Error patterns of native and non-native listeners’ perception of speech in noise

Benjamin D. Zinszer[1,2], Meredith Riggs[2], Rachel Reetzke[2,3], & Bharath Chandrasekaran[2,4]

1. Department of Linguistics and Cognitive Science, University of Delaware, Newark, DE, USA
2. Department of Communication Sciences and Disorders, University of Texas at Austin, Austin, TX, USA
3. Department of Psychiatry and Behavioral Sciences, UC Davis MIND Institute, Sacramento, CA, USA
4. Department of Communication Science and Disorders & Center for Neural Bases of Cognition, University of Pittsburgh, Pittsburgh, PA, USA

bzinszer@gmail.com*, mfriggs@gmail.com, rdreetzke@ucdavis.edu, b.chandra@pitt.edu
* Corresponding author

Running head: Word and morpheme speech perception errors

Abstract

Speech perception in noise requires both bottom-up sampling of the stimulus and top-down reconstruction of the masked signal from a language model. Previous studies have provided mixed evidence about the exact role that linguistic knowledge plays in native and non-native listeners' perception of masked speech. This paper describes an analysis of whole utterance, content word, and morphosyntactic error patterns to test the prediction that non-native listeners are uniquely affected by energetic and informational masks because of limited information at multiple linguistic levels. The results reveal a consistent disadvantage for non-native listeners at all three levels in challenging listening environments.

Keywords: speech processing, speech-in-noise, bilingualism, morphosyntax

1. Introduction

The cognitive demands of speech perception are often amplified in natural listening conditions where persistent speech and non-speech noise causes informational and energetic masking of the speech signal (Mattys et al., 2012). By all accounts, the demands of speech perception in noisy conditions (hereafter, SPIN) are even greater for non-native listeners, who have less experience with the language they are hearing and interference from knowledge of their first language (Meador, Flege, & MacKay, 2000). Across studies, non-native listeners reliably perform worse than native listeners in perceiving most types of masked speech (see Garcia Lecumberri, Cooke, & Cutler, 2010 for review). Both informational masks and energetic masks adversely affect non-native listeners' comprehension more than native listeners overall (Cooke, 2008), but the exact mechanism underlying this disadvantage are not entirely clear. In this study, we expand upon an analytic approach described by Smith and Fogerty (2017) to classify and compare different types of errors committed by native and non-native listeners under different masking conditions.

Evidence from both native and non-native listeners highlights the importance of synthesizing top-down linguistic knowledge with bottom-up acoustic information for SPIN (Rönnberg et al., 2013). Glimpsing is one proposed mechanism by which listeners integrate a degraded bottom-up signal with a top-down language model. When listeners perceive fragments of interrupted speech, they actively reconstruct the surrounding context based on their linguistically-informed predictions (Cooke, 2006). Listeners with less language knowledge will have less such top-down information to apply to the inference, and thus, an important contribution to the non-native listener disadvantage in SPIN is likely incomplete knowledge of the target language (Garcia Lecumberri et al., 2010). This view is supported by the elimination of native vs. non-native differences in energetic masking when linguistic cues to the stimulus (e.g., syntactic, semantic, or phonotactic context) are not available in the signal (Garcia Lecumberri et al., 2010).

Nevertheless, non-native listeners also appear to apply top-down knowledge in speech perception to overcome an energetic mask, though to less advantageous ends than native-listeners (Bradlow & Alexander, 2007). This application of language knowledge also presents a second obstacle: non-native listeners' knowledge of their native language (not presented in the stimulus) can lead to interference in speech perception (Golestani, Rosen, & Scott, 2009). Conflicting cues from a listener's native language and the (non-native) target language may thus be a source of disadvantage for the non-native listener. Cooke and colleagues (2008) reported that while the overall native listener advantage is preserved for both steady-state noise and multi-talker masks, non-native learners were more adversely affected by a talker mask than could be accounted for by its energetic contribution alone, indicating a greater sensitivity to interfering information from the talker mask for non-native listeners relative to native listeners.

