Spectral integration of English speech for non-native English speakers

Lauren Calandruccio^{1,a)} and Emily Buss²

¹Department of Psychological Sciences, Case Western Reserve University, 11635 Euclid Avenue, Cleveland, Ohio 44106, USA

²Department of Otolaryngology/Head and Neck Surgery, 170 Manning Drive, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

(Received 15 March 2017; revised 7 September 2017; accepted 8 September 2017; published online 26 September 2017)

When listening in noisy environments, good speech perception often relies on the ability to integrate cues distributed across disparate frequency regions. The present study evaluated this ability in non-native speakers of English. Native English-speaking and native Mandarin-speaking listeners who acquired English as their second language participated. English sentence recognition was evaluated in a two-stage procedure. First, the bandwidth associated with $\sim 15\%$ correct was determined for a band centered on 500 Hz and a band centered at 2500 Hz. Performance was then evaluated for each band alone and both bands combined. Data indicated that non-natives needed significantly wider bandwidths than natives to achieve comparable performance with just the low or just the high band alone. Further, even when provided with wider bandwidth within each frequency region, non-natives were worse than natives at integrating information across bands. These data support the idea that greater bandwidth requirements and a reduced ability to integrate speech cues distributed across frequency may play an important role in the greater difficulty non-natives often experience when listening to English speech in noisy environments. © 2017 Acoustical Society of America.

[<http://dx.doi.org/10.1121/1.5003933>]

[JFL]

Pages: 1646–1654

I. INTRODUCTION

In the mid-1960s, a serendipitous finding by Moe Bergman and colleagues revealed that participants who had learned English as a second language were at a significant disadvantage recognizing English speech in degraded listening conditions compared to their native English-speaking counterparts. These researchers did not set out to test non-native English speech recognition; rather, they had recruited a cohort of listeners from one specific housing development in the greater New York City area to participate in experiments on masked speech perception. After analyzing their data, they found that this cohort had significantly poorer scores in several degraded listening conditions than listeners recruited from other areas. The only notable difference between the participants from this housing development and other participants was that they spoke accented English, albeit with a high degree of proficiency. These participants had lived in the United States (U.S.) an average of over 50 years, and their English speech recognition in quiet was nearly perfect. However, they had significant difficulties recognizing speech with reverberation, increased presentation rate, low- and high-pass filtering, or temporal interruptions (Bergman, 1980). Since these discoveries were made, there have been numerous reports of the difficulty even highly proficient non-native speakers of a language have recognizing noisy speech in their non-native language (e.g., Mayo *et al.*, 1997; Rogers *et al.*, 2006; Shi, 2010). The current study evaluates what role speech bandwidth requirements

and the ability to integrate speech cues across frequency play in the greater susceptibility to masking in non-natives.

When trying to recognize noisy speech, all listeners are faced with two main challenges. First, listeners need to stream the auditory signal of interest as one object, separating it from the competing sounds (Bregman, 1990). Second, listeners need to contend with only having access to portions of the auditory target, due to energetic masking effectively eliminating regions in which the energy of the competing sounds dominate the signal-to-noise ratio (SNR; e.g., Brungart *et al.*, 2006; Howard-Jones and Rosen, 1993). Due to the redundant nature of speech, native speakers of the target language with normal hearing are quite good at recognizing speech with significant portions of the spectrum missing (e.g., Cooke, 2006; Warren *et al.*, 1995), as they are able to combine cues across both frequency and time (Cherry and Wiley, 1967; Miller and Licklider, 1950). For sentence recognition, normal-hearing listeners who are native speakers of the target language can maintain close to 90% accurate word recognition with more than 70% of spectral components missing (Greenberg *et al.*, 1998).

Warren *et al.* (1995) provided a striking example of spectral integration. In that study, speech was filtered into narrow bands, which were then presented alone or in combination. For bands that were 1/20 of an octave wide, speech perception was very poor for a single band at either 370 or 6000 Hz, with scores of $<1\%$ and 10% correct, respectively. However, when these two bands were presented together, performance rose to 28% correct. The benefit of two bands presented together was defined as a *synergistic integration* of speech cues. Synergistic integration is most pronounced

^{a)}Electronic mail: Lauren.Calandruccio@case.edu

when the speech cues available to the listener are complementary rather than redundant (Grant *et al.*, 1991; Warren *et al.*, 2005).¹

Whereas adults are able to recognize speech in their native language based on spectrally sparse information, children require a broader spectrum. Eisenberg *et al.* (2000) evaluated speech perception using a noise-vocoded stimulus and found that young children (ages 5–7 yr) required more spectral bands than older children (ages 10–12 yr) and adult listeners. These results appeared to be due to a combination of factors: younger children appeared to need more spectral detail to recognize phonetic information, and once these cues were provided they were less able to take advantage of the sentence context.

A subsequent study by Mlot *et al.* (2010) provided further support for the idea that child age is negatively correlated with the bandwidth needed to achieve a specified performance level. In the first stage of testing, Mlot *et al.* evaluated performance for a band of speech centered on either 500 or 2500 Hz, adaptively varying filter bandwidth to estimate the criterion bandwidth associated with 15%–20% correct (Hall *et al.*, 2008; Noordhoek *et al.*, 1999). Children needed wider bandwidths compared to adult listeners to attain similar performance values, and child age was negatively correlated with the criterion speech bandwidth. When criterion bandwidths were used for individual listeners, performance for each band alone was ~20% correct, but performance when both bands were presented together was over 80% correct for both children and adults. In other words, spectral integration was comparable across groups after normalizing performance for each band alone.

