Recognition of foreign-accented speech in noise: The interplay between talker intelligibility and linguistic structure[a)]

Dorina Strori,[1,b)] Ann R. Bradlow,[2] and Pamela E. Souza[1]

[1]Department of Communication Sciences and Disorders, Northwestern University, 2240 Campus Drive, Evanston, Illinois 60208, USA

[2]Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, Illinois 60208, USA

ABSTRACT:

Foreign-accented speech recognition is typically tested with linguistically simple materials, which offer a limited window into realistic speech processing. The present study examined the relationship between linguistic structure and talker intelligibility in several sentence-in-noise recognition experiments. Listeners transcribed simple/short and more complex/longer sentences embedded in noise. The sentences were spoken by three talkers of varying intelligibility: one native, one high-, and one low-intelligibility non-native English speakers. The effect of linguistic structure on sentence recognition accuracy was modulated by talker intelligibility. Accuracy was disadvantaged by increasing complexity only for the native and high intelligibility foreign-accented talkers, whereas no such effect was found for the low intelligibility foreign-accented talker. This pattern emerged across conditions: low and high signal-to-noise ratios, mixed and blocked stimulus presentation, and in the absence of a major cue to prosodic structure, the natural pitch contour of the sentences. Moreover, the pattern generalized to a different set of three talkers that matched the intelligibility of the original talkers. Taken together, the results in this study suggest that listeners employ qualitatively different speech processing strategies for low- versus high-intelligibility foreign-accented talkers, with sentence-related linguistic factors only emerging for speech over a threshold of intelligibility. Findings are discussed in the context of alternative accounts. © *2020 Acoustical Society of America.*

https://doi.org/10.1121/10.0001194

(Received 13 October 2019; revised 14 April 2020; accepted 17 April 2020; published online 5 June 2020)

[Editor: Megha Sundara] Pages: 3765–3782

I. INTRODUCTION

One of the most remarkable aspects of speech processing is our ability to understand a variable and complex signal like speech under a wide range of adverse conditions. Adverse conditions may originate in the speech source (i.e., the talker), the environment (i.e., the speech transmission channel), or the receiver (i.e., the listener) (Mattys *et al.*, 2012). A common source-related adverse condition is foreign-accented (i.e., second-language) speech, which arises from mismatches between first-language (L1) and second-language (L2) linguistic structures, and typically requires more processing effort on the part of the listener (e.g., Anderson-Hsieh *et al.*, 1992; Bradlow and Bent, 2008; Munro and Derwing, 1995, and many others).

Foreign-accented speech is encountered frequently in daily communications by virtually all listener populations, from toddlers in nurseries to elderly adults in assisted-living facilities. Recent estimations by the U.S. Census Bureau revealed that 21.8% of the population over the age of 5 in the United States speaks a language other than English at home, and 8.5% of this population reported speaking English "less than very well" (U.S. Census Bureau, 2017). Many of these individuals (although, not all) presumably speak English with a noticeable foreign accent.

Foreign-accented English is characterized by several deviations from native English at both the segmental and supra-segmental levels, such as changes in rhythm and tonal patterns (i.e., speech prosody) (Adams and Munro, 1978), as well as discrete spectro-temporal features that signal the identity of consonants and vowels (Flege and Eefting, 1988; Fox *et al.*, 1995; MacKay *et al.*, 2000).

The challenge for L1 English listeners of understanding L2-accented English speech most likely depends on a combination of factors related to the talker, environment, and linguistic nature of the message (Bradlow and Pisoni, 1999). The L2 proficiency level of the talker and the relevant L1–L2 typological differences may affect the extent to which L2-accented speech deviates from L1 norms, such that greater deviation results in increased processing load and understanding challenges on the part of the listener (Bradlow and Bent, 2008; Wilson and Spaulding, 2010). Further compounding these challenges may be the presence of environmental degradation, such as background noise.

[a)]Portions of this work were presented in "Sentence recognition in noise: Effects of talker intelligibility and linguistic complexity," Annual Meeting of the Psychonomic Society of America, Vancouver, British Columbia, Canada, November 2017, "Sentence recognition in noise: The interplay between talker intelligibility and linguistic complexity," Meeting of the Acoustical Society of America, Minneapolis, MN, May 2018, and "Recognizing foreign-accented speech in noise: Disentangling the effects of talker intelligibility, linguistic complexity and prosodic structure realization," Annual Meeting of the Psychonomic Society of America, Montréal, Quebec, Canada, November 2019.

[b)]Electronic mail: dorina.strori@northwestern.edu

0001-4966/2020/147(6)/3765/18/$30.00

Previous work has examined the role of talker accent in speech recognition in both quiet (Burda *et al.*, 2003; Gordon-Salant *et al.*, 2010a; Munro and Derwing, 1995) and noisy listening conditions (Ferguson *et al.*, 2010; Gordon-Salant *et al.*, 2015; Gordon-Salant *et al.*, 2013; Gordon-Salant *et al.*, 2010b; Munro, 1998; van Wijngaarden *et al.*, 2002; Wilson and Spaulding, 2010; Rogers *et al.*, 2004). The general pattern of findings has revealed that listeners display lower speech recognition performance for foreign-accented talkers compared to native-accented talkers (Burda *et al.*, 2003; Ferguson *et al.*, 2010; Gordon-Salant *et al.*, 2015; Gordon-Salant *et al.*, 2013; Gordon-Salant *et al.*, 2010a,b; Munro and Derwing, 1995), and that performance for foreign-accented talkers declines with increasing talker accentedness (Gordon-Salant *et al.*, 2010a; Gordon-Salant *et al.*, 2013; Gordon-Salant *et al.*, 2015). Further, background noise has been shown to have adverse effects on foreign-accented speech recognition. Listeners exhibit lower recognition accuracy in noise than in quiet and performance decreases with decreasing signal-to-noise ratio (SNR) (Munro, 1998; Gordon-Salant *et al.*, 2010b; van Wijngaarden *et al.*, 2002; Wilson and Spaulding, 2010).

However, the role of the linguistic nature of the speech message (isolated word, simple sentence, longer discourse, etc.), remains relatively unexplored. Notably, studies on foreign-accented speech recognition have typically employed linguistically simple testing materials, ranging from individual words (e.g., Burda *et al.*, 2003; Ferguson *et al.*, 2010; Gordon-Salant *et al.*, 2010a; Gordon-Salant *et al.*, 2015), to simple sentences with no supporting semantic context (e.g., Gordon-Salant *et al.*, 2010b; Gordon-Salant *et al.*, 2013). While the use of testing materials that have a simple, balanced linguistic structure with limited context is understandably convenient for controlled laboratory settings, a major disadvantage lies in the fact that linguistically simple sentences represent a narrow window into real-world speech experience. In everyday situations, listeners are exposed to speech samples of varying linguistic structure. Current literature shows a growing interest in tackling the issue of limited ecological validity linked to the use of linguistically impoverished speech materials for testing speech performance. To this end, there have been various efforts toward creating materials and tasks that are more representative of the processing and communicative demands of realistic situations. For instance, a line of work has targeted increasing the variability in testing materials in terms of both talkers and the linguistic structure of sentences (e.g., the Perceptually Robust English Sentence Test—Open set (PRESTO) sentences: Gilbert *et al.*, 2013; Tamati *et al.*, 2013; Tamati and Pisoni, 2014; Faulkner *et al.*, 2015; Plotkowski and Alexander, 2016). Other studies have designed dynamic tasks that emulate realistic conversations, such as the so-called "problem-solving" tasks (e.g., the "Map Task," Anderson *et al.*, 1991, the several versions of the "diapix" task [American English: Van Engen *et al.*, 2010; British English (diapixUK): Baker and Hazan, 2011; Spanish (diapixFL): Wester *et al.*, 2014; Lecumberri *et al.*, 2017], and a speech comprehension test that targets emulating features of realistic speech communication experiences (The National Acoustic Laboratories Dynamic Conversations Test (NAL-DCT): Best *et al.*, 2016; Best *et al.*, 2018). Studies have also examined the role of communicative intent in speech produced in adverse listening situations and have shown that talkers modify their speech to a greater extent in interactive situations and adapt it in ways that benefit their interlocutors (Garnier *et al.*, 2010; Hazan and Baker, 2011).

In a separate line of research, several studies have revealed that syntactic structure impacts the processing of and memory for sequences of words (Bever *et al.*, 1969; Bonhage *et al.*, 2014; Epstein, 1961; Jarvella and Herman, 1972; Miller and Isard, 1963; Marks and Miller, 1964; Potter and Lombardi, 1998). Thus, listeners' understanding of foreign-accented speech will likely not only hinge on talker-dependent variation in overall intelligibility (a signal-related factor) but also on variation in the linguistic structure of the utterance (e.g., a simple statement versus a more complex sentence with more than one clause).

The present study aimed at exploring the relationship between talker intelligibility and linguistic structure in the context of foreign-accented sentence recognition in noise. The manipulation of the linguistic structure involved creating a contrast in the sentences with regard to their linguistic structure, resulting in sentences deemed to be linguistically simple and sentences that were considered to be linguistically more complex. The operationalization of the notion of "linguistic complexity" was motivated by the heterogeneous and multifaceted linguistic nature of sentences in daily communications. To this end, "linguistic simplicity versus complexity" was construed along several linguistic criteria that were deemed to contribute to an increased processing difficulty for the sentences.

In regard to the main focus of the study, the relationship between talker intelligibility and linguistic structure, we anticipated two possible outcomes. One possibility is that the effect of sentence complexity is driven by talker-independent factors and is primarily dependent on the abstract syntactic structure of the sentences. In this case, the difference in recognition accuracy between simple and complex sentences would be comparable across all talkers regardless of native versus non-native status and regardless of the degree of foreign-accentedness and overall intelligibility for non-native talkers. More specifically, we may anticipate that listeners will be more likely to misperceive words in complex sentences than in simple sentences as a result of the extra processing burden required to form mental representations of complex sentences.

Alternatively, the effect of sentence complexity may interact with talker-dependent variation in overall intelligibility, such that a certain threshold of intelligibility is required to evoke complex syntactic processing. In this case, a difference in recognition accuracies for simple versus complex sentences will be more pronounced for talkers with higher overall intelligibility than for low intelligibility talkers because, without accurate recognition at the segment and word levels, listeners may not be able to reconstruct the intended hierarchical syntactic structure of a complex sentence. When intelligibility at the segment and words levels is low and leads to more speech recognition errors and effortful listening (as in the case of low intelligibility foreign-accented speech), listeners may resort to a processing strategy that generally precludes syntactic constituent "chunking" and effectively treats all sentences as non-hierarchically ordered word sequences. A somewhat counter-intuitive prediction of this hypothesis is that the typical decrease in recognition accuracy for complex (i.e., multi-clausal) sentences relative to simple (i.e., mono-clausal) sentences will not be observed for relatively low intelligibility foreign-accented speech.

