Comparing Native Chinese Listeners' Speech Reception Thresholds for Mandarin and English Consonants

Jian Gong*, Yameng Yu, William Belllamy, Feng Wang, Xiaoli Ji and Zhenzhen Yang
School of Foreign Languages, Jiangsu University of Science and Technology, Zhenjiang, China
* E-mail: j.gong@just.edu.cn Tel/Fax: +86-0511-84401945

***Abstract*—The presence of noise can greatly affect listeners' speech perception. Previous studies have demonstrated that non-native listeners' speech perception performance is reduced more than natives' in noise conditions. Most previous studies have focused on the effects of different noise types on non-native speech perception, and using a fixed signal to noise ratio level in different perception tasks. However, the masking effect of noise may be different for individual speech sounds, therefore leaving an incomplete picture of non-native speech perception in noise conditions. The current study applies an adaptive procedure to dynamically adjust the signal to noise ratio to measure listeners' Speech Reception Threshold (SRT) in noise conditions. More specifically, a group of native Chinese listeners' SRTs for Mandarin and English consonants in Speech Shaped Noise were measured and compared. The results showed that Chinese listeners' mean SRT for Mandarin consonants was 3.6dB lower than that for English consonants, indicating a general native language advantage. However, detailed analysis has revealed the mean SRT for the 5 most noise-tolerable consonants in Mandarin was 2.6dB higher than that in English. This result suggests that non-native speech perception in noise conditions may not always be more difficult than native ones. The acoustic features of different sounds could affect their intelligibility in noise conditions.**

I. Introduction

The presence of noise can greatly affect listeners' speech perception. Previous studies have demonstrated that non-native listeners' speech perception performance is reduced more than natives' in noise conditions [1], [2], [3]. Even for high level L2 learners or bilinguals, performance dropped significantly in noise conditions despite having native-like performance in quiet conditions [4], [5]. This native advantage may be due to the fact that native speakers can better use context and other perceptual cues in noise, because they have larger vocabulary, higher grammar and syntax level, and more experience in speech perception in adverse environments than L2 learners [6]. Several studies have demonstrated that the native advantage increases with the increase of noise level, particularly in sentence perception. However, on the level of phoneme perception, contradictory results have been obtained [7], [8], [9]. In a large-scale study involving 8 groups of different L1 listeners, a "language-independent" processing was evident in acoustic and auditory considerations alongside the L1 influence, playing an important role in English consonant perception in noise [10]. This may partly explain the various degrees of native advantage observed in previous studies.

Fixed signal-to-noise ratio (SNR) levels are normally used in speech perception in noise experiments. However, this may raise problems when the purpose of the experiment is to quantify listeners' ability to perceive individual phonemes. Previous studies have demonstrated that to reach equal intelligibility, a large range of SNRs is required for different consonants [11]. Therefore using several fixed SNRs for all consonants may not truthfully reflect the intelligibility for some consonants. Speech Reception Threshold (SRT) is commonly used to measure speech intelligibility in noise. However, most studies employing the SRT procedure are about native speech perception, with few trying to use it for non-native speech perception. More recently, researchers employed SRT procedures to investigate native and non-native sentence perception in noise [12] . However, to the authors knowledge, no study has ever employed the SRT procedure to compare native and non-native consonant perception in noise.

The purpose of the current study is to compare native Chinese listeners' consonant perception in both native Mandarin and non-native English in noise. The SRT procedure is followed to dynamically change the SNRs for different consonants. In this way, the listeners' ability to identify individual consonants in Mandarin and English can be obtained and compared. Several research questions were of interest: (1) Do listeners perform better in their native or non-native language? (2) What are the general SRT patterns for native and non-native consonants? (3) What phonetic-articulation or acoustic feature plays the most important role in consonant perception in noise? A test of English consonant identification in quiet conditions was also included as a reference.

II. Methods

A. Listeners

A group of 40 native Chinese listeners, including 19 males and 21 females, participated in the current study. These listeners were undergraduate and postgraduate students from Jiangsu University of Science and Technology, ranging in age from 19 to 30 years ($M = 23$). No listener had reported hearing or language problems, and all the listeners had passed a hearing test with pure-tone thresholds $\leq$ 15dB HL at octave intervals between 250 and 8000Hz [13]. Listeners were all from Jianghuai Mandarin dialect spoken region (central-east China) and had certification in level II grade B or above in

the National Proficiency Test of Putonghua (Mandarin). These listeners were studying various courses in university, and most of them had passed the College English Test Band 6 (CET-6). Listeners were paid for their participation.