The present experiment used the methods of Hall *et al.* (2008) and Mlot *et al.* (2010) to determine whether adult non-native speakers of English are able to effectively integrate spectrally distant information for a sentence recognition task. We hypothesized that similar to the results observed for children, adult non-native speakers would require broader bandwidths than adult native speakers to achieve similar performance due to their linguistic inexperience with the English language. To test that prediction, we compared results obtained with non-native speakers of English to a group of native English speakers. We also anticipated that non-natives would be able to combine cues across bands that are separated in frequency; however, it was unclear whether they would be able to benefit to a similar degree as their native English-speaking counterparts. If non-native speakers require broader criterion bandwidths to recognize sentences and are less good at integrating information across bands, perhaps this could account for some of the increased difficulty non-native speaking listeners tend to have while recognizing speech in noisy environments.

II. METHODS

A. Listeners

Data are reported for 19 native Mandarin speakers (4 males, 15 females; 20–36 years of age, mean of 26 yr) and 10 native English speakers (2 males, 8 females; 19–28 years of age, mean of 22 yr). One additional Mandarin

speaker was recruited but did not finish the experiment due to scheduling limitations. All listeners had normal audiometric thresholds (≤ 20 dB hearing level) at octave frequencies between 250 and 8000 Hz, bilaterally, as tested using standard audiometric procedures (ASHA, 2005). All listeners provided informed consent, following procedures approved by the Institutional Review Board at the University of North Carolina at Chapel Hill, and were paid for their participation.

All non-native listeners completed a linguistic and demographic questionnaire developed by the Linguistics Department at Northwestern University (Chan, 2012). This questionnaire assesses English language proficiency focusing on five areas: language status, language stability, language competency, language history, and demand for language usage (see von Hapsburg and Peña, 2002). Table I provides listener specific information regarding linguistic history for the non-native English-speaking listeners. All 19 of them reported higher proficiency and competency in Mandarin than English. They rated their reading, writing, speaking, and listening ability in Mandarin as “perfect” (10) to “excellent” (9), and in English as “slightly less than adequate” (4) to “very good” (8). All participants reported Mandarin to be their dominant language, and indicated stability and competency in both Mandarin and English. All participants indicated they were born in Mainland China, and their *language history* indicated the average age of English acquisition was ~9 years old, with one subject indicating English language exposure by the age of three years old. Demand of language usage indices indicated that 18 non-native listeners reported using both languages daily, whereas one reported using only English daily. On average, these listeners reported using English 53.7% of the time. It was reported that Mandarin is mainly spoken with family and friends, while English is primarily used with classmates, professors, friends, and co-workers.

In addition, all non-native English speakers completed a Versant English test (Pearson, San Antonio, TX). This is an automated speech recognition test completed over the phone that is used to assess non-native English speakers’ English proficiency. The Versant English test is often used by businesses when hiring potential employees who are non-native English speakers to help ensure efficient English communication upon employment. It is also used at major U.S. universities as an assessment for English language skills to ensure an adequate level of English communication for potential teaching and graduate assistants who are non-native English speakers. This test provides the examiner with scores for sentence mastery, fluency, vocabulary, and pronunciation, and an overall English assessment score ranging between 20 and 80 points. Versant scores are positively correlated with non-native English speakers’ ability to understand English speech in noise (Calandruccio *et al.*, 2014; Rimikis *et al.*, 2013). To provide a sense of privacy, listeners were seated in a double-walled, sound isolated room while performing the test. Individual Versant scores for all non-native English speaking listeners are shown in Table I, along with listeners’ linguistic history information.

TABLE I. English language history is provided for each participant. Asterisks indicate an English-speaking country other than the U.S. Also, Versant English test scores for all 19 non-native English-speaking listeners are provided. An overall score is shown for each listener, as well as sub-category scores for sentence mastery, fluency, vocabulary, and pronunciation. Scores range between 20 and 80 points.

ID	Language questionnaire				Versant English test				
	Age of English acquisition (years old)	Secondary Education instruction in English (percent)	Age moved to English speaking country (years old)	Years of English experience	Overall Versant	Sentence mastery	Vocabulary	Fluency	Pronunciation
NN01	10	10	32	26	67	70	63	67	67
NN02	11	20	22	11	49	54	50	48	40
NN03	10	0	21	12	59	56	63	61	55
NN04	3	0	29	20	51	61	56	48	38
NN05	15	25	15	8	80	80	77	80	80
NN06	8	40	16	12	65	63	68	66	62
NN07	6	10	22	19	63	56	64	67	68
NN08	10	15	34	24	39	44	42	35	35
NN09	11	20	36	25	64	59	72	62	65
NN10	9	25	18	11	60	68	75	52	43
NN11	12	0	21*	16	50	59	59	43	39
NN12	10	10	24	16	37	48	38	30	33
NN13	7	10	19	13	63	68	69	60	57
NN14	13	15	30	18	46	58	53	36	39
NN15	7	0	18*	15	80	80	80	76	80
NN16	7	10	19	15	59	64	61	56	52
NN17	11	15	—	22	44	54	48	34	38
NN18	9	10	17*	14	56	69	61	49	42
NN19	5	30	20*	18	53	61	73	40	39