In several sentence-in-noise recognition experiments, we examined the effect of simultaneously varying the intelligibility of the talker and the linguistic structure of the sentences on the recognition of foreign-accented English sentences by native listeners. The first experiment consisted of three separate conditions that explored the interplay between talker intelligibility and sentence structure in two SNRs (one low: Experiment 1 A, and one higher: Experiment 1B) and in two formats of stimulus presentation (mixed: Experiments 1 A and 1B, and blocked-by-talker: Experiment 1 C). We reasoned that the inclusion of two SNRs would allow us to compare the relationship between talker intelligibility and linguistic complexity across difficult and easier listening conditions, and also aid to rule out any potential confounds related to potential floor and/or ceiling recognition performance levels. Similarly, the two types of stimulus presentation (mixed versus blocked-by-talker) aimed at testing the relationship of interest in two different listening situations: one in which the talker was uniform within a block (blocked presentation: three blocks of trials, one block per talker) and one situation in which the talker varied across trials (mixed presentation: one block of trials for all the talkers). The first listening situation reduces talker variability on a trial-by-trial level and provides the listener with the opportunity to become accustomed to the talker within the corresponding block of sentences by means of consistent exposure to the same talker for the duration of the block. In comparison, the second listening situation involves talker variability on a trial-by-trial level and as such, does not provide the same opportunity to the listener to become accustomed to a talker via consistent exposure. We were interested to examine whether the relationship between talker intelligibility and linguistic structure would display differences across these two listening situations.

The second experiment examined the generalizability of the patterns found in the first experiment, namely, whether they extended to a different set of three talkers that matched the intelligibility levels of the first set of talkers. It consisted of two conditions, one involving the low (Experiment 2A) and the other involving the higher (Experiment 2B) SNR values employed in the first experiment. We reasoned that the replicability of patterns of interest observed in the first experiment to a different set of talkers would further consolidate these findings.

The third experiment sought to investigate the relationship between talker intelligibility and linguistic structure in the absence of one of the main cues to prosodic structure: the natural pitch contour of the sentences. Studies have shown that natural pitch variations (the natural pitch contour) play an important role in facilitating sentence recognition in both quiet (e.g., Laures and Weismer, 1999) and noisy environments (e.g., Binns and Culling, 2007; Laures and Bunton, 2003; Miller *et al.*, 2010; Shen and Souza, 2017). We were interested in examining whether besides intelligibility, potential differences across talkers in conveying this important cue to prosodic structure also played a role in the relationship between talker intelligibility and linguistic structure. Talkers who are not proficient at producing acoustic cues to prosodic structure, such as non-native talkers, may not adequately convey the necessary prosodic cues that aid the listener in forming mental representations of the hierarchical syntactic structure of complex sentences. This could lead to the listener employing a processing strategy that does not dissociate between simple and structurally more complex sentences, resulting in the neutralization of any potential differences in recognition accuracy between these two types of sentences. Removing the natural pitch contour from the sentences (i.e., modifying the natural pitch contour of the sentences to be flat) allows us to investigate the consequences for speech recognition of neutralizing this potential production difference across the three talkers. More specifically, we asked whether the removal of pitch cues to prosody would affect the relationship between talker intelligibility and linguistic structure. We reasoned that findings from this experiment would provide insights on the possible role of prosodic structure realization in the relationship of interest. Namely, a different pattern of findings compared to the other experiments would provide some support for the role of prosodic cue realization by the talker in evoking effects of linguistic structure on the listener's sentence recognition accuracy. Alternatively, the emergence of the same pattern of findings as in the other experiments would weaken to some extent the plausibility of the prosodic structure realization argument.

II. METHODS

A. Participants

The first phase of the study, consisting of Experiment 1 A–1C, was conducted in the laboratory. Participants were students at Northwestern University who received course credit or payment for their participation. There were 31 participants in Exp. 1 A (Age range: 18–22 years; Mean age: 19.42 years), 30 participants in Exp. 1B (Age range: 18–24 years; Mean age: 20.32 years) and 35 participants in Exp. 1 C (Age range: 18–21 years; Mean age: 19.39 years). Three participants had to be excluded from Exp. 1 C due to a technical problem with the experimental software, resulting in 32 participants included in the data analysis. All participants consented to participate in the study and identified as native speakers of American English. Participants completed a questionnaire that included questions about their language background and their speech, hearing, and cognitive abilities. None of the participants reported any hearing, speech, language or cognitive problems.

The second and third phases of the study, consisting of Experiment 2A–2B (Phase 2) and Experiment 3 (Phase 3), were conducted online via the crowd-sourcing online platform Amazon's Mechanical Turk (AMT), with listeners recruited throughout the United States. The benefits of using crowdsourcing for speech processing research have been reviewed (e.g., Eskénazi *et al.*, 2013) and researchers working with experimental paradigms in phonetics, phonology, psycholinguistics, and sociolinguistics have successfully executed tasks through AMT, including discrimination, recognition, classification, and identification in noise of phonetically detailed stimuli (e.g., D'Onofrio, 2019; Walker and Campbell-Kibler, 2015; Yu and Lee, 2014).[1] In order to ensure that participants were completing the experiment in good faith and to the best of their ability, checks were put in place during recruitment and analysis. Participants were restricted to those who specified US origin in their "worker" profiles on the site, and every participant was required to complete a sound check to ensure that they were, in fact, listening to the auditory stimuli prior to beginning the task.[2] Through the embedded online interface, no participant was allowed to take the experiment more than one time. All participants identified as native speakers of American English and completed a questionnaire that included questions about their language background and their speech, hearing, and cognitive abilities. Participants who reported a hearing, speech, language, or cognitive problem were excluded. Forty-five participants participated in Exp. 2A (Age range: 21–50; Mean age: 32.96 years). Three participants had to be excluded from data analysis: two who declared a hearing problem and one who declared a speech-related problem. This resulted in 42 participants included in the data analysis. Forty-five participants (Age range: 18–46 years; Mean age: 30.84 years) participated in Exp. 2B, one of whom declared a cognitive problem and was excluded from the study, resulting in 44 participants included for data analysis. Forty-five participants participated in Experiment 3 (Age range: 18–50 years; Mean age: 30.71 years). Two participants were excluded from data analysis: one participant who declared a cognitive problem and one participant who provided null responses to every trial, suggesting inability to perform the task. This resulted in 43 participants included in the data analysis. All AMT participants completed an informed consent process and were paid for their time.

B. Sentences and noise

The sentence stimuli consisted of 93 English sentences obtained from the Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings (ALLSSTAR) created and maintained by the Speech Communication Research Group at Northwestern University (Bradlow; for description of the corpus, see also Bradlow *et al.*, 2017; Bradlow *et al.*, 2018). This comprehensive corpus includes digital speech recordings from over 120 talkers producing a variety of speech materials in their native language (L1) and non-native language (L2), English. The materials for the present study consisted of 60 sentences from the Hearing in Noise Test (HINT) (Nilsson *et al.*, 1994) and 33 sentences that were a combination of sentences from the Universal Declaration of Human Rights (DHR:17 sentences) and the story "The Little Prince" (LPP:16 sentences). The HINT sentences are short, semantically plausible sentences that have a linguistically simple structure (mono-clausal, canonical declarative syntax).[3] These sentences constituted the group of "Simple" sentences. The rest of the sentences represented the group of "Complex" sentences. Complex sentences were longer and had a more complex syntactic structure (multi-clausal, non-canonical syntax, and/or passive voice) compared to the HINT sentences. Specifically, in order to be classified as complex, a sentence had to be longer than the maximum length of a HINT sentence (seven words), as well as meet some, or all of the other linguistic criteria outlined in Table VII (Table VII and the list of all sentences are provided in the Appendix). The goal was to create two contrasting sets of sentences: one that consisted of sentences typically used in speech recognition tasks due to their conveniently simple linguistic structure, and one set that was more representative of realistic speech experiences and included sentences with more diverse and complex linguistic structures. In this respect, the HINT sentences represented a homogenous set, whereas the rest of the sentences constituted a heterogeneous set of sentences.

The same sentences were used in all experiments and the corresponding digital speech files were all leveled to equate root-mean-square (rms) amplitude across the full set. The sentences were produced by a total of six talkers, divided into two sets of three talkers each. One set of talkers was used in the first and third experiments (set 1), and the other set of talkers was used in the second experiment (set 2). Within each talker set, one talker was a native speaker of American English (L1: English, the "Native" talker) and the other two talkers were native speakers of Mandarin Chinese and non-native speakers of English (L1: Mandarin Chinese, L2: English, the "Accented" talkers). Further, the two non-native talkers varied with respect to their overall intelligibility in English. Namely, one talker had a high intelligibility score in English ("Accented-High") and the other talker had a low intelligibility score in English ("Accented-Low"). These intelligibility scores were obtained in a separate, unrelated sentence-in-noise recognition study with native listeners of American English (Bradlow *et al.*, 2018) and are presented in Table VIII as percentage correct recognition accuracies (Table VIII can be found in the Appendix). The two sets of talkers were matched in terms of the intelligibility of each talker category (i.e., Native, Accented-High, Accented-Low). There were six versions of a sentence, each spoken by one of the talkers, resulting in a total of 558 (6 talkers × 93 sentences) stimulus files.

In Experiment 3, the natural pitch ($F0$) contours of all the 279 stimulus files (3 talkers × 93 sentences) were manipulated and the sentence files were resynthesized using the

Praat software (Boersma and Weenink, 2018) via the method of pitch-synchronous overlap-and-add (Moulines and Charpentier, 1990). The purpose of the manipulation was to remove one of the major cues to prosodic structure, the natural variations in the pitch ($F0$) contour, and neutralize differences in production related to this cue across the three talkers. For each stimulus file, a "flat pitch contour" version was created via the following formula:

$$\text{Instant } F0 = \text{Sentence average } F0. \tag{1}$$

The pitch contour manipulation was implemented prior to the sentences being mixed with noise, as explained below.

The sentences were embedded in steady-state broadband noise, which was the speech-shaped noise used in the original HINT (Nilsson *et al.*, 1994). The sentences were mixed with the noise via a custom made program in MATLAB (release 2018b) (Mathworks, 2018) at the specified SNRs. Two SNRs were used across the experimental conditions: a "low" (–4 dB) and a "high" (0 dB). The choice of the initial SNR (–4 dB), implemented in the first experimental condition (Exp. 1A), was informed by relevant literature (e.g., Bradlow *et al.*, 2018) and by a pilot study run prior to the experiment.[4]

In Experiment 1, the first condition (Exp. 1A) involved the low SNR (–4 dB) and the second condition (Exp. 1B) involved the high SNR (0 dB). The third condition, Exp. 1 C, involved the low SNR, in parallel to Exp. 1A.[5]

In the second experiment, the first condition (Exp. 2A) involved the low SNR (–4 dB), whereas the second condition (Exp. 2B) involved the high SNR (0 dB), in parallel to the corresponding conditions in the first experiment (Exp. 1A and 1B, respectively).

The third experiment involved only the high SNR (0 dB). Since the sentence stimuli in this experiment had been acoustically modified to have a flat pitch contour prior to being embedded in noise, a process that may lead to lower overall sentence recognition accuracies even in quiet (e.g., Laures and Weismer, 1999), it was reasoned that the low SNR (–4 dB) could make the task too difficult for the listeners.