B. Stimuli

The English consonant stimuli used in the current study were nonsense vowel-consonant-vowel (VCV) tokens derived from the Interspeech 2008 Consonant Challenge corpus [14]. The vowel contexts for each VCV token in this corpus were the 9 combinations of the 3 vowels /æ, i, u/ in initial and final positions. Two stress types, namely front stress and end stress, were both recorded for each VCV context in this corpus. A subset of the corpus produced by 4 male and 4 female speakers, containing 23 British English consonants (/p, b, t, d, k, g, tʃ, ʤ, f, v, θ, ð, s, z, ʃ, ʒ, h, m, n, l, r, j, w/, [15]) were used in the English consonant identification in quiet and SRT tests. Similar VCV tokens from a Chinese corpus collected in a previous study [16] were used in the Mandarin consonant SRT test. The Chinese VCV tokens were produced by 3 male and 3 female speakers, including 23 Mandarin Chinese consonants (/p, pʰ, tʰ, t, kʰ, k, tsʰ, ts, t̻ʰ, t̻, t̻ʂʰ, t̻ʂ, f, s, ʂ, ɕ, x, m, n, l, ɻ̩, j, w/, [17]).

In the English consonant identification in quiet test, 16 VCV tokens were used for each of the 23 consonants, making 368 VCV tokens altogether. The vowel contexts and stresses were balanced for each consonant. Another 10 VCVs were used as practice items at the beginning of the test. In the English consonant SRT test, only the VCV tokens in /æCæ/ context with end stress were used in order to reduce the variability of phonetic context, which would make the SRTs more stable between listeners and more comparable across corpora [11]. Four VCV tokens for each consonant were selected, and were repeated 5 times during the test, which makes it 20 tokens for each consonant, and 460 tokens all together. Similarly, there were 460 VCV tokens in the Mandarin consonant SRT test, 4 tokens for each of the 23 consonants, and which were repeated 5 times. 10 practice tokens were given at the beginning in both English and Mandarin SRT tests.

Speech Shaped Noise (SSN) was used as the noise masker in the SRT tests. All stimuli were normalized to have equal root-mean-square (RMS) energy and the noise was added immediately prior to presentation. The signal-to-noise ratio (SNR) for each stimulus was adjusted dynamically according to the SRT measure procedure.

C. Procedure

All tests were carried out at a sound-treated audiology test lab at Jiangsu University of Science and Technology. Listeners were tested one by one. The presentation of stimuli and the collection of responses were controlled by a customized MATLAB [18] program. Listeners finished the English SRT test first, followed by the Mandarin SRT test and English consonant identification in quiet test. In the English tests, listeners were asked to assign the consonant they heard in each VCV token to one of the 23 English consonant categories by clicking the corresponding button on a 4×6 on-screen button grid. Real English words with capital letters to indicate the corresponding consonant were shown on the buttons. Similarly, in the Mandarin test, the listeners' task was to classify the consonants they heard in the VCV tokens into the Mandarin consonant categories. Chinese characters with corresponding consonant to the syllable initial position were shown on the buttons in order to reduce orthographic influence [19].

For the English and Mandarin SRT tests, the SNR for each VCV token was modified dynamically according to the history of listeners' perception responses, following a 2-down 1-up adaptive procedure [20], and the step size was fixed at 2dB. For example, if the current SNR for an "aba" token was -4dB, and the listener gave an incorrect answer, then the SNR for the next "aba" token would be increased to -2dB. If the listener gave a correct answer, then the SNR for the next "aba" token would be kept at -4dB. If the listener could correctly identify the "aba" token at -4dB again, then the SNR for the third "aba" token would again be decreased to -6dB. The SRT for each consonant was calculated by averaging the SNR values for the last 5 tokens for that consonant.

Previous studies have demonstrated that different consonants have various reception thresholds [11], [21], and if the initial SNR is set too high, the SRT might not be reliable due to the lack of convergence in the last 5 SNR values (i.e., SNR values have been still continuously going down for the last few tokens) [9]. In the current study, a pilot test with an initial SNR fixed at 0dB was carried out to find out the proper initial SNR values for different English and Mandarin consonants. The initial SNRs were set based on individual consonants rather than using a fixed value for all sounds (see TABLE I).