B. Speech materials

Stimuli were Bamford-Kowal-Bench (BKB) sentences (Bench *et al.*, 1979) spoken by a male talker with no remarkable regional dialect (Auditec, Inc., St. Louis, MO). The BKB materials include 21 lists of 16 sentences. Each sentence has either three or four keywords, and each list has 50 keywords total. Keywords are either monosyllabic, disyllabic, or trisyllabic (average within list keyword syllable count = 66%, 32%, and 2%, respectively). The sentences were developed based on a lexicon of speech spoken by children with hearing loss and are simple in their grammar and vocabulary, making these materials a suitable choice to use with non-native English speakers. An example of a BKB sentence is, “The clown has a funny face.” Keywords for this example sentence are underlined.

C. Procedure

Listeners were comfortably seated in a single-walled (IAC, North Aurora, IL) isolated room. Stimuli were presented monaurally to the left ear via Sennheiser HD 25–1 II (Wedemark, Germany) earphones. Listeners were instructed that they would be listening to a male talker and their task was to repeat back every word the male talker said. Listeners were encouraged to guess and/or repeat back what they heard even if they thought the sentence did not make sense. Listeners were encouraged to speak loudly, slowly, and to enunciate to minimize examiner difficulty understanding listeners’ accented speech. No feedback was provided to the listener.

Following the methods outlined in Hall *et al.* (2008) and Mlot *et al.* (2010), the first stage of testing was to determine criterion bandwidth for a single band arithmetically centered

on 500 or 2500 Hz; these are referred to below as the low and high bands, respectively. These frequencies were chosen based on a compromise of three considerations: (1) speech cues in these frequency regions tend to be informative for sentences (Pavlovic, 1994), (2) the two bands are sufficiently well separated in frequency to accommodate a range of bandwidths while maintaining spectral separation (Grant *et al.*, 2007), and (3) previous studies have used these frequencies (Hall *et al.*, 2008; Mlot *et al.*, 2010). Filtering was performed using finite impulse response (FIR) filters with 12-Hz resolution. Prior to filtering, the root-mean-square (RMS) speech level was 76 dB sound pressure level (SPL). The order of presentation was randomized: approximately half of the listeners completed the 500-Hz condition before the 2500-Hz condition, and the other half completed testing in the opposite order. The first stage of testing resulted in criterion speech bandwidths for each listener at each of the two frequency bands. The *criterion speech bandwidth* refers to individually tailored frequency bandwidths associated with a common percent correct performance level (Noordhoek *et al.*, 1999).

For non-native speakers of English, the criterion speech bandwidth was determined adaptively. The width of the band was decreased by a factor of 1.21 when one or more keywords were reported correctly, and increased following two sentences in which all keywords were reported incorrectly. Each track continued until eight reversals were obtained, and the final estimate of criterion bandwidth was the geometric mean of the bandwidth at the last six-track reversals. Based on previous data, this adaptive rule was expected to result in a bandwidth associated with ~15% correct recognition. Three or four threshold estimates were

collected for each listener at each frequency, and the final threshold estimate was the geometric mean of all thresholds obtained.

Whereas the adaptive method estimated the criterion speech bandwidth associated with 15% correct in non-native listeners, pilot data obtained from native English speakers using the adaptive bandwidth estimation procedure yielded bandwidths that were associated with a higher percent correct (~22%). This difference in scores could reflect differences between groups in the ability to use sentence context to improve performance. This possibility is evaluated in Sec. III. The primary goal of the first stage of testing was to determine bandwidths associated with matched performance across groups, so an alternative approach for determining the criterion speech bandwidth in the native English speakers was pursued. For native English speakers, the criterion bandwidth was obtained by trial and error, with each listener providing data in blocks of 16 sentences (1 list) at a fixed bandwidth. Using this approach, each listener completed between three and eight blocks in each condition before a bandwidth associated with ~15% correct was identified.

In the second stage of testing, performance was evaluated for the low band, the high band, and the two bands (both low and high) combined, using individualized criterion speech bandwidths. Each estimate was based on data from three sentence lists. The differences in procedures used to estimate performance for native and non-native listeners in the first stage of testing could theoretically have introduced differences in the precision of threshold estimates. However, the pattern of results in the single-band conditions of stage 2 of testing indicate comparable performance in the two groups. Percent correct scores (total keywords correct/150 total keywords presented) were transformed to rationalized arcsine units (RAUs; Studebaker, 1985) prior to statistical analysis to normalize variance across conditions. This transformation was especially important for these data, as many listeners' criterion speech bandwidths in the independent band conditions (i.e., the low frequency band or the high frequency band alone) resulted in performance <20% correct.

Sentence lists were randomly selected for each listener and each condition, and care was taken that listeners never heard the same sentences more than once. All data were collected in a single test session with breaks.