The general pattern for non-native listeners to perform worse than native listeners on SPIN tasks may belie the nuanced interaction between different sources of information that listeners apply or selectively inhibit in the reconstruction of linguistically-relevant information from glimpses of the target. Smith and Fogerty (2017) provided further evidence for this interaction in native speakers with a microscopic analysis of error patterns in SPIN to discriminate between phonemic part-word errors and whole-word errors in sentence perception as a function of glimpse size. This study identified two important patterns in SPIN errors: (1) As glimpse sizes diminished (i.e., energetic mask coverage increased), word omission and—to a lesser extent—word substitution errors increased, while word additions remained rare at all masking levels, confirming the tendency of listeners to reconstruct words from perceived fragments. (2) At higher energetic mask coverage (67-100%), native listeners were nearly twice as likely to substitute whole, syntactically- and semantically-plausible words instead of substituting less relevant--but phonologically similar--words that would match the fragments of input. This finding highlights the importance of higher level lexical or syntactic predictions in SPIN, even as the expense of lower-level phonemic similarity. Theirs is the first study to analyze these "microscopic" error patterns in sentences by contrasting whole-word from phoneme-level errors.

In the present study, we compare listeners' error patterns at a level of linguistic knowledge not previously examined in SPIN experiments. This study examines the error rates in whole utterances, in content words, and in morphosyntactic affixes and closed-class words for native and late-L2-onset, non-native listeners. Like the sound-based substitutions in Smith and Fogerty's study, many morphosyntactic errors result in partial changes of the target word (e.g., substituting *cat* for *cats*) and to closed-class words (e.g., omissions of definite article *the*), but these errors may be mitigated by making morphosyntactic inferences from the context. While the lexical knowledge that appears to underlie the results in Smith and Fogerty's (2017) paradigm may be quickly attained by second language learners, masked morphosyntactic information could be only inferred by using broader sentential context, providing a stronger contrast of native- vs. non-native listeners. We ask how native and non-native listeners' error patterns differ across these levels and across four types of mask, with two specific predictions and a third open-ended question:

(1) We predicted that non-native listeners to English speech would be *more* susceptible to morphosyntactic errors than native listeners, as a result of their incomplete language knowledge. Mandarin Chinese has very different morphosyntactic rules from English, particularly in pluralization, tense markers, subject-verb agreement, and use of articles. (2) We predicted that informational masks (1-Talker and 2-Talker) would increase the magnitude of non-native disadvantage relative to the energetic masks (SSN and 8-Talker; see Cooke et al., 2008). (3) Finally, because these three types of errors draw on linguistic features with different base rates (number of content words, number of morphemic affixes, etc.), we made no prediction about the main effect of error type, but we asked whether the interaction between mask type and error type could differ between native and non-native listeners (i.e., a three-way interaction) indicating that word and morphosyntactic errors were differently sensitive to mask types.

2. Method

2.1 Participants

We acquired archived behavioral SPIN data from the pre-testing regimen in a recent electrophysiological study (Reetzke et al., 2017) for this study. The dataset included fifteen native- and fifteen non-native speakers of American English, recruited at the University of Texas at Austin. One non-native speaker was excluded from the present analysis because they did not complete the behavioral tasks. The native group was composed of speakers of American English who reported no significant experience learning or speaking another language. Non-native participants were sequential Mandarin-English bilinguals. All non-native participants were born and raised in mainland China, spoke Mandarin Chinese as their native language, did not begin learning English formally until after the age of 6 (range = 7-16 years, *mean* = 10.1 years, *s.d.* = 2.6 years), and lived in the United States no more than 6 years (range = 1 to 6 years, *mean* = 2. 0 years, *s.d.* = 1.6 years).

