Effects of Phonetic Training with Noise on EFL learners' English Vowel Reception Thresholds

Yuyao Qin
School of Foreign Languages
Jiangsu University of Science and Technology
Zhenjiang, China
211111101116@stu.just.edu.cn

Jian Gong*
School of Foreign Languages
Jiangsu University of Science and Technology
Zhenjiang, China
j.gong@just.edu.cn

***Abstract*—To investigate the effectiveness of perceptual training with noise on learners' foreign vowel perception, a cohort of native Chinese listeners underwent intensive phonetic training to identify English vowels. Speech Shaped Noise was presented during the training procedure. Listeners' English vowel perception in noise ability was measured before and after training, by using a Vowel Reception Thresholds (VRT) metric, where the noise level was adjusted dynamically in an adaptive manner. Listeners' vowel identification in quiet condition was also measured. Significant VRT decrease was evident after training, along with a significant improvement of vowel identification in quiet. A significant negative correlation between individual vowel's VRT decrease and identification improvement was also found. These results indicated that high variability training with noise is effective and VRT can be used to quantify listeners' vowel perception ability change.**

Keywords—English vowel perception, phonetic training, speech shaped noise, vowel reception threshold,

I. Introduction

Spoken communication in noisy conditions presents problems for all listeners, but non-native listeners typically have more difficulties than native listeners [1], [2], [3]. One of the main reasons for this is because non-native listeners are less effective in processing of context and speech cues in adverse conditions than native listeners. Also, the perceptual cues that non-native listeners rely on might be different from that of native listeners' and are more easily masked by noise [4].

Normally the younger a learner starts learning L2, the better she/he be able to acquire L2 phonetic categories accurately. However, theoretical models in L2 perception claims that learners' L2 sound leaning ability remains intact over the life span [5]. Over the past few decades, a large body of studies had demonstrated that high-variability perceptual training (i.e., using multi-talker, multi-phonetic environments, natural speech, etc.) is an effective method to improve adult learner' L2 perception of consonants [6], [7], [8], vowels [9], [10], [11],[12], tones [13] and for production as well [14], [15]. Several recent studies have extended this high-variability training paradigm to learners' L2 speech perception in adverse environment. For example, native Spanish speakers were trained to perceive English consonants with background noise, and their English consonant identification was improved after training [16], [17]. Malaysian listeners' English phoneme discrimination under multi-talker babble was significantly improved by using high-variability training with multi-talker babble [18]. In another training study, native Chinese speakers were trained to identify English vowels in adverse acoustics background, and a significant performance improvement was evident after training [19]. The results from aforementioned studies all demonstrated that the high-variability training paradigm is also effective in improving non-native speaker's speech perception under noise conditions, and can help them to form more robust L2 sound categories [16], [17].

Usually, fixed noise levels are used in experiments of speech perception in noise, with just a few different signal-to-noise ratios (SNR). This may call into question when the purpose of the experiment is to quantify listeners' ability to perceive individual phonemes, because previous studies have demonstrated that a large range of SNRs is required for different phonemes to reach equal intelligibility [20]. Using some fixed SNRs for all phonemes may not accurately reflect the intelligibility for some sounds. Speech Reception Threshold (SRT) is commonly used to measure speech intelligibility in noise. However, most studies have applied the SRT procedure were to test native speech perception, with few trying to use it for non-native speech perception [21]. Recently, researchers employed SRT procedure to investigate non-native consonant perception in noise [22], and found that SRT was a useful metric in quantifying the non-native consonant perception ability changes in training [23]

Several recent studies indicated that at the phonemic level, vowel perception was significantly more difficult than consonant perception for English learners in China and the US [24], [25]. To the author's knowledge, no study has ever adopted SRT procedure to measure the perception of English vowel in a training study. Therefore, the purpose of the current study is to investigate the effectiveness of perceptual training with noise on Chinese EFL learners' English vowel perception, and to explore whether SRT can be a useful metric to quantify the vowel perception ability changes in noise.