III. RESULTS

Thresholds from the first stage of testing were normalized by dividing the criterion bandwidth (in Hz) by the center frequency of the band. These normalized criterion bandwidths are shown in Fig. 1 as a function of band center frequency. Open and filled circles show data for individual non-native and native English speakers, respectively. Boxes span the 25th–75th percentiles, horizontal lines indicate the medians of each distribution, and vertical lines mark the 10th and 90th percentiles.

The most striking trend in these data is the wider normalized criterion bandwidth for non-native than native English speakers. The mean normalized criterion bandwidths differed across groups by a factor of 1.24 for the low band

FIG. 1. Distributions of individual listeners' normalized criterion bandwidths for the low band (centered on 500 Hz) and the high band (centered on 2500 Hz). Open circles show results for individual non-native English-speaking listeners, and filled circles show results for individual native English-speaking listeners. Boxes span the 25th–75th percentiles, horizontal lines indicate the median, and vertical lines indicate the 10th–90th percentiles.

and a factor of 1.32 for the high band. These data were evaluated using a random-intercepts mixed-model regression analysis (nlme, R) with subject as a random variable. The model tested the fixed effect of band frequency (low and high), the fixed effect of native-language group (non-native and native), and the interaction of band frequency and native-language group. A log transform was applied to individual values of normalized criterion bandwidth prior to this analysis. As reported in Table II, there was no main effect of band frequency ($p = 0.19$), and no significant interaction between band frequency and group ($p = 0.44$). There was a significant effect of group ($p < 0.001$). Criterion bandwidths were significantly wider for non-native than native speakers for both band conditions (mean low frequency normalized bandwidths = 0.44 and 0.35, and mean high frequency normalized bandwidths = 0.43 and 0.32, for non-natives and natives, respectively).

The performance measured in the second stage of testing is shown in Fig. 2. Individual listeners' percent correct scores for the low band alone, the high band alone, and both bands presented together were transformed to RAUs. Plotting conventions follow those of Fig. 1. As observed previously, performance was relatively poor for each band alone and relatively good when both bands were presented together (Mlot *et al.*, 2010). Individualized criterion bandwidths supported similar performance for the low and high band alone conditions for both groups of listeners. However, scores for non-native listeners tended to be lower than those for native English speakers when the low band and the high

TABLE II. Mixed-model regression results for group and band frequency effects on criterion bandwidths. df = degrees of freedom.

	Estimate (β)	Standard error (SE)	df	t -value	p -value
Intercept	0.32	0.023	27	13.4	<0.001
Criterion speech bandwidth (low)	0.04	0.028	27	1.3	0.19
Group (non-native)	0.11	0.029	27	3.7	<0.001
Criterion speech bandwidth (low) $\times$ group (non-native)	-0.03	0.034	27	-0.8	0.44

FIG. 2. Distributions of individual listeners' performance in RAUs for the low band presented alone, the high band presented alone, or the combination of the low band plus the high band. Plotting conventions follow those of Fig. 1.

band were presented together. These trends were evaluated using a random intercepts mixed-model regression analysis with subject as a random factor, the fixed effect of band condition (low, high, and low + high), the fixed effect of native-language group (non-native and native), and the interaction of the two fixed effects. Here, we adjusted for unequal variances across groups and band conditions. The results of the analysis are shown in Table III. There was a significant interaction between band and group ($p < 0.001$). Thus, the difference in performance between native and non-native speakers depended on band condition. For both groups, *post hoc* Tukey testing indicated that performance was not significantly different for the low and high bands alone, but performance was significantly better when both bands were presented together compared to either band alone. Further, native speakers of English performed significantly better than non-natives when both the low and high bands were presented together (*estimated difference* = 15.3, Standard Error = 3.5, *t*-statistic = 4.3, *p*-value = 0.002); no significant group differences were observed for either of the two band alone conditions. See supplemental materials for all pairwise contrasts results.²

Whereas performance with each band alone was relatively poor, performance for two bands together was quite good. For non-native speakers of English, mean performance was 14.9% for each band alone and 65.1% for both bands together. For native speakers of English, mean performance was 12.5% for each band alone and 79.7% for both bands together. If the low and high bands provided independent information, then the percent correct when both bands are present (PC_{both}) could be predicted based on the percent

correct observed when just the low or high band is present (PC_{low} and PC_{high} , respectively) as follows:

$$PC_{\text{both}} = 100[1 - (1 - PC_{\text{low}}/100)(1 - PC_{\text{high}}/100)].$$

Improvement in excess of this expected performance indicates that the information provided by each band is not independent, but rather combines synergistically. The performance observed when both bands were present exceeded the percent correct expected based on a combination of independent information for all individuals, and the difference between observed and expected performance was larger for the native than the non-native group. The observed-minus-predicted difference is plotted for individual listeners in Fig. 3, following the plotting conventions of Fig. 1. The mean difference was 37.5% for the non-natives and 56.3% for the natives, indicating significantly greater synergistic integration for the native-speaking listeners than the non-native speakers ($t_{1,27} = 6.47$, $p < 0.001$).³

A. Performance and English language experience

Data reported in Mlot *et al.* (2010) indicated a significant correlation between a child's criterion bandwidth for the low and high frequency bands; children requiring a wider bandwidth for the low frequency band also tended to require a wider bandwidth for the high frequency. In the first stage of testing, which estimated individual criterion bandwidths, no significant association between a non-native listener's criterion bandwidth for the low and high frequency bands was observed, as evaluated using a two-tailed Pearson's product-moment correlation test ($r = 0.28$, $p = 0.249$). The association between *age of English acquisition* and criterion bandwidth for the low and high bands was also evaluated using one-tailed Pearson's product-moment correlations and Bonferroni adjusted alpha levels of 0.025 per test (0.05/2). There was an association between age of English acquisition and criterion bandwidth for the high band ($r = 0.56$, $p = 0.012$); participants who acquired English later in life, generally, required a wider high frequency band to achieve

FIG. 3. The difference between observed and expected performance for both bands presented together is shown for the two groups of listeners. Plotting conventions follow those of Fig. 1.