The native and non-native groups were comparable on age, sex, and non-verbal intelligence. See Table 1 for demographic details. Participants reported no previous history or diagnosis of speech, language, or neurodevelopmental disorders. All participants had normal hearing defined as air and bone conduction thresholds < 20 dB HL at octave frequencies from 250 to 8,000 Hz, as measured by an Interacoustics Equinox 2.0 PC-Based Audiometer. Participants had either no history of formal music training or no significant music experience (<6 years), according to a music and language questionnaire (Li et al., 2014). The groups did not significantly differ on this measure (*t*(27)=1.19, *p*=0.24).

Table 1. Participant demographics, group means and (standard deviations).

Group	N	Age	Sex	KBIT Score	Music Experience	TOAL-4
Native	15	22.5 (3.7)	9 F / 6 M	118 (8.4)	1.8 y (2.4)	107 (12)
Non-native	14	25.1 (3.4)	8 F / 6 M	124 (7.4)	0.9 y (1.8)	65 (10)

Participants also completed the Test of Adolescent and Adult Language (TOAL-4; Hammill, Brown, Larsen, & Wiederholt, 2007) as a standard measure of English language proficiency. The mean TOAL-4 composite for spoken English proficiency differed greatly between the native and non-native groups. Native speakers scored 107 (*s.d.*=12), and non-native speakers scored 65 (*s.d.*=10), *t*(27)=11.0, *p*<0.001.

2.2 Stimuli

The masked sentence stimuli were developed for previous speech in noise studies (for details, see: Chandrasekaran et al., 2015; Van Engen, 2012; Xie et al., 2015). Sixty-four target sentences from the Revised Bamford-Kowal-Bench Standard Sentence Test (Bamford & Wilson, 1979) were recorded by a female, native speaker of American English (Van Engen, 2012). Target sentence stimuli were organized into four masking conditions: sixteen 1-Talker (1T) informational mask trials, sixteen 2-Talker (2T) informational mask trials, twenty-four steady-state speech-shaped noise (SSN) energetic mask trials, and eight 8-Talker (8T) trials. Previous research has demonstrated that the 8T mask produces primarily energetic masking, at the same level as SSN (Brungart, Chang, Simpson, & Wang, 2009).

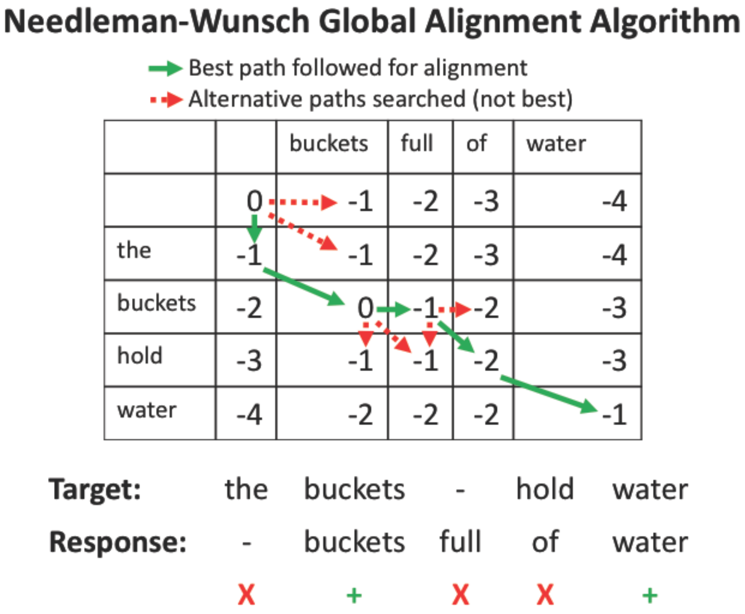
The SSN condition was composed from a steady-state white noise which was then shaped to a speech-like spectrum based on long-term average spectra acquired from 240 spoken sentences in the original corpus (Van Engen et al., 2010). Recordings of eight additional female, native speakers of American English reading a different set of sentences were used to generate the 1-, 2-, and 8-Talker masks (Van Engen et al., 2010; see Chandrasekaran et al., 2015 and Xie et al., 2015 for details). In all trials, the mask was 5 dB greater than the target sentences (SNR = -5 dB), consistent with the previous published use of these stimuli, wherein the SSN and 1T conditions elicited above-chance and below-ceiling performance (Chandrasekaran et al., 2015).