The desired SNR was achieved by keeping the rms amplitude of the sentences constant and scaling the rms amplitude of the noise. During the mixing process, random portions of the noise file were selected for each sentence, the mixed signal was preceded by 500 milliseconds of noise only, and the first and last 100 milliseconds of the mixed signal were tapered on and off to avoid noise-related artifacts to the listener. The stimuli were pre-processed (i.e., the sentences were mixed with the noise) before being presented to the listeners.

C. Procedure

The final stimuli were presented to listeners (in-lab and AMT participants) via a single page custom web application. Participants completed the study at their own pace. After listening to a trial, they typed their responses in a text input box in the web browser, then pressed the return key, after which the next trial was presented. A short tone preceded each trial to help focus the listener's attention to the arrival of the trial and minimize any element of surprise or abruptness. All stimulus files were preloaded to minimize wait times between items.

Listeners in the in-lab conditions (Exp. 1A–1C) were seated in a double-walled sound-treated booth, in front of a computer. They read instructions on the computer screen and were also verbally informed by the experimenter that they would listen to sentences in background noise. Participants were instructed to ignore the noise and to type in the sentence as accurately as possible, guessing if necessary. All the trials were presented to the listeners binaurally over headphones (Sennheiser HD 25–1 II), at 68 dB sound pressure level (SPL) using a professional headphone amplifier (APHEX HeadPod 454).

The AMT listeners (Exp. 2A–2B and Exp. 3) only read instructions on their computer screens and were instructed to complete the experiment in as quiet an environment as possible, use good quality headphones that had to be worn on both ears, and adjust the volume of their computers to the most comfortable level for them. These instructions aimed at ensuring as quiet and controlled listening environment as possible, in the absence of a well-controlled laboratory setting.

Experiment 1 consisted of three conditions, two of which involved a mixed design (varying the talker trial-to-trial), Exp. 1A (low SNR) and Exp. 1B (high SNR), and one that involved a blocked design, Exp. 1C (low SNR), in terms of the stimulus presentation method. In the mixed design, the sentences from all three talkers were presented in one, mixed block, whereas in the blocked design, the sentences were presented in three separate blocks, each of which included all the sentences (simple and complex) spoken by one talker.[6]

Experiment 2 consisted of two conditions, Exp. 2A (low SNR) and Exp. 2B (high SNR), both of which involved stimuli produced by a different set of three talkers from Experiment 1 (same talkers for both Exp. 2A and 2B) and the mixed design of stimulus presentation.

Experiment 3 consisted of one condition (high SNR, same talkers as in Experiment 1, flattened pitch contour of the sentences) that involved the mixed design of stimulus presentation.

The presentation of the experimental trials was counterbalanced and randomized across participants in all the experimental conditions. Counterbalancing the distribution of sentences among the three talkers resulted in six different stimulus presentation lists (3! possible orders for the 3 talkers), such that each sentence was produced by each talker and a sequence of three sentences (consecutively numbered) was assigned to a different order of the three talkers across the lists (same order of three talkers for sequences of three sentences within a list). The same six counterbalancing lists were involved in all the experimental conditions (mixed and blocked experimental designs) and a listener was randomly assigned to one. In all the experimental conditions, the 93 sentences were evenly distributed among the three talkers in each presentation list, such that each talker spoke a total of 31 sentences (20 simple and 11 complex sentences). The randomization of trial presentation was different between the mixed and blocked designs. In the mixed design (varying the talker trial-to-trial), a presentation list consisted of one block that included all 93 sentences (each of the three talkers producing 31 sentences), randomized for their order of presentation. In the blocked design (Exp. 1C), a presentation list consisted of three separate blocks, each including all the 31 sentences produced by one talker, randomized for their order of presentation within the individual block. Listeners heard a sentence only once throughout the duration of the experiment. Prior to the experimental trials, listeners completed six practice trials, three of which were simple sentences and three that were complex sentences. The practice trials were different from the experimental trials and they were spoken by a different set of three talkers that matched the intelligibility range of the three talkers that produced the experimental trials.

D. Data scoring

Recognition scores were determined using a word correct count. While the typical scoring procedure for tests of English intelligibility in both research and clinical settings involves measures of keyword recognition (i.e., recognition of content words to the exclusion of intervening grammatical function words), we counted both content word and function word recognition accuracy. This was primarily motivated by the variable linguistic context of the complex sentences, as well as the fact that this study constituted the first to obtain recognition scores for these sentences. Further, Bradlow *et al.* (2018) compared the all-word scoring method and a keyword-only scoring method for some of their experimental conditions and found a very high correlation between the two scoring methods. In the present study, a word was counted as correctly recognized (scored as 1) if it was transcribed perfectly, including all affixes (e.g., plural "s" and past tense "ed"). Obvious spelling errors or homophones were also counted as correct. Incorrect words were scored as 0 and the recognition accuracy for each sentence was calculated as the proportion of correctly recognized words, n/N, where n is the number of correctly recognized words in a sentence and N is the total number of words in the sentence. In order to assess scoring reliability, a part of the responses of the first experimental condition (Exp.1A) were scored by two scorers. The inter-scorer agreement was 98.5%, which was deemed high enough to rely on only a single scorer for the rest of the collected listener responses.

III. ANALYSIS AND RESULTS

A. Talker intelligibility and sentence complexity

Listeners' sentence recognition performance was analyzed in relation to the two experimental manipulations: talker intelligibility and sentence structure, both manipulated within subjects. Sixty-two listener responses out of a total of 16 647 (<0.5%) responses (60 in Exp. 1A and 2 in Exp. 2A) had to be excluded from analysis due to a software problem traced after the data collection was completed.[7] Data analysis was conducted in R (R Core Team, 2019, version 3.6.1), using generalized linear mixed effects models (GLMMs) (Baayen *et al.*, 2008) that involved logistic regression with beta-binomial distribution for the response variable.[8] GLMMs that use logistic regression and assume a binomial distribution of the response variable have been recommended over their linear counterparts to analyse proportion (or percentage) data (Bolker *et al.*, 2009; Jaeger, 2008). However, using a strictly binomial distribution for proportion data that represent sentence recognition accuracy does not adequately capture the fact that the probability p of correctly recognizing a word in a sentence of N words may vary across the words in the sentence due to contextual effects. In addition, the binomial distribution is also prone to the *overdispersion* problem, which means that the model underestimates the actual variance in the data (see Ferrari and Comelli, 2016, for more detailed explanations and comparison of different linear and generalized linear mixed models). An alternative to overcome these issues and provide a more accurate model fitting of proportion data is using the beta-binomial distribution for the proportion response variable.

Unlike the binomial distribution, the aforementioned p is not fixed in each trial[9] in the beta-binomial distribution, but instead varies randomly following a beta distribution, thus accounting for the possibility of contextual effects. The beta-binomial regression has been proposed as an alternative to linear regression for analysing proportion data since a long time and its suitability to the analysis of this type of data has been shown in various contexts (Crowder, 1978, Hilbe, 2013; Muniz-Terrera *et al.*, 2016; Prentice, 1986). Recent advances in statistical software have made it possible to analyze proportion data with generalized linear mixed effects regression models that implement the beta-binomial distribution.

In the present models, fixed effects of talker (Talker, 3 levels: Native, Accented-High, and Accented-Low), sentence complexity (Complexity, 2 levels: Simple and Complex), and their interaction served as the primary effects of interest. Random intercepts of participant and item (sentence) were included in the models. The dependent variable was sentence recognition accuracy (proportion correct). The binary factor Complexity was dummy coded (0: Simple, 1: Complex, default = Simple) and the factor Talker was contrast coded using the forward difference coding system (default = Native). Forward difference coding, a strategy for coding categorical variables in mixed effects modelling, compares the mean of the dependent variable on a specific level of the independent variable to the mean of the dependent variable for the next (adjacent) level of the independent variable. In the present models, Talker was contrast coded as follows: The first contrast compared differences between the native and high intelligibility non-native talkers (2/3,

TABLE I. Summary of model comparisons for assessing the interaction between talker intelligibility and sentence complexity in each experimental condition. “Talker Set” refers to whether the three talkers were in the first, or the second set.

						Talker x Complexity		
Experiment	Medium	Description	SNR	Design	Talker Set	χ^2	*df*	*p*
1A	In-lab	Difficult	−4 dB	Mixed	1	125.75	2	<0.001 (***)
1B	In-lab	Easy	0 dB	Mixed	1	101.00	2	<0.001 (***)
1C	In-lab	Uniform Talker	−4 dB	Blocked	1	144.27	2	<0.001 (***)
2A	AMT	New Talkers—Difficult	−4 dB	Mixed	2	127.40	2	<0.001 (***)
2B	AMT	New Talkers—Easy	0 dB	Mixed	2	84.00	2	<0.001 (***)
3	AMT	Flat Pitch Contour	0 dB	Mixed	1	174.73	2	<0.001 (***)

–1/3, –1/3); the second contrast compared differences between the high and low intelligibility non-native talkers (1/3, 1/3, –2/3). The main focus of the analysis, the interaction between talker intelligibility and sentence complexity, was assessed by performing likelihood-ratio tests between the models with and without the interaction term in each experimental condition. These model comparisons revealed the presence of an interaction in all conditions. The interaction term significantly improved the model, which represented the best fit for the data. Table I provides a summary of the model comparisons and Table II displays summaries of the predictors in the models with the interaction term across all experimental conditions. The base category (in Intercept) in the models corresponds to the combination of native talker and simple sentences. Sentence recognition accuracies for each talker and sentence type combination are presented in Table III.

As displayed in Table III, the overall sentence recognition accuracy dropped with increasing sentence complexity and decreasing talker intelligibility across all conditions.

TABLE II. Summaries of the best fit models of talker intelligibility and sentence complexity, for each experimental condition. Coefficient values are given in the log-odds ratio scale. “Talker1” and “Talker2” represent the two levels of the contrast-coded predictor “Talker.”