TABLE III. Mixed-model regression results for group (native, non-native) and band (low, high, low + high) effects on RAU performance.

	Estimate (β)	Standard error (SE)	df	<i>t</i> -value	<i>p</i> -value
Intercept	80.03	2.7	54	29.7	<0.001
Band (high)	-69.4	2.6	54	-27.0	<0.001
Band (low)	-69.7	2.5	54	-27.7	<0.001
Group (non-native)	-15.3	3.5	27	-4.3	<0.001
Band (high) $\times$ group (non-native)	18.1	3.6	54	5.0	<0.001
Band (low) $\times$ group (non-native)	18.5	3.4	54	5.4	<0.001

criterion performance. There was no association between age of acquisition and criterion bandwidth for the low band ($r = 0.27$, $p = 0.273$).

Overall Versant scores have been associated with non-native English sentence recognition performance when listening in noise (Rimikis *et al.*, 2013). In the first stage of testing, there was no association between normalized criterion bandwidth for either band and overall Versant score, evaluated using a two-tailed Pearson's product-moment correlation and Bonferroni adjusted alpha levels of 0.025 per test (0.05/2; $r = 0.01$, $p = 0.997$ and $r = -0.23$, $p = 0.353$, for the low and high band, respectively). In contrast, the overall Versant score was significantly correlated with "observed-predicted" scores when both bands were presented together ($r = 0.48$, $p = 0.038$). A two-tailed Pearson's product-moment correlation was used to test this relationship using an alpha level of 0.05.

Given the pilot data indicating that non-native listeners may be using sentence context less effectively than native English-speaking listeners, an analysis of context effects was undertaken using data from the second stage of testing. One way of quantifying context effects is based on the j -factor, which is calculated as the log of the probability of recognizing the whole sentence, divided by the log of the probability of recognizing each word within the sentence (Boothroyd and Nittrouer, 1988). For the condition in which both the low and high bands were present, the j -factor was not significantly different for native and non-native listeners ($t_{27} = 0.754$, $p = 0.457$), with mean values of $j = 2.7$ and $j = 2.8$, respectively. This result suggests a comparable ability to use context in the two-band condition.

An analysis based on the j -factor could not be performed for conditions with the low band alone and high band alone, due to the very small probability of getting all keywords in a sentence correct in these conditions. However, there was evidence of a group effect in the pattern of keywords correct per sentence. If a listener is able to use context, then getting one word correct would tend to increase the likelihood of getting other words correct. Omitting trials when all words were incorrect, non-native listeners got three or more keywords in a sentence correct an average of 6% and 9% of the time (low and high band conditions, respectively). For native English speakers, those values were 13% and 14%.⁴ That is, when groups were matched for overall performance, the correct responses provided by native English-speaking participants were more likely to co-occur within sentences, whereas those of non-native participants were more distributed across sentences. This result is consistent with non-natives relying less on sentence context than native English speakers. The significance of this group difference was confirmed using a random intercepts mixed-model regression analysis after percent correct in each condition was transformed into RAU. Subject was included as a random factor, and the model indicated a significant effect of group (estimate = -1.20 , Standard Error = 0.43 , $df = 27$, t -value = -2.76 , $p = 0.010$). No significant effect of band (low, high) or group-by-band interaction was found.

IV. DISCUSSION

Sentence recognition was measured for non-native and native English-speaking listeners in two stages of testing.

The first stage of testing assessed the individualized criterion bandwidth needed to achieve $\sim 15\%$ performance at either 500 or 2500 Hz. Results indicated that non-native speakers of English needed a significantly *wider* bandwidth for both the low and high frequency bands compared to native English-speaking adults. In the second stage of testing, it was revealed that non-native speakers were less adept than native speakers at combining speech cues across the two distinct bands to support sentence recognition.

A. Effects of linguistic inexperience

The two stages of testing reported above for non-native speakers of English were based on the methods and stimuli used in Mlot *et al.* (2010). However, direct comparisons between child listeners and non-native English-speaking adults should be made with caution. The children tested by Mlot *et al.* spanned a relatively wide range of ages (6–14 years old), with the goal of observing developmental effects of spectral integration. In the current study, care was taken to maximize homogeneity within our population of non-native English speakers to increase the study power with respect to a native/non-native comparison. That being said, a few general observations can be made.