TABLE I
INITIAL SNRS FOR ENGLISH AND MANDARIN CONSONANTS

English	p	b	t	d	k	g	tʃ	ʤ	f	v	θ	ð
Initial SNR	-6	0	-6	0	0	-6	-6	-4	-2	4	4	4
English	s	z	ʃ	ʒ	h	m	n	l	r	j	w	
Initial SNR	-4	0	-6	0	-6	-2	-4	0	-4	-2	-6	
Mandarin	p	b	t	d	k	g	c	z	ch	zh	q	j
Initial SNR	-4	0	4	-4	0	0	-4	0	-4	-4	-4	-4
Mandarin	f	s	sh	x	h	m	n	l	r	y	w	
Initial SNR	-2	-4	-4	-4	-6	-4	0	-4	-6	-4	-2	

III. RESULTS AND ANALYSIS

Native Chinese listeners' SRTs for Mandarin and English consonants are shown in TABLE II and III. As expected, native Chinese listeners' mean SRT over all Mandarin consonants (-6.5dB) was 3.6dB lower than their mean SRT over all English consonants (-2.9dB). Statistical analysis confirmed the difference was significant [$t(78) = -6.9844, p < 0.001$]. This result tends to indicate a general native language advantage for Chinese listeners. That is, they could tolerate more noise in their native language perception than in non-native language perception. However detailed observation of the SRTs on individual consonants revealed rather complicated patterns. It

can be seen from TABLE II that the SRTs for some English consonants (/ð, θ, ʒ, v/) were extremely high. In fact, the mean SRT for the 5 English consonants with the highest SRTs was significantly 15.6dB higher than the mean SRT for those in Mandarin $[t(78) = 16.6868, p < 0.001]$. For the 13 consonants in the middle in both languages, the mean SRT difference was reduced to 1.5 dB, though it was still significant $[t(78) = 2.4592, p < 0.05]$. An interesting and unexpected result was that the mean SRT for the 5 English consonants with the lowest SRTs being significantly 2.6 dB lower than that from the Mandarin consonants $[t(78) = -5.3484, p < 0.001]$, which means the Chinese listeners were able to better identify these non-native sounds than their native sounds in noise condition. Another interesting result was the great SRT difference between Mandarin /tʰ/ and English /t/.

TABLE II and III also list some phonetic-articulatory information such as place, manner and voicing for Mandarin and English consonants. The aspiration information and the correct identification rates in quiet conditions are also given for Mandarin and English consonants respectively. For the Mandarin consonants, it is difficult to generalize the conclusion that any particular feature of place, manner, voicing or aspiration for a consonant would affect their SRTs. Consonants from the same feature class could have varied SRTs. A similar situation can be found for most of the English consonants. However, it is notable that 4 (/ð, θ, ʒ, v/) of the 5 most difficult English consonants to perceive in noise were fricatives. TABLE II shows that these fricatives were also the most difficult sounds to identify in quiet conditions for Chinese listeners. In fact, a significant high negative correlation was found between English consonants' SRTs and their correct identification rate in quiet conditions ($r = -0.823, p < .001$), indicating language experience may have some influence on non-native consonant perception in noise.

The sibilant features for Mandarin and English consonants are also given in TABLE II and III. In the current study, the SRTs for the 9 Mandarin and 6 English sibilant sounds were relatively low (English /ʒ/ was an exception), consistent with previous studies' results, which indicates that sibilant sounds are found to be relatively better perceived in noise conditions [10], [11].

TABLE II
SRTs FOR ENGLISH CONSONANTS

Cons	SRT (dB)	M	P	V	Quiet (%)	Sib
ð	27	fricativ	dental	+	32	-
θ	23	fricativ	dental	-	45	-
ʒ	14	fricativ	post-alv	+	40	+
v	13	fricativ	lab-den	+	68	-
b	1	plosive	bilabial	+	98	
f	-1	fricativ	lab-den	-	94	-
k	-2	plosive	velar	-	81	
l	-2	lateral	alveolar	+	95	
z	-5	fricativ	alveolar	-	73	+
n	-5	nasal	alveolar	+	91	
m	-6	nasal	bilabial	+	98	
d	-6	plosive	alveolar	+	97	
s	-6	fricativ	alveolar	-	81	+
ʤ	-7	affricat	post-alv	-	72	+
p	-8	plosive	bilabial	-	96	
ʧ	-9	affricat	post-alv	-	88	+
r	-10	approxi	post-alv	+	86	
w	-10	approxi	approxi	+	79	
h	-12	fricativ	glottal	-	95	-
g	-12	plosive	velar	+	90	
j	-12	approxi	approxi	+	85	
ʃ	-15	fricativ	post-alv	-	91	+
t	-16	plosive	alveolar	-	93	
mean	-2.9				81.2	