2.3 Procedure

Participants were seated in a SoundEgg sound-attenuated seat with a personal computer and Sennheiser HD280 headphones. They were instructed that they would be listening to several recorded sentences in different types of noise, and the target sentence would begin about half a second after the noise. Participants were asked to type the target sentence or their best guess into the computer for each trial. The computer volume was adjusted to a comfortable level, and participants listened to the 64 stimulus trials in a uniquely randomized order for each participant. Responses were scored according to the error analysis described below.

2.4 Error Analysis

In contrast with keyword counts often used for speech in noise tasks (e.g., Chandrasekaran et al., 2015; Reetzke et al., 2016; Smayda et al., 2016; Van Engen et al., 2012; Xie et al., 2014, 2015), this analysis examined whole utterance, content word, and morphosyntactic level errors. Participants' typed responses and their respective target sentences were aligned by a custom implementation of the Needleman-Wunsch algorithm, a dynamic global alignment algorithm primarily used for aligning protein and nucleotide sequences in bioinformatics (Needleman & Wunsch, 1970). In the original algorithm, three scores govern the optimal alignment: a match award, a mismatch penalty, and a gap penalty. The optimal alignment between two sequences is chosen by maximizing matches between the sequences' elements and minimizing mismatches and gaps. This search process is illustrated in Figure 1.


Needleman-Wunsch Global Alignment Algorithm
Best path followed for alignment
Alternative paths searched (not best)

		buckets	full	of	water
	0	-1	-2	-3	-4
the	-1	-1	-2	-3	-4
buckets	-2	0	-1	-2	-3
hold	-3	-1	-1	-2	-3
water	-4	-2	-2	-2	-1

Target:	the	buckets	-	hold	water
Response:	-	buckets	full	of	water
	X	+	X	X	+

Figure 1. Procedure for aligning target and response sentences. Each word pair matched by the aligner is compared for word-level or morphosyntactic errors.

Our implementation compares each word between two sentences (target and response) and adds one additional parameter to award pairs of words with Levenshtein distances ≤ 2. This adjustment encourages alignment for words with similar spellings such as some homonyms and rhymes when an exact match is not found. The alignment step produces two optimally-aligned sentences with equal numbers of words, allowing for gaps, which are represented by the token "_" (see Figure 1). Errors are then calculated by examining each word pairing in the alignment.

After alignment all words were lemmatized and tagged for part of speech by the Pattern module (De Smedt & Daelemans, 2012), a Python toolkit which quickly and accurately performs part-of-speech tagging, lemmatization, spelling suggestions, and other natural language processing (NLP) tasks in Python 2.7.10. If a pair of aligned words matched in their root forms but not in the original target and response, a morphosyntactic error was recorded. Additionally, if both words were function words or if a function word was aligned with a gap token, a morphosyntactic error was recorded. If both words were content words or if a content word was aligned with a gap token, a content word error was recorded. If one word in a pairing was a function word and the other a content word, one morphosyntactic error and one content word error were each recorded.

Finally, if none of the roots for content words from the key sentence appeared in the response, the entire trial was reclassified as "Did Not Hear" (DNH). This distinction is important because sometimes participants failed to transcribe any response at all or fully transcribed a masker sentence instead of the target. These whole utterance level errors are counted separately, as they do not reflect specific word-level or morphosyntactic changes in the perceived signal.

The code for performing this error analysis and the de-identified dataset are both publicly available at https://spin-scorcerer.github.io.

3. Results & Discussion

3.1 "Did Not Hear" Trials

The "Did Not Hear" (DNH) data were coded as 1 for a trial with a DNH error and 0 for all other trials, and they were fit to a logistic mixed-effects model (using the lme4 package for R; Bates et al., 2014) with a random effect of subject and fixed effects for non-native relative to native (reference level) listener groups and for each of the talker masks (1T, 2T, 8T) relative to SSN (reference level). We also estimated the interaction between group and each talker mask. The left panel of Figure 2 depicts the mean subject-level proportion of DNH trials in each group for each type of mask.