Similar to children, adult non-native speakers of English needed significantly wider bandwidth compared to native English-speaking adults. Mlot *et al.* (2010) reported that children also required significantly wider bandwidths than native English-speaking adults to achieve similar performance for sentences bandpass filtered into either a low or high frequency band. In addition, the criterion bandwidth was negatively correlated with child age for bands in both frequency regions. One possible explanation for this result is related to linguistic experience; the greater experience of older children allows them to recognize speech based on more spectrally limited cues than younger children. If that is the case, then one might expect that criterion bandwidths in non-native listeners would also be associated with language experience and/or proficiency.

For non-native speakers of English, data on the association between criterion bandwidth and English language experience and proficiency were mixed. The overall Versant score did not predict the criterion bandwidth needed for the non-native listeners for either frequency band. For the high band alone condition, there was an association between age of acquisition and the criterion bandwidth, but no relationship was observed for the low band. One explanation for these results is that greater English language exposure allows non-native speakers of English (specifically native speakers of Mandarin) to better utilize high frequency English speech cues. However, it is also possible that homogeneity within the group of non-native English speakers in the present study prevented us from seeing a relationship between performance and English language ability. There was limited diversity in the age of English acquisition, all of these listeners learned English in a similar manner (predominately for higher education), and all were either in their late teens or adults when they immigrated to the U.S. Nevertheless, the main conclusion from these data is that non-native English

speakers require wider criterion bandwidths to achieve similar performance levels compared to native speakers of English. This result could be different for a group of simultaneous-bilingual speakers or for a group of early sequential-bilingual speakers who learned—and became fluent with the English language—at a much younger age.

Although both children (Mlot *et al.*, 2010) and non-native adults (present study) needed wider bandwidths than native English-speaking adults in the first stage of testing, only the children benefitted from the presentation of both bands to a comparable extent as adult native English-speaking listeners. In the present work, a significant correlation was observed for Versant score and performance with both the low and high bands. Further, the overall Versant score was also significantly correlated with observed-expected performance. These results are in agreement with the idea that greater language proficiency is associated with greater ability to integrate cues across bands that are separated in frequency. However, this idea needs to be further explored since the data for children (as young as 6 years old) reported in Mlot *et al.* did not show a deficit with respect to spectral integration. The results of this experiment taken together with those reported in Mlot *et al.* suggest that relying on sparse cues (e.g., low alone or high alone conditions) and integrating sparse cues across distant frequency bands (e.g., low + high condition) may not necessarily rely on the same underlying speech perception skills.

B. Use of sentence context

An interesting difference between the data obtained for native English-speaking adults and children reported in Mlot *et al.* (2010) and the data reported here for non-native and native English-speaking adults has to do with performance when a single band of speech is present. For both the native English-speaking adults and children presented in Mlot *et al.*, performance with a single band was ~20% correct for the low band (children: 21.4%, adults: 23.7%) and for the high band (children: 23.1%, adults: 19.7%). In contrast, average scores for the non-natives tested in the present experiment were 14.9% for the low as well as the high band. One possible explanation for the non-native listeners' relatively poor performance in single-band conditions, with the adaptively estimated criterion speech bandwidth, could be related to how listeners with different linguistic experience make use of sentence context.

Previous studies have shown that non-natives are less adept than natives at taking advantage of sentence context to aid recognition (Bradlow and Alexander, 2007; Golestani *et al.*, 2009; Mattys *et al.*, 2010; Shi, 2014). For example, Shi and Koenig (2016) reported that non-natives (non-English dominant) listeners made less effective use of both syntax and semantic cues relative to their native peers. This means that partial information (e.g., one word heard correctly) improved a native listener's ability to recognize other words within a sentence more than a non-native listener. The adaptive bandwidth procedure reduced the stimulus bandwidth after the listener responded correctly to one or more words. If native listeners were better than non-native

listeners at using context, they would be less likely than non-natives to get just one word correct. That is, sparse information could allow them to get more than one word correct. The stepping rule in stage 1 is not sensitive to the use of this additional context information, but the use of context would tend to improve performance in stage 2 of testing. In the current data set, in the second stage of testing, it was observed that native English speakers were more likely to get all words in a sentence correct than non-native speakers (see final analysis in Sec. III). This was evident even though differences in overall performance between the two groups was normalized with the use of individualized criterion bandwidths.

C. Glimpsing degraded speech—Cue requirements

It has been suggested by many researchers that listeners use “glimpses” of speech signals to understand speech in noisy environments (see Cooke, 2006, for an overview). Assmann and Summerfield (2004) describe the process associated with glimpsing as requiring two steps: (1) tracking glimpsed cues, and (2) integrating glimpsed information. Non-natives appear to segregate sound streams as well as natives, as they derive a similar benefit from spatially separating the target and masker compared to native listeners (Ezzatian *et al.*, 2010; von Hapsburg *et al.*, 2004).⁵ Our data suggest that once non-natives are provided with speech cues in two distant frequency regions, they are rather good at integrating the distant spectral information into one coherent message, although they are not as efficient as a native-speaker of the language. Therefore, it appears that poorer sentence recognition in noise that is often observed for non-native English speakers (Mayo *et al.*, 1997; Shi, 2010) could be due, at least in part, to a reduced ability to integrate information across frequency. Further, due to the greater bandwidth needed in the single band alone conditions, it is also likely that for non-native speakers, a third factor in addition to segregation and integration is critical—more stringent cue requirements.