Experiment		Predictors	*β*	*SE*	*z*	*Pr(>/z/)*
1A		Talker1 (Native vs Accented-High)	1.79	0.11	16.74	<0.001 (***)
Difficult (In-lab)		Talker2 (Accented-High vs -Low)	2.27	0.08	26.63	<0.001 (***)
SNR	−4 dB	Complexity	−0.71	0.16	−4.34	<0.001 (***)
Design	Mixed	Talker1: Complexity	−0.64	0.14	−4.61	<0.001 (***)
Talker Set	1	Talker2: Complexity	−0.94	0.12	−7.67	<0.001 (***)
1B		Talker1 (Native vs Accented-High)	1.89	0.15	12.87	<0.001 (***)
Easy (In-lab)		Talker2 (Accented-High vs -Low)	1.78	0.09	20.47	<0.001 (***)
SNR	0 dB	Complexity	−0.83	0.20	−4.18	<0.001 (***)
Design	Mixed	Talker1: Complexity	−0.86	0.18	−4.81	<0.001 (***)
Talker Set	1	Talker2: Complexity	−0.73	0.12	−6.04	<0.001 (***)
1C		Talker1 (Native vs Accented-High)	1.82	0.10	17.87	<0.001 (***)
Blocked (In-lab)		Talker2 (Accented-High vs -Low)	2.3	0.08	27.92	<0.001 (***)
SNR	−4 dB	Complexity	−0.71	0.17	−4.24	<0.001 (***)
Design	Blocked	Talker1: Complexity	−0.62	0.13	−4.61	<0.001 (***)
Talker Set	1	Talker2: Complexity	−0.99	0.12	−8.48	<0.001 (***)
2A		Talker1 (Native vs Accented-High)	1.69	0.07	23.92	<0.001 (***)
New Talkers-Difficult (AMT)		Talker2 (Accented-High vs -Low)	1.53	0.07	22.63	<0.001 (***)
SNR	−4 dB	Complexity	−0.45	0.16	−2.87	0.004 (**)
Design	Mixed	Talker1: Complexity	−0.88	0.10	−9.07	<0.001 (***)
Talker Set	2	Talker2: Complexity	−0.23	0.10	−2.31	0.02 (*)
2B		Talker1 (Native vs Accented-High)	1.25	0.09	14.3	<0.001 (***)
New Talkers-Easy (AMT)		Talker2 (Accented-High vs -Low)	1.85	0.07	27.05	<0.001 (***)
SNR	0 dB	Complexity	−0.8	0.16	−4.9	<0.001 (***)
Design	Mixed	Talker1: Complexity	−0.54	0.11	−4.7	<0.001 (***)
Talker Set	2	Talker2: Complexity	−0.47	0.10	−4.84	<0.001 (***)
3		Talker1 (Native vs Accented-High)	1.84	0.08	23.09	<0.001 (***)
Flat Pitch Contour (AMT)		Talker2 (Accented-High vs -Low)	1.98	0.07	28.38	<0.001 (***)
SNR	0 dB	Complexity	−0.63	0.12	−5.43	<0.001 (***)
Design	Mixed	Talker1: Complexity	−0.88	0.11	−8.25	<0.001 (***)
Talker Set	1	Talker2: Complexity	−0.62	0.10	−6.05	<0.001 (***)

TABLE III. Mean sentence recognition accuracies across experimental conditions for each talker and sentence type combination, presented as percentage correct (%).

		Talker	Complex sentences	Simple sentences
Exp. 1A				
Difficult				
SNR	−4 dB	Native	78.13	92.83
Design	Mixed	Accented-High	54.27	69.60
Talker Set	1	Accented-Low	25.03	25.27
Exp. 1B				
Easy				
SNR	0 dB	Native	90.29	97.55
Design	Mixed	Accented-High	75.62	85.22
Talker Set	1	Accented-Low	53.06	55.21
Exp. 1C				
Uniform Talker				
SNR	−4 dB	Native	78.86	93.44
Design	Blocked	Accented-High	54.48	70.48
Talker Set	1	Accented-Low	25.57	25.35
Exp. 2A				
New Talkers—Difficult				
SNR	−4 dB	Native	65.57	81.93
Design	Mixed	Accented-High	46.08	51.10
Talker Set	2	Accented-Low	20.58	22.47
Exp. 2B				
New Talkers—Easy				
SNR	0 dB	Native	78.63	90.40
Design	Mixed	Accented-High	65.72	76.85
Talker Set	2	Accented-Low	37.45	43.70
Exp. 3				
Flat Pitch Contour				
SNR	0 dB	Native	69.56	87.66
Design	Mixed	Accented-High	47.90	59.40
Talker Set	1	Accented-Low	20.76	21.23

The more interesting finding, the interaction between talker intelligibility and sentence complexity displayed in Tables I and II, suggests that listeners did not benefit from recognizing simple sentences compared to more complex sentences to the same extent across the three talkers, as anticipated in our second hypothesis. To assess where these differences may lie, pairwise comparisons of estimated mean recognition accuracies of simple and complex sentences across the three talkers were conducted.[10] The results are summarized in Table IV.

These pairwise comparisons revealed that listeners were significantly more accurate at recognizing simple sentences compared to complex sentences when they were spoken by the native talker, in all experimental conditions. This was also the case for the high intelligibility non-native talker, except in Experiment 2A. In contrast, listeners did not benefit from simpler sentences (or suffer from more complex sentences) when they were spoken by the low intelligibility non-native talker, a pattern that was consistent across conditions.

The continuation of this pattern in Experiment 3 indicates that the removal of pitch cues to prosodic structure did not eliminate the difference between the high- and low-intelligibility non-native talkers in evoking an effect of the linguistic structure on sentence recognition accuracy.

Figure 1 graphically illustrates the observed interaction between talker intelligibility and sentence structure.

The effect of SNR on overall sentence recognition accuracy and on the interaction between talker intelligibility and linguistic structure was assessed in each pair of the relevant experimental conditions, namely, Exp. 1A and 1B, and Exp. 2A and 2B. The datasets in each pair of conditions were collapsed and a new factor, SNR, was added to the analysis. The two levels of the SNR factor were dummy coded as 0 (Low) and 1 (High). The model of interest included the interaction among three factors: Talker, Complexity, and SNR, whereby the first two represented within-subjects and the latter between-subjects manipulations. The presence of an interaction was assessed by performing likelihood-ratio tests between the models with and without the interaction term. Random intercepts for participant and item (sentence) were included in the models. Results revealed an effect of SNR in overall recognition accuracy in Exp. 1A–1B, $\beta = 1.13$, $\mathrm{SE} = 0.11$, $z = 9.94$; $p < 0.0001$, and in Exp. 2A–2B, $\beta = 1.12$, $\mathrm{SE} = 0.16$, $z = 6.80$; $p < 0.0001$. As it can be expected, listener's overall sentence recognition accuracy was higher in the high SNR condition compared to the low SNR condition. However, no interaction between SNR and the other two factors was found in either Exp. 1A–1B, $\chi^2(2) = 2.06$, $p = 0.36$, or in Exp. 2A–2B, $\chi^2(2) = 5.90$, $p = 0.052$, suggesting that the effect of linguistic structure was modulated by talker intelligibility to a similar extent in both difficult and easier listening conditions.

The effect of the stimulus presentation method (mixed versus blocked design) on overall sentence recognition accuracy and on the interaction between talker intelligibility and linguistic structure was assessed in a similar way. The datasets of the corresponding conditions, Exp. 1A and 1C, were collapsed and a new factor, Design, was added to the analysis. The two levels of the Design factor were dummy coded as 0 (Mixed) and 1 (Blocked). The model of interest included the interaction among three factors: Talker, Complexity, and Design, whereby the first two represented within-subjects and the latter between-subjects manipulations. The presence of an interaction was assessed by performing likelihood-ratio tests between the models with and without the interaction term. Random intercepts for participant and item (sentence) were included in the models. Results revealed no effect of the method of stimulus presentation (Design) on overall sentence recognition accuracy, $\beta = 0.03$, $\mathrm{SE} = 0.11$, $z = 0.32$; $p = 0.75$, and no interaction between Design and the other two factors, $\chi^2(2) = 0.17$, $p = 0.92$. This indicates that using a uniform talker in the blocked condition did not help listeners' overall sentence recognition accuracy and it did not lead to a different interaction pattern between talker intelligibility and sentence complexity compared to varying the talker from trial-to-trial.

TABLE IV. Pairwise comparisons of the estimated marginal means for each talker and sentence complexity combination. Results are given on the log-odds ratio scale. *P*-values are adjusted by the Tukey method for multiple comparisons.

Experiment		Simple vs complex sentence recognition accuracy	*β*	*SE*	*t*	*Pr(>/t/)*
1A - Difficult		*Talker*				
SNR	−4 dB	Native	1.45	0.19	7.56	<0.0001 (***)
Design	Mixed	Accented-High	0.81	0.17	4.63	0.0001 (***)
Talker Set	1 (Lab)	Accented-Low	−0.12	0.18	−0.69	0.98
1B - Easy						
SNR	0 dB	Native	1.65	0.24	6.76	<0.0001 (***)
Design	Mixed	Accented-High	0.79	0.21	3.77	0.002 (**)
Talker Set	1 (Lab)	Accented-Low	0.06	0.20	0.30	0.999
1C - Uniform Talker						
SNR	−4 dB	Native	1.45	0.19	7.53	<0.0001 (***)
Design	Blocked	Accented-High	0.83	0.18	4.70	<0.0001 (***)
Talker Set	1 (Lab)	Accented-Low	−0.16	0.18	−0.92	0.94
2A - New Talkers - Difficult						
SNR	−4 dB	Native	1.11	0.17	6.64	<0.0001 (***)
Design	Mixed	Accented-High	0.23	0.16	1.40	0.73
Talker Set	2 (AMT)	Accented-Low	0.0004	0.17	0.003	1.00
2B - New Talkers - Easy						
SNR	0 dB	Native	1.31	0.18	7.24	<0.0001 (***)
Design	Mixed	Accented-High	0.78	0.17	4.50	0.0001 (***)
Talker Set	2 (AMT)	Accented-Low	0.31	0.17	1.83	0.45
3 - Flat Pitch Contour						
SNR	0 dB	Native	1.42	0.14	10.45	<0.0001 (***)
Design	Mixed	Accented-High	0.54	0.13	4.29	0.0003 (***)
Talker Set	1 (AMT)	Accented-Low	−0.07	0.13	−0.56	0.99

B. Talker intelligibility and length of complex sentences

As described above, a distinctive feature of the complex sentences is their heterogeneous nature with respect to several linguistic features, aimed as an analogy to realistic speech experiences. One of these features, sentence length varied considerably across the set of complex sentences, which provided motivation for a separate analysis of recognition accuracy as a function of sentence length and talker intelligibility within this set. Since length and complexity are essentially inseparable features in a sentence (an increase in one feature will often involve an increase in the other one, too), this analysis sought to convey a more nuanced view of the relationship between talker intelligibility and the linguistic nature of sentences (simple/shorter versus complex/longer). To this end, this additional analysis involved length as a continuous variable (i.e., number of words in the sentence) instead of the simple/complex dichotomy in the previous analysis. Given that length is not a variable feature in the homogenous set of simple sentences (length range of only 4–6 words, with 56.67% of sentences having length = 5), this analysis was confined to the set of complex sentences (length range of 8–20 words, with mean of 14 and median of 14).

Identical model fitting and comparisons as in the main analysis described above were performed, with Length as a continuous numerical predictor (centered around its mean value), instead of the binary factor Complexity. Model comparisons revealed an interaction between Talker and Length in all experimental conditions, indicating that the effect of sentence length on listeners' recognition accuracies was modulated by the talker. Table V presents a summary of the model comparisons. Figure 2 illustrates the effect of sentence length for each talker, across experimental conditions. It is important to note that, due to removal of the simple sentences, the analyses in Table V and the data in Fig. 2 are based on considerably fewer data points (lower statistical power) than the prior analysis with the dichotomous variable.