Effects of listener group and each of the talker masks in the model were statistically significant (p<0.001), as well as two of the three interaction terms: A negative interaction term between non-native and 1T (z=-2.95, p=0.003) indicated a smaller non-native disadvantage relative to the SSN condition. Likewise, in the 2T condition, a negative interaction with non-native (z=-4.98, p<0.001) also indicated a smaller non-native disadvantage in 2T as compared to SSN. A much smaller interaction (half magnitude relative to 2T) was observed for 8T (z=-2.05, p=0.04), providing weak evidence of a change in the non-native disadvantage between 8T and SSN conditions in the frequency of DNH responses.

Therefore, we find support for a non-native disadvantage at this whole-utterance level. Non-native listeners were significantly more likely to fail to transcribe any sentence at all under the purely energetic masking condition (SSN). This effect was attenuated for informational masks, 1T and 2T, suggesting that the non-native listeners benefitted from the extra glimpses of target speech more than they were adversely affected by distractors in the informational mask.

Beyond the weak interaction with non-native status, the fixed effect of 8T was significant (z=10.02, p <0.001), indicating a higher probability of producing DNH responses than SSN across the two groups. The fixed effects of non-native (vs. native) group, 1T (vs. SSN) and 2T (vs. SSN) were all larger than their respective interaction terms. Thus, while the non-native disadvantage was attenuated for 1T and 2T masks, the non-native listeners were still more likely to produce DNH trials overall, and all three talker masks resulted in more DNH trials across native and non-native listeners than SSN.

In contrast with the 1T and 2T findings, the weaker interaction for non-native disadvantage between the 8T and SSN suggests that 8T mask coverage was more similar to steady-state noise (consistent with Brungart et al., 2009). However across groups, the 8T mask yielded significantly more errors than SSN at the same signal to noise ratio, which indicates that some additional masking occurred in the 8T condition besides the purely energetic contribution, contrary to the findings of Freyman et al. (2004).

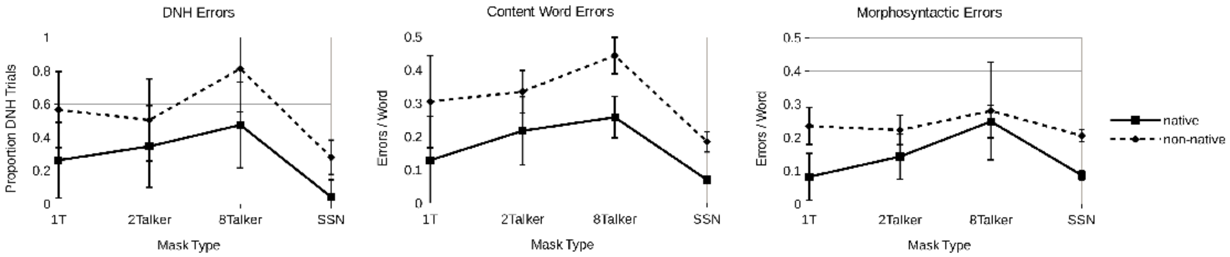


Figure 2. Mean error rates in native and non-native listeners. Left panel, the proportion of "did not hear" (DNH) responses per condition for both listener groups. Middle panel, content word-level error rates in native (solid line) and non-native (dashed line) listeners. Right panel, morphosyntactic errors for native (solid line) and non-native (dashed line) listeners. Error bars denote two standard errors of the mean.

3.2 Content Word and Morphosyntactic Errors

After excluding all DNH trials from the dataset, we fit a linear mixed-effects model (using the lmerTest package for R; Kuznetsova, Brockhoff, & Christensen, 2017) to the content word and morphosyntactic error rates per word in each trial, with a random effect of subject and fixed effects for non-native relative to native listener, morphosyntactic error relative to content word error, and each of the talker masks (1T, 2T, 8T) relative to SSN. We summarized the fixed effects in this model with a three-way ANOVA: 2 (group) by 2 (error type) by 4 (mask type).