If natives can recognize speech based on fewer cues, perhaps due to greater familiarity with the language, then non-natives may require more cues, consistent with the broader criterion bandwidth in the first stage of testing. Whereas native and non-native speakers of English appeared to use context to a comparable degree in the two-band condition, the non-natives appeared to benefit from context less than native English speakers when listening to the low or high band alone. However, there are numerous reports in the literature indicating that listeners use different strategies, or rely on different types of cues, depending upon the listening situation (see Mattys, 2004; Hillenbrand, 2003; Calandruccio *et al.*, 2010, for examples). It is possible that in severely band limited listening scenarios, non-natives are more reliant on acoustic cues than contextual cues. However, when the listening scenario is more favorable, contextual cues may be easier for this group to access. Support for this idea can be found in the literature. There is evidence that non-natives can take advantage of sentence context cues when the sentence is spoken clearly, as opposed

to in a conversational speaking style (Bradlow and Alexander, 2007). Further, there are also data supporting the idea that the ability of non-native listeners to use sentence context cues decreases as the SNR becomes more unfavorable (Shi, 2014; Shi and Koenig, 2016). Further exploration is needed to better understand spectral integration of cues in relation to the ability to use context for non-native listeners. Experiments using semantically anomalous sentences may provide insight into this relationship.

V. CONCLUSIONS

Non-native English speakers need wider bandwidths in both low and high frequency regions to perform comparably to their native English-speaking counterparts while recognizing band limited English sentences in quiet. Once sufficient bandwidth is provided to achieve ~15% correct with each band alone, both groups are able to efficiently combine information across bands to improve their overall speech recognition performance; however, non-native speakers benefit less than native speakers from access to two bands. Within the group of non-native English speakers, there was a significant correlation between overall Versant scores and performance with both the low and high bands present. These results indicate that language experience may impact the ability to combine cues across frequency. Greater bandwidth requirements and reduced ability to integrate cues across frequency in non-native English speakers with normal hearing suggests that audibility requirements could be substantially higher in non-native hearing-impaired listeners compared to their native English-speaking peers. More research is needed to better understand the effects of linguistic experience for the designing and fitting of amplification devices for listeners with hearing loss who are using such devices while listening in their second language.

ACKNOWLEDGMENTS

Portions of this work were presented at the American Auditory Society meeting in Scottsdale, AZ, March 2016. Funding support provided by the National Institutes of Health, National Institute of Deafness and Other Communication Disorders (NIDCD) Grant No. R01 DC007391. Sincere thanks to Dr. Lu-Feng Shi and Dr. Ken Grant for helpful suggestions and stimulating conversations while working on this project.

¹Warren *et al.* (2005) defined synergy (+Δ%) based on the proportion of correct responses observed in the dual-band condition and the proportion of correct responses expected as follows: $+Δ\% = 100[(\text{observed}/\text{expected}) - 1]$.

²See supplementary material at <http://dx.doi.org/10.1121/1.5003933> for all pairwise contrasts results for the random intercepts mixed-model regression analysis for group (native, non-native) and band condition (low, high, low + high) effects on RAU performance.

³A significant effect of group is also observed when synergy, as defined by Warren *et al.* (2005), is compared across groups ($t_{1,27} = 5.48, p < 0.001$).

⁴Omitting trials where all words were incorrect, three or more words were reported correctly 73% of the time for native English speakers and 54% of the time for non-natives in the low + high condition. This result was not reported in the main text due to group differences in mean performance. The better performance of the native English speakers in the low + high condition would tend to increase the likelihood of three or more correct

keywords within a sentence relative to results obtained with the non-native listeners. This is not a concern for the low and high band alone conditions, due to the very similar percent correct in the two groups. If anything, the slightly better performance of the non-natives (14.9%) than the natives (12.5%) would tend to introduce a bias in the opposite direction.

⁵The one caveat is that the two published reports on spatial release from masking in natives and non-natives assessed performance at different SNRs (Ezzatian *et al.*, 2010; von Hapsburg *et al.*, 2004) due to poorer performance overall in non-natives. Therefore, while these data indicate comparable benefits of spatial (or perceived spatial) separation between groups, the higher baseline thresholds for the co-located condition in non-natives may limit the amount of observed masking release for the non-native listener group (Bernstein and Grant, 2009).