To assess the differences among talkers in regard to the effect of sentence length on recognition accuracy, the data was first divided into three sets, each corresponding to the sentences produced by one talker. Afterward, the main effect of sentence length on recognition accuracy was assessed in each data set via likelihood-ratio tests that compared the models with and without the Length factor. Results across experimental conditions are displayed in Table VI.

Despite some variation across conditions, the overall pattern displayed in Table VI and Fig. 2 is that the influence of sentence length on complex sentence recognition accuracy is strongest and most consistently observed for the native talker, and weakest for the low intelligibility non-native talker.

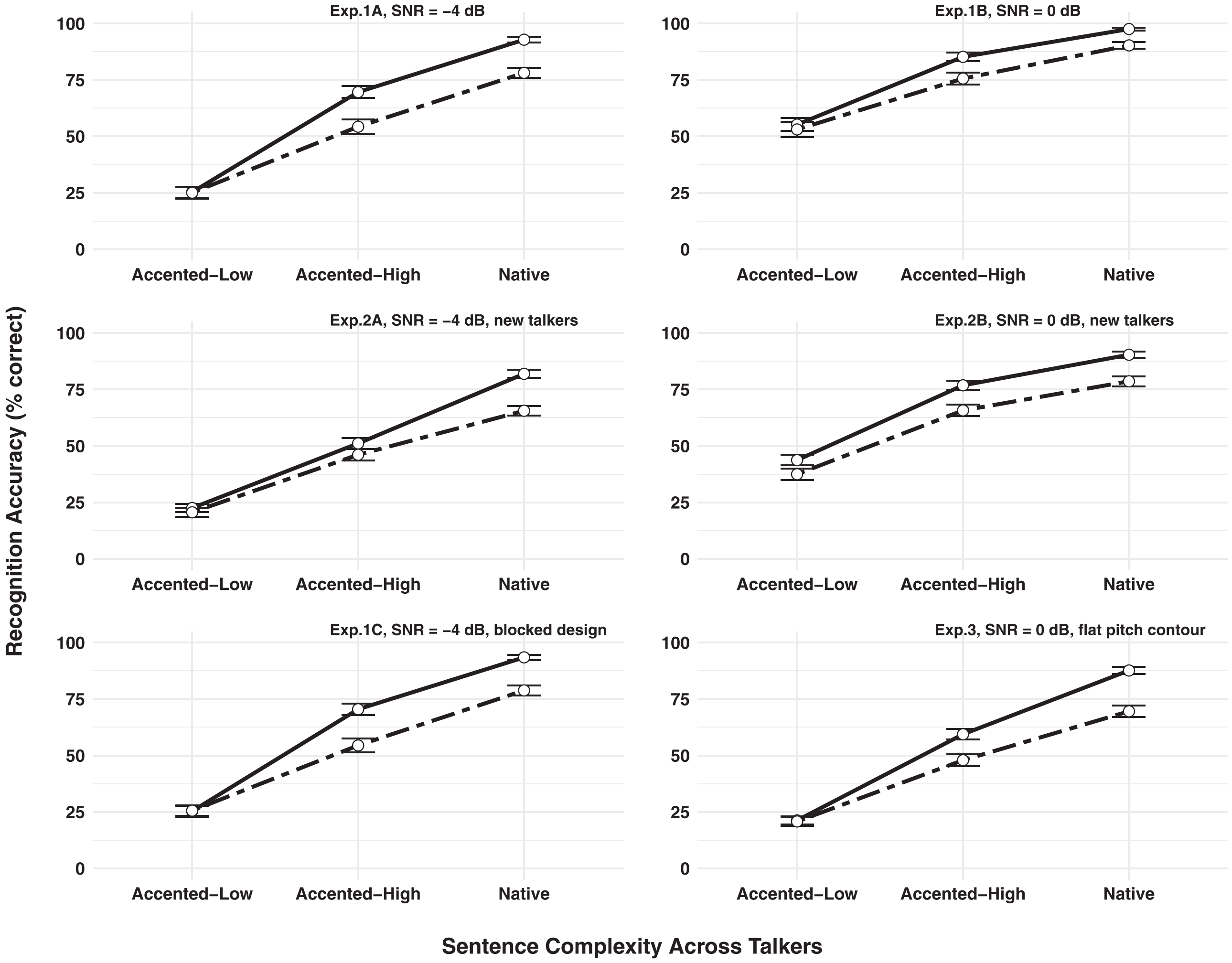

FIG. 1. Mean recognition accuracies across all talker and sentence combinations for all experimental conditions. The solid line represents the simple sentences and the dashed line represents the complex sentences. Error bars are displayed as 95% confidence intervals (CI).

IV. DISCUSSION

A. Talker intelligibility and linguistic structure

This study explored the relationship between a talker-dependent factor (intelligibility) and a sentence-dependent factor (linguistic structure) in several listening conditions. The findings suggest that the linguistic structure of the uttered message has an effect on listeners' recognition performance and that this effect is modulated by the intelligibility of the talker. Recognition accuracy was higher for simple sentences compared to more complex sentences, however, this recognition benefit, or lack thereof, was modulated by the intelligibility of the talker. Specifically, listeners recognized simple sentences more accurately than complex

TABLE V. Summary of model comparisons for assessing the interaction between talker intelligibility and sentence length in the set of complex sentences for each experimental condition.

						Talker x Complex sentence length		
Experiment	Medium	Description	SNR	Design	Talker Set	χ^2	*df*	*p*
1A	In-lab	Difficult	−4 dB	Mixed	1	16.09	2	<0.001 (***)
1B	In-lab	Easy	0 dB	Mixed	1	18.51	2	<0.001 (***)
1C	In-lab	Uniform Talker	−4 dB	Blocked	1	7.74	2	0.02 (*)
2A	AMT	New Talkers—Difficult	−4 dB	Mixed	2	17.35	2	<0.001 (***)
2B	AMT	New Talkers—Easy	0 dB	Mixed	2	16.46	2	<0.001 (***)
3	AMT	Flat Pitch Contour	0 dB	Mixed	1	37.15	2	<0.001 (***)

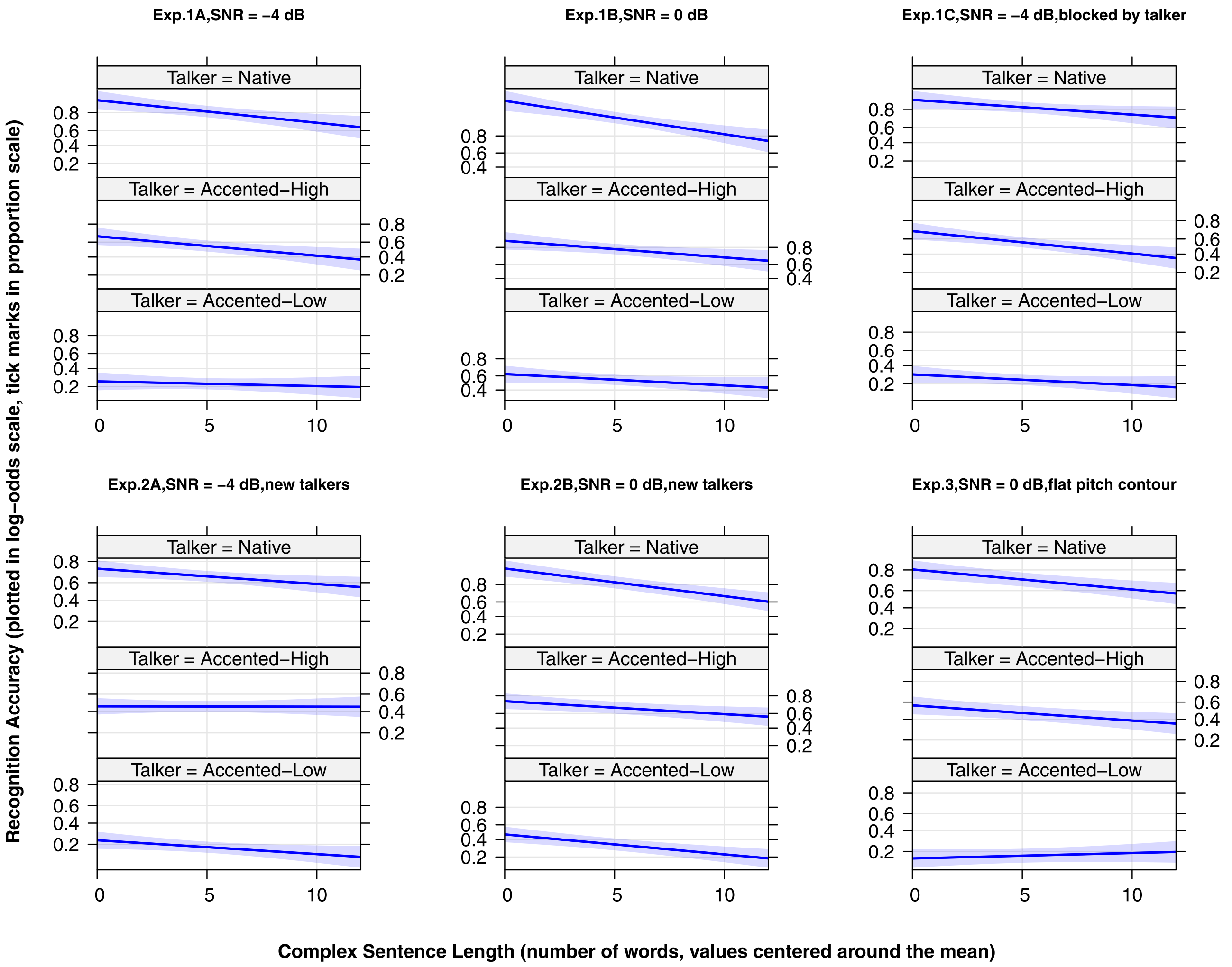

FIG. 2. (Color online) The effect of sentence length for each talker in the set of complex sentences across conditions. The plotted effect was derived from the same generalized linear mixed effect model with the interaction between talker and sentence length that was implemented in the corresponding data analysis. The shaded areas represent confidence intervals, visually shown as Scheffe-type confidence envelopes, at 99% CI. The values on the horizontal axis correspond to the length of the complex sentences, centered around the mean value. The values on the *y*-axis represent sentence recognition accuracy, plotted in the scale of the linear predictor (sentence length) in the generalized linear mixed effects model, i.e., log-odds, with the tick marks shown in the mean (proportion correct) scale (see Fox *et al*., 2019, the tutorial of the "effects" R package, for more details on *y*-axis specification).

sentences only when they were produced by the native and high intelligibility non-native talkers. In the case of the low intelligibility non-native talker, the linguistic structure of the sentences did not play a role on the listener's recognition accuracy. This pattern was present across the different conditions tested.