The ANOVA found a marginally significant three-way interaction between group, mask, and error type ($F(3,2383)=2.39$, $p=0.07$). The two-way interactions between group and error type and between mask type and error type were both significant (group x error $F(1,2383)=4.79$, $p=0.03$; mask x error $F(1,2383)=10.82$, $p<0.001$). The two-way interaction between group and mask was not significant ($p=0.91$), although its variance may have been explained by the three-way interaction term. All main effects were statistically significant ($p<0.001$) but interpretable only in relation to their significant interaction terms, addressed in the follow-up analyses.

In the combined content words and morphosyntax errors, we found further confirmation of a non-native disadvantage in SPIN. However, our prediction that morphosyntax would be especially sensitive to non-native status due to additional linguistic information required to resolve morphosyntactic ambiguity (Meador, Flege, & MacKay, 2000) is only equivocally supported. A significant ($p=0.03$) two-way interaction between group (native vs. non-native) and error type, coupled with a marginally significant three-way interaction (group x error type x mask type) suggests that the non-native disadvantage might differ between the content word and morphosyntactic error measures. Bonferroni-corrected post-hoc *t* tests on the subject-level mean error rates (estimated first for each mask type, and then averaged across conditions to balance against the uneven number of trials per mask type) for native vs. non-native listeners revealed a significant non-native disadvantage in both content word errors (Welch's two-sample $t(27)=3.46$, $p_{Bonf}=0.004$, Cohen's $d=1.28$) and morphosyntactic errors (Welch's two-sample $t(25)=2.88$, $p_{Bonf}=0.016$, Cohen's $d=1.08$). However, while it appears that both error measures yield highly significant non-native disadvantages, the effect size in this dataset slightly favors content words over morphosyntax (although the two effect sizes did not statistically differ).

Because the differences between number of content word errors and number of morphosyntactic errors could be attributed to the base rate differences in the number of content words vs. the number of morphosyntactic affixes and function words, we selected this factor to partition the data over and perform two-way ANOVAs to further investigate the possible interactions. We re-estimated the linear mixed effects models separately for content word errors and for morphosyntactic errors, applying a 2 (group) by 3 (mask type) ANOVA for each model. For content word errors, the main effects of group ($F(1,35)=14.21$, $p<0.001$) and mask type ($F(3,1190)=35.07$, $p<0.001$) were statistically significant. No significant interaction was observed ($p=0.59$), see middle panel of Figure 2. The same was true for the morphosyntactic errors. Main effects of group ($F(1,38)=16.76$, $p<0.001$) and mask type ($F(3,1193)=9.28$, $p<0.001$) were statistically significant, while the interaction was not ($p=0.11$), see right panel of Figure 2. The main effects of group supported the non-native disadvantage in each error type.

We performed planned comparisons between SSN and the three talker mask types using Bonferroni corrected paired-*t* tests on subject-level averages. In the content word errors, 2T and 8T masks significantly exceeded SSN (both $p_{Bonf}<0.001$), but in the morphosyntactic errors, only the 8T significantly differed from SSN ($p_{Bonf}=0.008$). This finding aligns with the DNH results, wherein 8T produced significantly more errors than SSN. In both error conditions, listeners' 8T performance seems more closely linked to linguistic interference (i.e., number of talkers) than to mask energy (i.e., mask coverage, when signal-to-noise is held constant). However, the absence of significant group x mask type interactions in the two-way ANOVAs of either error type do not provide any support for Cooke et al.'s (2008) observation of increased non-native disadvantage under informational masking.