Cons-Consonants, M-Manner of articulation, P-Place of articulation, V-Voicing, Quiet-Correct identification in English consonant identification in quiet test, Sib-Sibilant
"+" means with the feature, "-" means without the feature

TABLE III
SRTs FOR CHINESE CONSONANTS

Cons	SRT (dB)	M	P	V	Asp	Sib
tʰ	9	plosive	den-alvo	-	+	
n	-2	nasal	den-alvo	+		
p	-2	plosive	bilabial	-	-	
f	-2	fricativ	lab-den	-		-
t	-3	plosive	den-alvo	-	-	
ts	-3	affrica	t den-alvo	-		+
k	-5	plosive	velar	-	-	+
l	-5	lateral	den-alvo	+		
tɕʰ	-7	affricat	palatal	-	+	+
w	-7	approxi	bilabial	+		
s	-8	fricativ	den-alvo	-		+
pʰ	-8	plosive	bilabial	-	+	
ʃ	-8	fricativ	post-alv	-		+
tsʰ	-9	affricat	den-alvo	-	+	+
ʧʰ	-9	affricat	post-alv	-	+	+
ʧ	-9	affricat	post-alv	-	-	+
ɕ	-9	fricativ	palatal	-		+
m	-9	nasal	bilabial	+		
kʰ	-10	plosive	velar	-		
x	-10	fricativ	velar	-		-
tɕ	-11	affricat	palatal	-	-	+
ɹ	-11	approxi	post-alv	+		
j	-12	approxi	palatal	+		
mean	-6.5					

Cons-Consonants, M-Manner of articulation, P-Place of articulation, V-Voicing, Asp-Aspiration, Sib-Sibilant
"+" means with the feature, "-" means without the feature

IV. DISCUSSION

The current study investigated native Chinese listeners' SRT for their native Mandarin consonants and non-native English consonants. The results show that listeners' mean SRT for Chinese consonants was 3.6dB lower than that for English consonants, indicating a general native language advantage. These results are consistent with those of previous studies that indicate Chinese-English bilinguals' Mandarin sentence SRT was 2.8dB lower than their English sentence SRT [12]. In sentence perception in noise tests, listeners could benefit from a high level of language experience such as lexical, semantic and syntactic information [22]. The current study's results provides evidence that the native language advantage can be found in a low level of speech perception in noise as well.

Chinese listeners' consonant SRTs varied greatly in both Mandarin and English, even within the same phonetic-articulation category. No clear relation between a consonant's

SRT and its articulation features such as place, manner, and voicing was found in either language (except that non-sibilant fricatives are the most difficult in English, which will be discussed later), and no clear similar SRT pattern for different consonants between the two languages was found either. Consistent with the current study's findings, previous studies also demonstrated the great variability of consonant intelligibility in noise [11], indicating the masking effect of noise varies greatly due to different detailed acoustic features of different consonants rather than broad abstract phonetic-articulation classifications.

Sibilant sounds demonstrated relatively low SRTs for both Mandarin and English in the current study, while non-sibilant fricatives were among the most difficult sounds, especially in English. Similar results were reported in several previous studies where the intelligibility of English non-sibilant fricatives were the worst in noise conditions for both native and non-native listeners [10], [11], [23]. It has been suggested that this is a "language-independent" phenomenon. The reason that sibilant sounds are more intelligible in noise might due to the fact that the high frequency cues for sibilant sounds survive better in SSN. However, a noteworthy fact is that Chinese listeners' identification for English non-sibilant fricatives were also quite poor in quiet conditions. In fact, Chinese listeners' SRTs for English consonants were highly correlated with their identification performance for the same sounds in quiet conditions. This suggests that, in the current study, the intelligibility of non-sibilant fricatives in noise for Chinese listeners was affected by their L2 experience.

The mean SRT for the 5 most noise tolerable consonants in Mandarin was 2.6dB higher than that in English, suggesting that non-native speech perception in noise may not be always more difficult than the native one. This is an rather unexpected result. A possible explanation could be that the different detailed acoustic features of sounds from different languages may affect their intelligibility in noise conditions, even though they may be quite similar to each other in more general features. The large SRT difference between Mandarin /t^h/ and English /t/ in the current study is a good example. Although the Mandarin /t^h/ and English /t/ are quite similar to each other in inter-vocalic positions, however, the production of English /t/ has more dental feature than the Mandarin /t^h/, which gives the English /t/ more high-frequency energy and makes it more intelligible in SSN [11].

