Dutch surrounding trials of the mixed condition in Experiment 1 were replaced with English-in-quiet surrounding trials in the English-quiet condition, resulting in 66 English-in-English test trials and 62 English-in-quiet surrounding trials. For the Dutch-quiet condition, the English-in-English test trials of the English-quiet condition were replaced with English-in-Dutch test trials, resulting in 66 English-in-Dutch and 62 English-in-quiet trials. The procedure was almost identical to the one used in Experiment 1, except that listeners were orally instructed to listen to English sentences spoken by a native English female speaker in the presence or absence of background speech. Of the eight practice trials, half were presented with background speech (SNR of +5 and 0 dB, two trials each), and half were presented without background speech.

The analysis was similar to Experiment 1: A 2×2 model of recognition accuracy was constructed with condition as one contrast-coded fixed effect (pure vs quiet) and background language as the other (Dutch vs English), and the condition by language background interaction. Random intercepts were included for participants and items, along with a random slope for condition by items.

3.2 Results

Figure 1 (right panel) shows recognition accuracy scores for both English-in-English and English-in-Dutch test trials across the pure and the quiet condition. The dotted

line connects performance on the subset of English-in-English trials ($n = 66$) of the pure English condition ($M = 61\%$, Exp. 1) with identical trials of the mixed-quiet condition ($M = 66\%$). The solid line connects the subset of English-in-Dutch trials ($n = 66$) of the pure Dutch condition ($M = 84\%$) with identical trials of the mixed-quiet condition ($M = 85\%$). The analysis showed a main effect of background language [$\beta = 1.36$, s.e. $= 0.20$, $\chi^2(1) = 3.80$, $p < 0.0001$] and no other effects ($ps > 0.1$).

The findings of Experiment 2 established that the detrimental effect of masker uncertainty that we observed in the mixed-language condition of Experiment 1 for the English-in-Dutch test trials was reversed in the Dutch-quiet condition of Experiment 2. Thus, in Experiment 2 we observed equivalent recognition accuracy for the English-in-Dutch test trials in the pure and mixed condition, despite the fact that the mixed condition in Experiment 2 had the same degree of trial-to-trial masker uncertainty as the mixed condition of Experiment 1 where we observed a decline in recognition accuracy for the English-in-Dutch test trials in the mixed versus in the pure conditions. As for Experiment 1, recognition of the relatively difficult English-in-English test trials remained stable across the pure and mixed conditions.

4. General Discussion

The present study examined how listeners deal with variation in the target-background language relationship at the contextual level in a speech-in-speech recognition task. Crucially, the approach in this study involved comparisons of identical trials across conditions, thereby offering comparisons with controlled energetic masking characteristics. Moreover, we examined the influence of masker uncertainty separately for relatively easy (English-in-Dutch) and relatively difficult (English-in-English) trials.

The results of this study replicated the mismatched language benefit (Brouwer *et al.*, 2012) for English targets with English and Dutch maskers, and demonstrated further that background language variation at the contextual level influenced speech-in-speech recognition only in the particular situation of relatively easy, mismatched language English-in-Dutch test trials surrounded by relatively difficult, matched language English-in-English trials. Specifically, Experiment 1 showed that English-in-Dutch recognition declined when the test trials were presented in a mixed condition that included surrounding trials with greater masking (English-in-English) than the test trials (English-in-Dutch). Experiment 2 then showed that this decline in recognition accuracy could be reversed by mixing the relatively easy English-in-Dutch test trials with surrounding trials that had no masking at all (English-in-quiet test trials). In contrast, relatively difficult matched language English-in-English recognition remained stable across pure and mixed blocks even though, in both experiments, the surrounding trials had less effective maskers than the test trials. We suggest that this stability in recognition accuracy may be due to the predominant influence of energetic masking factors that could not be overcome by reduced masking on surrounding trials. It remains for future research to determine whether the masking in English-in-English trials can be modulated by contextual factors (e.g., if there is a SNR at which contextual factors can exert an influence in a parallel way to what we have observed for the mismatched language English-in-Dutch trials).

The exact mechanism by which contextual variation exerts its influence on speech-in-speech recognition has yet to be specified. It is, however, clear from the present research that speech-in-speech recognition accuracy can display sensitivity to contextual variation as it extends beyond the time-frame of the target signal itself. In contrast to previous work that showed little or no effect of cross-trial variation in acoustic or spatial characteristics known to influence speech-in-speech recognition (e.g., Brungart and Simpson, 2004; Freyman *et al.*, 2007; Jones and Litovsky, 2008), this study has shown sensitivity to contextual variation in terms of the matching or mismatching of the language being spoken in the target and in the background. This sensitivity extended to languages as similar in sound structure as English and Dutch. Importantly, while previous studies examined the influence of masker uncertainty on

overall speech-in-speech recognition across a test session, the present study demonstrated that the influence of masker uncertainty is constrained by the relative ease or difficulty of the test trials in relation to the surrounding trials. While masking appeared to spread from relatively difficult surrounding trials to relatively easy test trials, relatively difficult test trials remained resistant to a release from masking when surrounded by relatively easy trials. Thus, this study has demonstrated that the cost of masker uncertainty for background adaptation (tuning out) can be added to the cost of greater surrounding trial masking for target adaptation (tuning in). It remains for future work to determine if there are conditions in which the cost of masker uncertainty for background adaptation (tuning out) may be outweighed by a benefit of less surrounding trial masking for target adaptation (tuning in).

Acknowledgments

This work was supported by Grant No. R01-DC005794 from NIH-NIDCD. We thank Chun Liang Chan, Vanessa Dopker, and Lindsay Valentino for technical and research assistance.

References and links

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**, 390–412.

Bamford, J., and Wilson, I. (**1979**). "Methodological considerations and practical aspects of the BKB sentence lists," in *Speech-hearing Tests and the Spoken Language of Hearing-impaired Children*, edited by J. Bench and J. Bamford (Academic, London), pp. 148–187.

Barr, D., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**(3), 255–278.

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.

Brungart, D. S., and Simpson, B. D. (**2004**). "Within-ear ad across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," J. Acoust. Soc. Am. **115**(1), 301–310.

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

Freyman, R. L., Helfer, K. S., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**(2), 1040–1046.

Garcia Lecumberri, M. L., and Cooke, M. (**2006**). "Effect of masker type on native and non-native consonant recognition in noise," J. Acoust. Soc. Am. **119**(4), 2445–2454.

IEEE Subcommittee on Subjective Measurements. IEEE Recommended Practices for Speech Quality Measurements (**1969**). IEEE Trans. Audio Electroacoust. **17**, 227–246.

Jones, G. L., and Litovsky, R. Y. (**2008**). "Role of masker predictability in the cocktail party problem," J. Acoust. Soc. Am. **124**(6), 3818–3830.

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.