3.3 Balanced mask types

In a final follow-up analysis, we addressed the possible role of the unequal number of trials for different mask types. Previous research (see Brouwer & Bradlow, 2014; Freyman, Helfer, & Balakrishnan, 2007; Watson, 1987) demonstrated significant effects of uncertainty in performance on tone or speech perception, which may have affected the relative difficulty of each mask condition based on its frequency in a given task. To address this issue, we selected the first eight trials of each mask type for every subject, so that every mask type had the same degree of exposure and opportunities for participants to learn the noise properties. We then repeated all of the analyses described in the foregoing Results sections.

The results were nearly identical, with the following two exceptions: In the 2x2x4 (group x error type x mask type) ANOVA, the interaction of group x error type was no longer statistically significant (original: $F(1,2383)=4.79$, $p=0.03$; subset: $F(1,1013)=1.51$, $p=0.22$). The planned comparison of 1T and SSN content word errors significantly differed (original: $t(26)=1.94$, $p_{Bonf}=0.188$; subset: $t(25)=2.61$, $p_{Bonf}=0.046$). Thus, our follow-up analysis on the balanced subset of data did not change the interpretation of our main findings.

4. Summary & Conclusions

In this study, we examined three types of errors made by native and non-native listeners to spoken English sentences under energetic and informational masking. Based on previous research, we expected non-native listeners to perform differently from native listeners across these conditions as a result of their incomplete language knowledge, which limits the top-down resources that non-native listeners can draw upon to make inferences about the masked speech (Garcia Lecumberri et al., 2010). This prediction was supported by lower non-native listener performance in all three error types, and the present study identifies significant differences at all three levels separately. The evidence in this study tentatively suggested that content word errors increased more for non-native (relative to native) listeners than morphosyntactic errors, but effects at both levels were large and did not greatly differ in magnitude. In contrast to some previous studies of sentence (Cooke et al., 2008) and consonant (Garcia Lecumberri & Cooke, 2006) perception, we did not find evidence that the informational masks were disproportionately more difficult for non-native listeners relative to the non-native disadvantage in a steady-state noise energetic mask, although the 8T mask produced more errors on all three measures than a steady-state speech-shaped noise mask, suggesting that it provided both energetic and informational masking (contrary to Brungart et al., 2009; Freyman et al., 2004).

The present findings should be considered in the context of a few important limitations. The non-native listeners in this study were native speakers of one language (Mandarin Chinese), selected for its particular differences from English in morphosyntax. However, these non-native listeners were relatively diverse in their experience with English, likely adding variance to their error rates that was not accounted for in this brief investigation. Further, the number of trials per mask type (constrained by the availability of the archival data) were not balanced and mask types were randomly ordered, contributing to listeners' uncertainty about noise properties in any given trial. Our follow-up analysis did not find serious consequences of imbalance, and effects of mixing noise types were not observed for other-language masks but not same-language masks (Brouwer & Bradlow, 2014), likely limiting this effect to the SSN condition, if any. Further applications of our proposed error analyses to new participant groups and mask types would help to clarify these issues as well as potentially offering important new insights on non-native speech in noise perception. To that end, we have made the analysis code and sample data publicly available at https://spin-scorcerer.github.io to enable other researchers to implement this approach with their own data.

Acknowledgements

This work was supported by a grant from the National Institute on Deafness and other Communication Disorders of the National Institutes of Health to B.C. (R01 DC015504).

References

Andrillon, T., Kouider, S., Agus, T., & Pressnitzer, D. (2015). Perceptual learning of acoustic noise generates memory-evoked potentials. *Current Biology*, *25*(21), 2823-2829. https://doi.org/10.1016/j.cub.2015.09.027

Bamford, J., & Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. In J. Bench & J. Bamford (Eds.), *Speech-hearing tests and the spoken language of hearing-impaired children* (pp. 148-187). London: Academic Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339-2349. https://doi.org/10.1121/1.2642103

Brouwer, S., & Bradlow, A. R. (2014). Contextual variability during speech-in-speech recognition. *The Journal of the Acoustical Society of America, 136*(1), EL26-EL32. https://doi.org/10.1121/1.4881322