- American Speech-Language-Hearing Association (ASHA). (2005). "Guidelines for manual pure-tone threshold audiometry" [Guidelines], available at www.asha.org/policy (Last viewed February 20, 2015).
- Assmann, P. F., and Summerfield, Q. (2004). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York), Vol. 14, pp. 231–308.
- Bench, J., Kowal, A., and Bamford, J. (1979). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," *Brit. J. Audiol.* **13**, 108–112.
- Bergman, M. (1980). *Aging and the Perception of Speech* (University Park Press, Baltimore, MD), pp. xiii–173.
- Bernstein, J. G. W., and Grant, K. W. (2009). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **125**(5), 3358–3372.
- Boothroyd, A., and Nittrouer, S. (1988). "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.* **84**(1), 101–114.
- Bradlow, A. R., and Alexander, J. A. (2007). "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners," *J. Acoust. Soc. Am.* **121**(4), 2339–2349.
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (The MIT Press, Cambridge, MA), pp. viii–773.
- Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.* **120**(6), 4007–4018.
- Calandruccio, L., Buss, E., and Hall, J. W. (2014). "Effects of linguistic experience on the ability to benefit from temporal and spectral masker modulation," *J. Acoust. Soc. Am.* **135**(3), 1335–1343.
- Calandruccio, L., Dhar, S., and Bradlow, A. (2010). "Speech-on-speech masking with variable access to the linguistic content of the masker speech," *J. Acoust. Soc. Am.* **128**(2), 860–869.
- Chan, C. L. (2012). "NU-subdb: Northwestern University Subject Database [web application]," Department of Linguistics, Northwestern University.
- Cherry, C., and Wiley, R. (1967). "Speech communication in very noisy environments," *Nature* **214**(5093), 1164.
- Cooke, M. (2006). "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.* **119**(3), 1562–1573.
- Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., and Boothroyd, A. (2000). "Speech recognition with reduced spectral cues as a function of age," *J. Acoust. Soc. Am.* **107**(5), 2704–2710.
- Ezzatian, P., Avivi, M., and Schneider, B. A. (2010). "Do nonnative listeners benefit as much as native listeners from spatial cues that release speech from masking?," *Speech Commun.* **52**(11–12), 919–929.
- Golestani, N., Rosen, S., and Scott, S. (2009). "Native-language benefit for understanding speech-in-noise: The contribution of semantics," *Bilingual, Lang. Cognit.* **12**(3), 385–392.
- Grant, K. W., Braida, L. D., and Renn, R. J. (1991). "Single band amplitude envelope cues as an aid to speechreading," *Q. J. Exp. Psychol.* **43**(3), 621–645.
- Grant, K. W., Tufts, J. B., and Greenberg, S. (2007). "Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals," *J. Acoust. Soc. Am.* **121**, 1164–1176.
- Greenberg, S., Arai, T., and Silipo, R. (1998). "Speech intelligibility derived from exceedingly sparse spectral information," in *International Conference on Spoken Language Processing*, pp. 2–5.

- Hall, J. W., III, Buss, E., and Grose, J. H. (2008). "Spectral integration of speech bands in normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **124**(2), 1105–1115.
- Hillenbrand, J. M. (2003). "Some effects of intonation contour on sentence intelligibility," *J. Acoust. Soc. Am.* **114**, 2338.
- Howard-Jones, P. A., and Rosen, S. (1993). "Unmodulated glimpsing in 'checkerboard' noise," *J. Acoust. Soc. Am.* **93**(5), 2915–2922.
- Mattys, S. L. (2004). "Stress versus coarticulation: Toward an integrated approach to explicit speech segments," *J. Exp. Psychol. Human Percept. Perform.* **30**, 397–408.
- Mattys, S. L., Carroll, L. M., Li, C. K. W., and Chan, S. L. Y. (2010). "Effects of energetic and informational masking on speech segmentation by native and non-native speakers," *Speech Commun.* **52**(11–12), 887–899.
- Mayo, L. H., Florentine, M., and Buus, S. (1997). "Age of second-language acquisition and perception of speech in noise," *J. Speech Hear. Res.* **40**(3), 686–693.
- Miller, G. A., and Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**(2), 167–173.
- Mlot, S., Buss, E., and Hall, J. W., III (2010). "Spectral integration and bandwidth effects on speech recognition in school-aged children and adults," *Ear Hear.* **31**(1), 56–62.
- Noordhoek, I. M., Houtgast, T., and Festen, J. M. (1999). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. I. Normal-hearing listeners," *J. Acoust. Soc. Am.* **105**(5), 2895–2902.
- Pavlovic, C. V. (1994). "Band importance functions for audiological applications," *Ear Hear.* **15**(1), 100–104.
- Rimikis, S., Smiljanica, R., and Calandruccio, L. (2013). "Nonnative English speaker performance on the Basic English Lexicon (BEL) sentences," *J. Speech Hear. Res.* **56**(3), 792–804.
- Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (2006). "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," *Appl. Psycholinguist.* **27**(3), 465–485.
- Shi, L.-F. (2010). "Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition," *J. Speech Hear. Res.* **53**(4), 821–835.
- Shi, L.-F. (2014). "Measuring effectiveness of semantic cues in degraded English sentences in non-native listeners," *Int. J. Audiol.* **53**(1), 30–39.
- Shi, L.-F., and Koenig, L. L. (2016). "Relative weighting of semantic and syntactic cues in native and non-native listeners' recognition of English sentences," *Ear Hear.* **37**(4), 424–433.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- von Hapsburg, D., Champlin, C. A., and Shetty, S. R. (2004). "Reception thresholds for sentences in bilingual (Spanish/English) and monolingual (English) listeners," *J. Am. Acad. Audiol.* **15**(1), 88–98.
- von Hapsburg, D., and Peña, E. D. (2002). "Understanding bilingualism and its impact on speech audiometry," *J. Speech Hear. Res.* **45**(1), 202–213.
- Warren, R. M., Bashford, J. A., and Lenz, P. W. (2005). "Intelligibilities of 1-octave rectangular bands spanning the speech spectrum when heard separately and paired," *J. Acoust. Soc. Am.* **118**(5), 3261–3266.
- Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (1995). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.* **57**(2), 175–182.