The first experiment discovered this relationship and investigated it in different listening scenarios related to the level of noise (low vs high SNR) and the type of stimulus presentation (mixed vs blocked), all of which displayed the same overall pattern. The presence of the same robust interaction between talker intelligibility and sentence complexity in both difficult (Exp. 1A) and easier (Exp. 1B) listening conditions ruled out the possibility that the initial pattern of findings may have occurred due to a floor recognition performance for the low intelligibility non-native talker. Further, a comparative inspection of the pattern of results in the low and high SNR conditions provides insights on a potential confound, the "indexical load" introduced by a non-native accent. More specifically, in the high SNR condition, the overall recognition accuracy for complex sentences spoken by the low-intelligibility L2 talker was almost equivalent to the overall recognition accuracy for complex sentences spoken by the high-intelligibility L2 talker in the low SNR condition (Table III: approximately 53% and 54% correct accuracy for complex sentences, respectively). Nevertheless, while there was an increase in recognition accuracy for simple sentences in the case of the high-intelligibility L2 talker in the low SNR, no such increase was observed for the low-intelligibility L2 talker in the high SNR condition (Table III: approximately 70% and 55% correct accuracy for simple sentences, respectively). This comparison within the L2 talkers, who presumably involve the same indexical load, suggests that the effect of linguistic structure is more likely related to the intelligibility of the talkers than to "indexical load."

TABLE VI. Summary of model comparisons for assessing the main effect of sentence length on recognition accuracy in the set of complex sentences for each talker and across all experimental conditions.

			Complex sentence length		
Experiment	Description	Talker	χ^2	*df*	*p*
1A	**Lab - Difficult**				
SNR	−4 dB	Native	5.85	1	0.01 (*)
Design	Mixed	Accented-High	3.43	1	0.06
Talker Set	1	Accented-Low	0.42	1	0.51
1B	**Lab - Easy**				
SNR	0 dB	Native	13.78	1	<0.001 (***)
Design	Mixed	Accented-High	4.25	1	0.04 (*)
Talker Set	1	Accented-Low	3.26	1	0.07
1C	**Lab - Uniform Talker**				
SNR	−4 dB	Native	2.42	1	0.12
Design	Blocked	Accented-High	5.23	1	0.02 (*)
Talker Set	1	Accented-Low	0.62	1	0.43
2A	**AMT - New Talkers**				
SNR	−4 dB	Native	8.00	1	0.005 (**)
Design	Mixed	Accented-High	0.004	1	0.95
Talker Set	2	Accented-Low	1.00	1	0.32
2B	**AMT - New Talkers**				
SNR	0 dB	Native	22.29	1	<0.001 (***)
Design	Mixed	Accented-High	2.26	1	0.13
Talker Set	2	Accented-Low	5.24	1	0.02 (*)
3	**AMT - Flat Pitch Contour**				
SNR	0 dB	Native	8.47	1	0.004 (**)
Design	Mixed	Accented-High	1.59	1	0.21
Talker Set	1	Accented-Low	0.35	1	0.55

The second experiment examined whether the pattern of findings observed in the first experiment generalized to a different set of talkers that matched the intelligibility range of the first set of talkers. Results revealed that the interaction between talker intelligibility and sentence complexity discovered in the first experiment generalized to the new set of three talkers in the second experiment. The pattern was consistently replicated in the second condition (high SNR) and partly replicated in the first condition (low SNR) of the second experiment. More specifically, while an interaction between talker intelligibility and sentence complexity was found in the low SNR condition (Exp. 2A) consistent with all the other conditions, a significant recognition benefit for simple over more complex sentences only emerged for the sentences spoken by the native talker. Unlike in the other experimental conditions, the high intelligibility non-native talker did not elicit a similar recognition benefit for the simple sentences. A possible explanation for this discrepancy in the pattern of findings across the two experiments may rely on talker-related differences. That is, although the two high intelligibility non-native talkers have similar intelligibility scores, talker-related differences may still lead to differences in recognition patterns, especially in more challenging listening conditions, where there is more uncertainty and/or ambiguity regarding the speech message. This view is supported by the fact that the original pattern of findings found in the first experiment was replicated in the high SNR condition of the second experiment (Exp. 2B). Listeners showed a recognition benefit for the simple sentences spoken by the high intelligibility non-native talker when listening conditions became more favorable in the high SNR condition. The role of talker-related differences on listeners' speech recognition performance, widely referred to as indexical effects, is well-documented in the literature, where a plethora of studies has shown that changing the talker may lead to differences in performance (e.g., Bradlow *et al.*, 1999; Church and Schacter, 1994; Luce and Lyons, 1998; Goldinger, 1996; Nygaard *et al.*, 1994; Palmeri *et al.*, 1993; Strori *et al.*, 2018).

However, the aforementioned discrepancy may also be ascribed in part to factors related to different mediums of testing across the two experiments. Namely, the listening environment was well-controlled in the laboratory setting of the first experiment, whereas the listening environment in the web-based platform of the second experiment was not controlled. Further, factors related to the participants in each experiment (such as differences in the age range, cognitive ability, hearing sensitivity, years of education), may have also played a role. While several measures that aimed at minimizing the effects of this variability to a good extent were implemented, including a larger number of participants, sound checks, participant background questionnaires, and rigorous instructions, the fact remains that it is not possible to have the same level of control over the two testing mediums.

The third experiment investigated whether, besides intelligibility, potential differences across talkers in conveying a major cue to prosodic structure also played a role in the relationship between talker intelligibility and linguistic structure. Results revealed the same overall pattern of findings observed in the other experiments, suggesting that the removal of pitch cues to prosodic structure did not eliminate the difference between the high and low intelligibility non-native talkers in evoking a cost/benefit of the linguistic structure on sentence recognition accuracy.

The analysis of recognition accuracy as a function of sentence length and talker intelligibility within the set of complex sentences revealed a relatively similar pattern to the analysis on sentence complexity and talker intelligibility. This finding, together with observations regarding the relationship between sentence length and complexity in the present study, suggest a more nuanced interpretation of the results. That is, length and complexity are virtually inextricable features of a sentence, and the fact that they varied together in the present study makes it difficult to attribute the observed differences in performance between the simple and complex sentences to either differences in sentence complexity or length. Nevertheless, regardless of whether this sentence-by-talker interaction is driven by sentence complexity per se, or by the number of words in the sentence (i.e., sentence length), the present results indicate that low intelligibility non-native speech involves an attenuation of the influence of sentence-dependent (i.e., linguistic structure) variation on recognition of words in sentence contexts.

B. What drives the modulated effect of linguistic structure on foreign-accented speech recognition?

A plausible explanation for the present pattern of findings relies on speech intelligibility. Sentence recognition requires the listener to successfully identify word boundaries and group constituent words and phrases into superordinate and subordinate clauses (i.e., to form mental representations of hierarchical syntactic structures). Studies have shown that speech is processed in terms of a hierarchy of units, rather than on a sequential, segment by segment basis (e.g., Wingfield, 1975). It is reasonable to assume that the more complex the structure of the sentence becomes, the more difficult it is for listeners to form the mental representations of these sentences. As a result, recognition of complex/longer sentences could be expected to be more vulnerable to disruption, such as that from background noise, than recognition of simple/shorter sentences, which is what we observe when non-native speech is above a certain threshold of overall intelligibility. Namely, words in simpler sentences are less prone to recognition error than words in more complex sentences. In contrast, when overall intelligibility is very low, as in the case of the low-intelligibility non-native talkers, listeners seem to adopt a listening strategy that does not differentiate between words in simple sentences and words in more complex sentences. In this case, "sentences" are not perceived as syntactic constructs, but are effectively reduced to sequences of words with no internal, hierarchical organization that helps the listener to form mental representations of syntactic structures and benefit from simple sentences (or alternatively, be disadvantaged by more complex sentences). That is, if not enough words are accurately recognized (i.e., intelligibility is too low), then listeners do not receive enough linguistic information (i.e., sufficient number of words, or the right words) to fit into syntactic structures that would in turn guide sentence recognition. In this case, the difference between recognizing simple versus more complex sentences disappears.

Another possible explanation is based on the ability of the talkers to convey crucial cues to prosodic structure. Namely, talkers who have not mastered the production of acoustic cues to prosodic structure (e.g., non-native talkers) may produce sentences (both complex/long and simple/short) that lack systematic, syntactically-informed hierarchical structure. In the present study, it could be the case that unlike the high-intelligibility non-native talkers, the low-intelligibility non-native talkers are not efficiently conveying the acoustic cues to prosody that help the listener build hierarchical sentence structure. Consequently, listeners adopt a so-called "shallow" listening strategy that remains impervious to variation in sentence structure, displayed in similar recognition accuracies for simple and more complex sentences. The results of the third experiment weaken the plausibility of this explanation, since neutralizing the difference across talkers in conveying pitch cues to prosodic structure did not alter the observed interaction between talker intelligibility and linguistic structure. Similar to the native talker, the high-intelligibility non-native talker continued to evoke an effect of linguistic structure on recognition accuracy in the absence of the natural pitch contour, as opposed to the low-intelligibility non-native talker who did not elicit such an effect. It is worth noting that the implications from this finding are limited to one of the acoustic cues to prosodic structure. While the pitch contour is considered a major cue to prosodic structure (e.g., Cole, 2015; Cutler *et al.*, 1997), it may still be the case that other important prosodic cues that were not manipulated here, such as stress and timing patterns, may be playing a role. In this respect, this result constitutes important, yet preliminary evidence on the role of prosodic structure realization.

A third plausible account that can accommodate the present findings is the Ease of Language Understanding (ELU) Model (Rönnberg *et al.*, 2013; Rönnberg *et al.*, 2008). Rönnberg and colleagues have proposed that when speech is processed in difficult listening conditions, listeners have difficulty matching the degraded input to stored semantic representations and as such, resort to the explicit use of cognitive resources, such as working memory capacity, to process the signal. On the assumption that cognitive resources constitute a finite pool (e.g., Just and Carpenter, 1992; Kahneman, 1973), higher reliance on these resources to understand a heavily degraded signal may result in fewer resources left for other tasks related to the signal. In the present case, listeners may be struggling more to understand the more degraded sentences produced by the low-intelligibility non-native talkers compared to the sentences produced by the high-intelligibility non-native (and native) talkers. Consequently, listeners may be allocating disproportionately more cognitive resources towards deciphering the sentences produced by the low-intelligibility non-native talkers as compared to the high-intelligibility non-native talkers, resulting in fewer resources left for constructing mental representations of hierarchical, syntactic sentence structure. In the absence of syntactic structures that guide sentence recognition, the difference between recognizing simple versus more complex sentences is neutralized for low-intelligibility non-native speech.

V. CONCLUSION

The present study examined the relationship between talker intelligibility and linguistic structure on recognition accuracy for native and non-native sentences in noise. It revealed a robust interplay between these two crucial factors across different listening conditions. These results suggest that listeners seem to exhibit different listening strategies for high versus low intelligibility non-native talkers.

This work provides new insights on foreign-accented speech recognition and represents an initial step into understanding the challenge posed by this type of commonly encountered speech from a realistic perspective. It highlights the role of the structure of the linguistic message (simple/short versus more complex/longer sentences) by including sentences that go beyond the simplistic nature of the materials typically used in speech recognition studies, which offer limited ecological validity. The present findings may inform assessment and training approaches to foreign-accented speech in relevant settings, such as in the clinic, industry, the classroom, nursing centers, or any other real-world situations that present listening challenges. Specifically, the most effective and efficient strategies for assessing and enhancing speech communication under realistic conditions (e.g., through either listener or talker training) should incorporate speech materials of varying inherent intelligibility and/or linguistic structure. The data from this study indicate that overreliance on homogenous (and usually linguistically simple) test materials can obscure important material-dependent variation across foreign-accented talkers of varying overall intelligibility.