ACKNOWLEDGMENT

This study was supported by grants from the Humanities and Social Science Fund of the Ministry of Education of China (16YJC740020, 15YJC740034), the Jiangsu Social Science Fund (18YYB009) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX18_2292, KYCX18_2293).

REFERENCES

[1] D. Meador, J. E. Flege, and I. R. Mackay, "Factors affecting the recognition of words in a second language," *Bilingualism: Language and Cognition*, vol. 3, no. 1, pp. 55–67, 2000.

[2] M. Cooke, M. Garcia Lecumberri, and J. Barker, "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 414–427, 2008.

[3] S.-H. Jin and C. Liu, "English sentence recognition in speech-shaped noise and multi-talker babble for english-, chinese-, and korean-native listeners," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. EL391–EL397, 2012.

[4] L. H. Mayo, M. Florentine, and S. Buus, "Age of second-language acquisition and perception of speech in noise," *Journal of speech, language, and hearing research*, vol. 40, no. 3, pp. 686–693, 1997.

[5] C. L. Rogers, J. J. Lister, D. M. Febo, J. M. Besing, and H. B. Abrams, "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," *Applied Psycholinguistics*, vol. 27, no. 3, pp. 465–485, 2006.

[6] M. L. G. Lecumberri, M. Cooke, and A. Cutler, "Non-native speech perception in adverse conditions: A review," *Speech communication*, vol. 52, no. 11-12, pp. 864–886, 2010.

[7] M. G. Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2445–2454, 2006.

[8] A. Cutler, M. Cooke, M. L. G. Lecumberri, and D. Pasveer, "L2 consonant identification in noise: Cross-language comparisons," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[9] J. Gong, M. Cooke, and M. L. G. Lecumberri, "Can intensive exposure to foreign language sounds affect the perception of native sounds?," *Proc. Interspeech*, pp. 884–887, 2016.

[10] M. Cooke, M. L. G. Lecumberri, O. Scharenborg, and W. A. Van Dommelen, "Language-independent processing in speech perception: Identification of english intervocalic consonants by speakers of eight european languages," *Speech Communication*, vol. 52, no. 11-12, pp. 954–967, 2010.

[11] D. L. Woods, E. W. Yund, T. J. Herron, and M. A. U. Cruadhlaoich, "Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1609–1623, 2010.

[12] A. Stuart, J. Zhang, and S. Swink, "Reception thresholds for sentences in quiet and noise for monolingual english and bilingual mandarin-english listeners," *Journal of the American Academy of Audiology*, vol. 21, no. 4, pp. 239–248, 2010.

[13] A. N. S. Institute, "Specifications for audiometers (ansi s3. 6-2010)," 2010.

[14] M. Cooke and O. Scharenborg, "The interspeech 2008 consonant challenge," 2008.

[15] P. Roach, "British English: Received Pronunciation," *Journal of the International Phonetic Association*, vol. 34, no. 2, pp. 239–245, 2004.

[16] J. Gong, M. Cooke, and M. L. G. Lecumberri, "A computational modelling approach to the development of l2 sound acquisition.," in *ICPhS*, pp. 755–758, Citeseer, 2011.

[17] W. Lee and E. Zee, "Standard Chinese (Beijing)," *Journal of the International Phonetic Association*, vol. 33, no. 1, pp. 109–112, 2003.

[18] MATLAB, *version 9.2.0.538062 (R2017a)*. Natick, Massachusetts: The MathWorks Inc., 2017.

[19] J. Gong, M. Cooke, and M. Garcia Lecumberri, "Towards a quantitative model of mandarin chinese perception of english consonants," *Proc. NewSounds*, p. 128, 2010.

[20] H. Levitt, "Transformed up-down methods in psychoacoustics," *Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 467–477, 1971.

[21] D. L. Woods, Z. Doss, T. J. Herron, T. Arbogast, M. Younus, M. Ettlinger, and E. W. Yund, "Speech perception in older hearing impaired listeners: benefits of perceptual training," *PloS one*, vol. 10, no. 3, p. e0113965, 2015.

[22] J. Zaar and T. Dau, "Sources of variability in consonant perception of normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1253–1267, 2015.

[23] J. Gong, W. Zhou, and S. Zhang, "Effect of noise condition on the perception of l2 english consonants," *Language Education*, vol. 4, no. 2, pp. 44–52, 2016.