Chandrasekaran, B., Van Engen, K., Xie, Z., Beevers, C. G., & Maddox, W. T. (2015). Influence of depressive symptoms on speech perception in adverse listening conditions. *Cognition and Emotion*, *29*(5), 900-909. https://doi.org/10.1080/02699931.2014.944106

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America, 123*(1), 414-427. https://doi.org/10.1121/1.2804952

De Smedt, T. & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13*(Jun), 2063–2067. https://www.clips.uantwerpen.be/pages/pattern

Freyman, R. L., Helfer, K. S., & Balakrishnan, U. (2007). Variability and uncertainty in masking by competing speech. The Journal of the Acoustical Society of America, 121(2), 1040-1046. https://doi.org/10.1121/1.2427117

Golestani, N., Rosen, S., & Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Bilingualism: Language and Cognition, 12*(3), 385-392. https://doi.org/10.1017/S1366728909990150

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (2007). *TOAL-4: Test of Adolescent and Adult Language—Fourth Edition.*

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. https://doi.org/10.1121/1.2180210

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech communication*, *52*(11-12), 864-886. https://doi.org/10.1016/j.specom.2010.08.014

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. Journal of Statistical Software, 82(13). https://dx.doi.org/10.18637/jss.v082.i13

Li, P., Zhang, F., Tsai, E., Puls, B. (2014). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition, 17*(3), 673-680. https://doi.org/10.1017/S1366728913000606

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Meador, D., Flege, J. E., & MacKay, I. R. (2000). Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition*, *3*(1), 55-67. https://doi.org/10.1017%2FS1366728900000134

Reetzke, R., Lam, B. P. W., Xie, Z., Sheng, L., & Chandrasekaran, B. (2016). Effect of simultaneous bilingualism on speech intelligibility across different masker types, modalities, and signal-to-noise ratios in school-age children. *PloS one, 11*(12), e0168048. https://doi.org/10.1371/journal.pone.0168048

Reetzke, R., Xie, Z., & Chandrasekaran, B. (September 2017). Effects of selective attention and language experience on cortical entrainment to continuous speech. Poster presented at the *6th International Conference on Auditory Cortex*, Banff, Alberta, Canada.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., ... & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in systems neuroscience*, *7,* 31. https://doi.org/10.3389/fnsys.2013.00031

Smayda, K. E., Van Engen, K. J., Maddox, W. T., & Chandrasekaran, B. (2016). Audio-visual and meaningful semantic context enhancements in older and younger adults. *PloS one, 11*(3), e0152773. https://doi.org/10.1371/journal.pone.0152773

Smith, K. G., & Fogerty, D. (2017). Speech recognition error patterns for steady-state noise and interrupted speech. *Journal of the Acoustical Society of America*, *142*(3), 306–312. https://doi.org/10.1121/1.5003916

Van Engen, K. J. (2012). Speech-in-speech recognition: A training study. *Language and Cognitive Processes*, *27*(7-8), 1089-1107. https://doi.org/10.1080/01690965.2012.654644

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and speech*, *53*(4), 510-540. https://doi.org/10.1177/F0023830910372495

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PloS one, 7*(9), e43753. https://doi.org/10.1371/journal.pone.0043753

Viswanathan, J., Rémy, F., Bacon-Macé, N., & Thorpe, S. J. (2016). Long term memory for noise: evidence of robust encoding of very short temporal acoustic patterns. *Frontiers in Neuroscience*, *10*, 490. https://doi.org/10.3389/fnins.2016.00490

Watson, C. S. (1987). Uncertainty, informational masking, and the capacity of immediate auditory memory. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 267-277). Routledge.

Xie, Z., Yi, H. G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PloS one, 9(*12), e114439. https://doi.org/10.1371/journal.pone.0114439

Xie, Z., Maddox, W. T., Knopik, V. S., McGeary, J. E., & Chandrasekaran, B. (2015). Dopamine receptor D4 (DRD4) gene modulates the influence of informational masking on speech recognition. *Neuropsychologia*, *67*, 121-131. https://doi.org/10.1016/j.neuropsychologia.2014.12.01