ACKNOWLEDGMENTS

This research was supported by a grant from the Knowles Hearing Center at Northwestern University and NIH-NIDCD Grant No. R01-DC005794. The authors thank Chun Liang Chan for his technical support, Yasmine Elmasri for her help with data scoring, Daniel Turner for his help with data analysis, and Jing Shen for her help with the preparation of the stimuli for one of the experiments.

APPENDIX: SENTENCE AND TALKER DETAILS

The information provided below displays the linguistic criteria implemented in the categorization of sentences as simple versus complex, the complete list of the sentence stimuli, and the intelligibility scores of the six talkers used in the study.

1. Linguistic criteria for determining sentence complexity

TABLE VII. Some examples of simple and complex sentences, and the several linguistic criteria used for determining the “simple versus complex” distinction.

Example sentence	Length	No. of verbs	Modifiers (adjectives, adverbs)	Conjunctions (and, or, but)	Non-declarative syntax/Non-canonical word order	Passive verb	Negation	Sentence type
Everyone has the right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay.	19	2	Yes	Yes	No	No	No	Complex
How could he recognize me when he had never seen me before?	12	2	Yes	Yes	Yes	No	Yes	Complex
The wife helped her husband.	5	1	No	No	No	No	No	Simple
The food is expensive.	4	1	Yes	No	No	No	No	Simple

2. List of the sentence stimuli

Simple sentences (HINT)

A boy fell from the window.
The shoes were very dirty.
The fire was very hot.
The car is going too fast.
Swimmers can hold their breath.
The food is expensive.
Milk comes in a carton.
A towel is near the sink.
The janitor swept the floor.
The match boxes are empty.
The boy did a handstand.
They waited for an hour.
The milk is in a pitcher.
Her coat is on a chair.
The dog is chasing the cat.
The baby has blue eyes.
They wanted some potatoes.
The teapot is very hot.
The girl is washing her hair.
They called an ambulance.
Big dogs can be dangerous.
Somebody stole the money.
The picture came from a book.
The family bought a house.
They are pushing an old car.
They had two empty bottles.
The boy is running away.
He's skating with his friend.
She took off her fur coat.
The dog came home at last.
The young people are dancing.
They watched the scary movie.
The tall man tied his shoes.
The girl is fixing her dress.
The bus leaves before the train.
They are coming for dinner.
The waiter brought the cream.
The jelly jar was full.
The policeman knows the way.
He climbed up the ladder.
The wife helped her husband.
The player lost a shoe.
The team is playing well.
The painter uses a brush.
She stood near the window.
The children are walking home.
The dog sleeps in a basket.
Flowers can grow in the pot.
The football game is over.
The man is painting a sign.
They took some food outside.
The shirts are in the closet.
The truck drove up the road.
They are running past the house.
The milkman drives a small truck.
The bag fell off the shelf.
They knocked on the window.
The apple pie is good.
The girl played with the baby.
They are drinking coffee.

Complex sentences (grouped by their sources)

The Universal Declaration of Human Rights

No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment.

Everyone has the right to recognition everywhere as a person before the law.

Everyone has the right to own property alone as well as in association with others.

Everyone, without any discrimination, has the right to equal pay for equal work.

Parents have a prior right to choose the kind of education that shall be given to their children.

Everyone has the right to freedom of movement and residence within the borders of each State.

Everyone has the right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay.

No one shall be arbitrarily deprived of his property.

No one shall be arbitrarily deprived of his nationality nor denied the right to change his nationality.

Everyone has the right of equal access to public service in his country.

No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms.

Everyone has the right to freedom of peaceful assembly and association.

Everyone has the right to leave any country, including his own, and to return to his country.

Everyone has the right to form and to join trade unions for the protection of his interests.

Everyone has the right to life, liberty, and the security of person.

No one may be compelled to belong to an association.

No one shall be subjected to arbitrary arrest, detention or exile.

The Little Prince

Beside the well there was the ruin of an old stone wall.

Boa constrictors swallow their prey whole, without chewing it.

Bit by bit, I came to understand the secrets of your sad little life.

How could he recognize me when he had never seen me before?

The Earth is not just an ordinary planet.

And he lay down in the grass and cried.

Why should anyone be frightened by a hat?

I believe that for his escape he took advantage of the migration of a flock of wild birds.

The businessman opened his mouth, but he found nothing to say in answer.

The little prince crossed the desert and met with only one flower.

"I am right here," the voice said, "under the apple tree."

It took me a long time to learn where he came from.

A sheep - if it eats little bushes, does it eat flowers, too?

"The grown-ups are certainly very odd," he said to himself, as he continued on his journey.

All humanity could be piled up on a small Pacific islet.

And now six years have already gone by

3. Talker intelligibility scores

TABLE VIII. Talker intelligibility scores at two SNRs displayed as the percentage of correctly recognized words by native English listeners. In the case of the native talkers, the low and high SNRs were –8 and –4 dB, respectively, and for the foreign-accented talkers, the low and high SNRs were –4 and 0 dB, respectively (data from Bradlow *et al.*, 2018).

Talker	Language background	Task language	Recognition accuracy - low SNR	Recognition accuracy - high SNR
Native 1	English	English	79.56	96.50
Native 2	English	English	79.22	94.74
High Intel.1	Mandarin Chinese	English	68.55	86.89
High Intel.2	Mandarin Chinese	English	67.99	83.26
Low Intel.1	Mandarin Chinese	English	29.99	55.91
Low Intel.2	Mandarin Chinese	English	33.95	55.85

[1] Prior to collecting experimental data on AMT, we ran a pilot verification study with ten listeners on this platform with the materials used in the first in-lab condition (Exp. 1A, low SNR of –4 dB). The aim of the pilot study was to compare the patterns of findings between the two testing mediums and proceed with AMT testing if these patterns matched to a high extent. Results revealed identical patterns to those observed in the in-lab experiments (Exp. 1A–1C): (1) recognition accuracy decreased with decreasing talker intelligibility and increasing linguistic complexity), and (2) the effect of linguistic complexity was modulated by the intelligibility of the talker, such that recognition for complex sentences was lower than recognition accuracy for simple sentences, but only for the native and high-intelligibility foreign-accented talkers.

[2] In the audio test, the AMT participants transcribed two words (“donkey” and “zebra”) at the start of the experiment to ensure that their computers could play audio. They had to transcribe the two words correctly to proceed through to the next stage of the experiment. We also included detailed instructions for the participants regarding the listening environment and the use of good quality headphones that had to be worn binaurally, as in the laboratory testing.

[3] The length of HINT sentences ranges from four to seven words and the vast majority of them have only one verb, with few sentences having two verbs. All the HINT sentences used in the present study were from four to six words long and none of them had more than one verb.

[4] The pilot study was run with a different set of participants (five young normal-hearing listeners) and it measured speech reception thresholds (SRTs) for the HINT sentences spoken by one of the native talkers (Set 1) and the four foreign-accented talkers (talker sets 1 and 2). Scoring was performed on a keyword basis (three keywords/content words for each HINT sentence) and the targeted accuracy level was 66.67% correct (two out of three keywords correct), a level that was between 20% and 80% correct. The aim was to make the task difficult for the listener across talkers, while avoiding any potential floor/ceiling effects. The results revealed that an SNR of –4 dB would be a feasible choice for these purposes.

[5] The comparison of the patterns of results between Exp. 1A and 1B and Exp. 1A and 1C informed the decision to not use a second (“high”) SNR in Exp. 1C, since we did not expect to observe a different pattern at the high SNR.

[6] Comparisons between the patterns of results in Exp. 1A and 1C informed the decision to not implement the blocked design in the next experiments.

[7] In Exp. 1A, three sentences (two simple and one complex) were played twice for some of the presentation lists, instead of the intended sentences. This resulted in six sentences in total being excluded for each of the ten listeners in the presentation lists affected by this issue. In Exp. 2A, which took place on AMT, one (complex) sentence was played twice, instead of another sentence (simple) for one listener, which led to two sentences being excluded. Duplicate presentations may sometimes occur in crowd sourcing platforms like AMT and are virtually impossible to avoid.

[8] The models were implemented with the glmmTMB package (Brooks *et al.*, 2017).

[9] In our case, a trial represents a word in a sentence.

[10] These comparisons were implemented via the “emmeans” package (Lenth, 2019), previously known as “lsmeans.”

Adams, C., and Munro, R. R. (**1978**). “In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterance of some native and non-native speakers of English,” Phonetica **35**, 125–156.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (**1991**). “The HCRC map task corpus,” Language and Speech **34**, 351–366.

Anderson-Hsieh, J., Johnson, R., and Koehler, K. (**1992**). “The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody and syllable structure,” Lang. Learn. **42**, 529–555.

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

Baker, R., and Hazan, V. (**2011**). “DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs,” Behavior Research Methods **43**(3), 761–770.

Best, V., Keidser, G., Freeston, K., and Buchholz. J. M. (**2016**). “A dynamic speech comprehension test for assessing real-world listening ability,” Journal of the American Academy of Audiology **27**, 515–526.

Best, V., Keidser, G., Freeston, K., and Buchholz. J. M. (**2018**). “Evaluation of the NAL Dynamic Conversations Test in older listeners with hearing loss,” International Journal of Audiology **57**(3), 221–229.

Bever, T. G., Lackner, J. R., and Kir, R. (**1969**). “The underlying structures of sentences are the primary units of immediate speech processing,” Percept. Psychophys. **5**(4), 225–234.

Binns, C., and Culling, J. F. (**2007**). “The role of fundamental frequency contours in the perception of speech against interfering speech,” J. Acoust. Soc. Am. **122**, 1765–1776.

Boersma, P., and Weenink, D. (**2018**). Praat: doing phonetics by computer [Computer program]. Version 6.0.28, retrieved from http://www.praat.org/ (last viewed 14 October 2018).

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H., and White, J. S. (**2009**). “Generalized linear mixed models: A practical guide for ecology and evolution,” Trends Ecol. Evol. **24**(3), 127–135.

Bonhage, C. E., Fiebach, C. J., Bahlmann, J., and Mueller, J. L. (**2014**). “Brain signature of working memory for sentence structure: Enriched encoding and facilitated maintenance,” J. Cogn. Neurosci. **26**(8), 1654–1671.

Bradlow, A. R. “ALLSSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings,” retrieved from https://speechbox.linguistics.northwestern.edu/ALLSSTARcentral/#!/recordings (last viewed on 29 April, 2020).

Bradlow, A. R., and Bent, T. (**2008**). “Perceptual adaptation to non-native speech,” Cognition **106**, 707–729.

Bradlow, A. R., Kim, M., and Blasingame, M. (**2017**). “Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate,” J. Acoust. Soc. Am. **141**, 886–889.

Bradlow, A. R., Blasingame, M., and Lee, K. (**2018**). "Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech," J. Assoc. Lab. Phonology **9**(1), 1–20.

Bradlow, A. R., Nygaard, L. C., and Pisoni, D. B. (**1999**). "Effects of talker, rate, and amplitude variation on recognition memory for spoken words," Percept. Psychophys. **61**(2), 206–219.

Bradlow, A. R., and Pisoni, D. B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074–2085.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (**2017**). "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," The R Journal **9**(2), 378–400.

Burda, A. N., Scherz, J. A., Hageman, C. F., and Edwards, H. T. (**2003**). "Age and understanding speakers with Spanish or Taiwanese accents," Percept. Mot. Skills **97**, 11–20.

Church, B. A., and Schacter, D. L. (**1994**). "Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency," J. Exp. Psychol. Learn. Mem. Cogn. **20**, 521–533.

Cole, J. (**2015**). "Prosody in context: A review," Lang. Cogn. Neurosci. **30**(1–2), 1–31.

Crowder, M. J. (**1978**). "Beta-binomial ANOVA for proportions," J. R. Stat. Soc. Ser. C (Appl. Stat.) **27**, 34–37.

Cutler, A., Dahan, D., and van Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang. Speech **40**, 141–201.

D'Onofrio, A. (**2019**). "Complicating categories: Personae mediate racialized expectations of non-native speech," J. Sociolinguist. **23**(4), 346–366.

Epstein, W. (**1961**). "The influence of syntactical structure on learning," Am. J. Psychol. **74**(1), 80–86.

Eskénazi, M., Levow, G.-A., Meng, H., Parent, G., and Suendermann, D. (**2013**). *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment* (John Wiley & Sons Ltd., New York).

Faulkner, K. F., Tamati, T. N., Gilbert, J. L., and Pisoni, D. B. (**2015**). "List equivalency of PRESTO for the evaluation of speech recognition," J. Am. Acad. Audiol. **26**(6), 582–594.

Ferguson, S. H., Jongman, A., Sereno, J. A., and Keum, K. A. (**2010**). "Intelligibility of foreign- accented speech for older adults with and without hearing loss," J. Am. Acad. Audiol. **21**, 153–162.

Ferrari, A., and Comelli, M. (**2016**). "A comparison of methods for the analysis of binomial clustered outcomes in behavioral research," Journal of Neuroscience Methods **274**, 131–140.

Flege, J. E., and Eefting, W. (**1988**). "Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation," J. Acoust. Soc. Am. **83**, 729–740.

Fox, J., Weisberg, S., Price, B., Friendly, M., and Jangman, H. (**2019**). "Effect displays for linear, generalized linear, and other models," R package version 4.1-2.

Fox, R. A., Flege, J. E., and Munro, J. (**1995**). "The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis," J. Acoust. Soc. Am. **97**, 2540–2551.

Garnier, M., Henrich, N., and Dubois, D. (**2010**). "Influence of sound immersion and communicative interaction on the Lombard effect," J. Speech Lang. Hear. Res. **53**, 588–608.

Gilbert, J. L., Tamati, T. N., and Pisoni, D. B. (**2013**). "Development, reliability, and validity of PRESTO: A new high-variability sentence recognition test," J. Am. Acad. Audiol. **24**, 26–36.

Goldinger, S. D. (**1996**). "Words and voices: Episodic traces in spoken word identification and recognition memory," J. Exp. Psychol. Learn. Mem. Cogn. **22**, 1166–1183.

Gordon-Salant, S., Yeni-Komshian, G. H., and Fitzgibbons, P. J. (**2010a**). "Recognition of accented English in quiet by younger normal-hearing listeners and older listeners with normal hearing and hearing loss," J. Acoust. Soc. Am. **128**, 444–455.

Gordon-Salant, S., Yeni-Komshian, G. H., and Fitzgibbons, P. J. (**2010b**). "Perception of accented English in quiet and noise by younger and older listeners," J. Acoust. Soc. Am. **128**, 3152–3160.

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., and Cohen, J. I. (**2015**). "Effects of talker accent and age on recognition of multisyllabic words," J. Acoust. Soc. Am. **137**, 884–897.

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., Cohen, J. I., and Waldroup, C. (**2013**). "Recognition of accented and unaccented speech in different noise backgrounds by younger and older listeners," J. Acoust. Soc. Am. **134**, 618–627.

Hazan, V., and Baker, R. (**2011**). "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," J. Acoust. Soc. Am. **130**, 2139–2152.

Hilbe, J. M., (**2013**). "Beta binomial regression," in *The Selected Works of Joseph Hilbe* (bepress electronic repository), retrieved from http://works.bepress.com/joseph_hilbe/43/ (last viewed on 29 April, 2020).

Jaeger, T. F. (**2008**). "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," J. Mem. Lang. **59**(4), 434–446.

Jarvella, R. J., and Herman, S. J. (**1972**). "Clause structure of sentences and speech processing," Percept. Psychophys. **11**(5), 381–384.

Just, M. A., and Carpenter, P. A. (**1992**). "A capacity theory of comprehension: Individual differences in working memory," Psychol. Rev. **99**, 122–149.

Kahneman, D. (**1973**). *Attention and Effort* (Prentice-Hall, Englewood Cliffs, NJ).

Laures, J. S., and Bunton, K. (**2003**). "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions," J. Commun. Disord. **36**, 449–464.

Laures, J. S., and Weismer, G. (**1999**). "The effects of a flattened fundamental frequency on intelligibility at the sentence level," J. Speech Lang. Hear. Res. **42**, 1148–1156.

Lecumberri, M. L. G., Cooke, M., & Wester, M. (**2017**). "A bi-directional task-based corpus of learners' conversational speech," International Journal of Learner Corpus Research, **3**(2), 175–195.

Lenth, R. (**2019**). "emmeans: Estimated Marginal Means, aka Least-Squares Means," R package version 1.3.5.1, https://CRAN.R-project.org/package=emmeans last viewed on 5 March, 2020.

Luce, P. A., and Lyons, E. (**1998**). "Specificity of memory representation for spoken words," Mem. Cogn. **26**, 708–715.

MacKay, I. R. A., Flege, J. E., and Piske, T. (**2000**). "Persistent errors in the perception and production of word-initial English stop consonants by native speakers of Italian (A)," J. Acoust. Soc. Am. **107**, 2802–2802.

Marks, L. E., and Miller, G. A. (**1964**). "The role of semantic and syntactic constraints in the memorization of English sentences," Journal of Verbal Learning and Verbal Behavior **3**, 1–5.

Mathworks (**2018**). *MATLAB and Statistics Toolbox Release 2018b* (The MathWorks, Inc., Natick, MA).

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cogn. Process. **27**, 953–978.

Miller, G. A., and Isard, S. (**1963**). "Some perceptual consequences of linguistic rules," J. Verbal Learn. Verbal Behav. **2**, 217–228.

Miller, S. E., Schlauch, R. S., and Watson, P. J. (**2010**). "The effects of fundamental frequency contour manipulations on speech intelligibility in background noise," J. Acoust. Soc. Am. **128**, 435–443.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.

Muniz-Terrera, G., van den Hout, A., Rigby, R. A., and Stasinopoulos, D. M. (**2016**). "Analyzing cognitive test data: Distributions and non-parametric random effects," Stat. Methods Med. Res. **25**(2), 741–753.

Munro, M., and Derwing, T. (**1995**). "Foreign accent, comprehensibility and intelligibility in the speech of second language learners," Lang. Learn. **45**, 73–97.

Munro, M. J. (**1998**). "The effects of noise on the intelligibility of foreign-accented speech," Stud. Second Lang. Acquist. **20**, 139–154.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker contingent process," Psychol. Sci. **5**, 42–46.

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (**1993**). "Episodic encoding of voice attributes and recognition memory for spoken words," J. Exp. Psychol. Learn. Mem. Cogn. **19**, 309–328.

Plotkowski, A. R., & Alexander, J. M. (**2016**). "A Sequential Sentence Paradigm Using Revised PRESTO Sentence Lists," Journal of the American Academy of Audiology, **27**(8), 647–660.

Potter, M. C., and Lombardi, L. (**1998**). "Syntactic priming in immediate recall of sentences," J. Mem. Lang. **38**, 265–282.

Prentice, R. L. (**1986**). "Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors," J. Am. Stat. Assoc. **81**, 321–327.

R Core Team (2019). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ last viewed on 10 August, 2019.

Rogers, C. L., Dalby, J., and Nishi, K. (**2004**). "Effects of noise and proficiency on intelligibility of Chinese-accented English," Lang. Speech **47**, 139–154.

Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (**2008**). "Cognition counts: a working memory system for ease of language understanding (ELU)," International Journal of Audiology, **47**(Suppl. 2), S99–S105.

Rönnberg, J., Lunner, T., Zekveld, A., Sorqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. C., and Rudner, M. (**2013**). "The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances," Frontiers in Systems Neuroscience, **7**(31), 1–17.

Strori, D., Zaar, J., Cooke, M., and Mattys, S. L. (**2018**). "Sound specificity effects in spoken word recognition: The effect of integrality between words and sounds," Atten. Percept. Psychophys. **80**(1), 222–241.

Shen, J., and Souza, P. (**2017**). "The effect of dynamic pitch on speech recognition in temporally modulated noise," J. Speech Lang. and Hear. Res. **60**, 2725–2739.

Tamati, T. N., Gilbert, J. L., and Pisoni, D. B. (**2013**). "Some factors underlying individual differences in speech recognition on PRESTO: A first report," Journal of American Academy of Audiology **24**(7), 616–634.

Tamati, T. N., and Pisoni, D. B. (**2014**). "Non-native speech recognition in adverse listening conditions," Journal of American Academy of Audiology **25**(9), 869–892.

United States Census Bureau. (**2017**). "Selected characteristics of the total and native populations in the Unites States 2017 American Community Survey 1-Year Estimates" [Data file]. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_S0601&prodType=table (last viewed on 25 February 2020).

van Wijngaarden, S. J., Steeneken, H. J., and Houtgast, T. (**2002**). "Quantifying the intelligibility of speech in noise for non-native talkers," J. Acoust. Soc. Am. **112**, 3004–3013.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and 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**, 510–540.

Walker, A., and Campbell-Kibler, K. (**2015**). "Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task," Front. Psychol. **6**, 546.

Wester, M., Lecumberri, M. L. G., and Cooke, M. (**2014**). "DIAPIX-FL: A symmetric corpus of conversations in first and second languages," in Proceedings of Interspeech, Singapore, 509–513.

Wilson, E. O., and Spaulding, T. J. (**2010**). "Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English," J. Speech Lang. Hear. Res. **53**, 1543–1554.

Wingfield, A. (**1975**). "The intonation-syntax interaction: Prosodic features in perceptual processing of sentences," in *Structure and Process in Speech Perception. Communication and Cybernetics, Vol. 11*, edited by A. Cohen and S. G. Nooteboom (Springer, Berlin-Heidelberg).

Yu, A. C. L., and Lee, H. (**2014**). "The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study," J. Acoust. Soc. Am. **136**, 382–388.