II. Methods

A. Methods

A cohort of 20 native Chinese subjects, including 12 males and 8 females, participated in the current study. These subjects were students from Jiangsu University of Science and Technology with ages ranging from 19 to 30 years (M=23years.) No subject had reported any hearing or language problems, and all the subjects had passed a hearing test with pure-tone thresholds⩽15dB HL at octave intervals between 250 and 8000Hz[26]. Subjects were all from the Northern Mandarin dialect spoken region (north of the Yangtze River). They were all EFL learners but majoring in various courses, and most of them had passed a standardized English test required for college students in China (college English test band 4, CET-4). These subjects were further randomly assigned to a control group and a training group, each contained 10 people. The control group only participated

*Corresponding author

979-8-3503-3078-6/23/$31.00 ©2023 IEEE

in the pre- and post-tests, while the training group received intense phonetic training with noise in-between pre- and post-tests. Subjects were paid for their participation.

B. Speech Materials

There were two tasks in both pre- and post-tests, namely an English Vowel Reception Threshold (VRT) test and an English vowel identification in quiet test. The same materials were used in the pre- and post-tests. The English vowels used in current experiments were 10 monophthongs /ɪ, i, e, æ, ʌ, ɑ, ɒ, ɔː, ʊ, u ː/and 5 diphthongs /eɪ, aɪ, ɔɪ, əʊ, aʊ/. These 15 English vowels were put in /hVd/ frame and the corresponding 15 words were *heed, hid, head, hade, hard, had, hudd, hod, hoard, hood, who'd, hoad, hide, how'd, hoyd*. These 15 words were all real English words, produced by three native British speakers (2 males and one female) in a natural manner, yielding a total of 270 tokens (15 vowels×3 speakers×6 repetitions).

The training materials were different from those used in the pre-test and post-test, including several different CVC frames (e.g., /bVd/, /bVt/, /fVd/). Recordings of these monosyllabic English words were made from another six speakers of British English (4 males, 2 females) who did not appear in the pre-test and post-test phase. There were 16 training sessions in the current study. Each session included 4 blocks of 150 trials (15 vowels×10 tokens), with 16 sessions yielding a total of 2400 trials (16 session ×150 trails). Therefore, there were 2400 CVC stimuli altogether in the training.

Speech Shaped Noise (SSN) was used as the noise masker in the VRT test and training. The SSN was generated by passing white noise through a 50 coefficient filter with the shape of the long-term average spectrum derived from 200 sentences in a British English corpus [27]. 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 same SRT measure procedure in [22] and [28] for VRT test, while four levels of fixed SNRs were used in training (see sample stimuli in Fig.1).

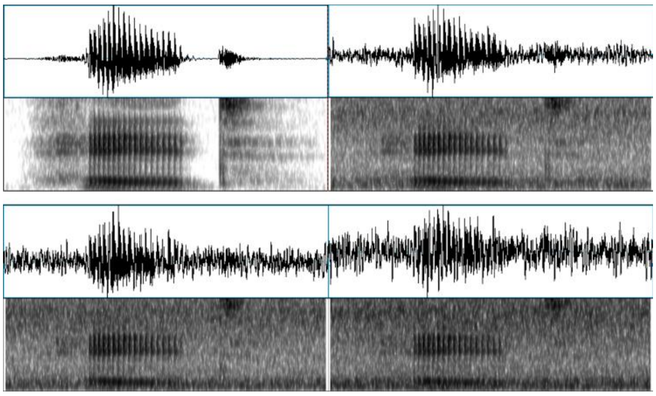

Fig.1. Samples of "hid" stimuli mixed with SSN. Top-left is the original clean speech, top-right, bottom-left and bottom-right are with SNR at 4dB, 0dB and -4dB respectively.

C. Procedures

1) Overall structure of the experiment: The experiment was carried out in a sound treated lab in Jiangsu University of Science and Technology, following the pre-test–training–post-test paradigm. Subjects were asked to finish all the tasks individually over 6 consecutive days. Pre- and post-tests were carried out in the first and last day, and 4 sessions of training were given in-between, one per day. Stimuli were delivered via a AKG K271 MkII headphone and a RME Fireface 800 sound card in all tasks, and the presentation of stimuli and collection of responses were controlled by a customized MATLAB [29] program.

2) Forced choice identification paradigm: In the pre- and post-tests, subjects finished the English vowel identification in quiet test first, followed by the English VRT test. The 15 alternatives forced choice identification paradigm was used in both tasks, in which the subjects were asked to assign the vowel they heard in each /hVd/ stimuli to one of the 15 English vowel categories by clicking the corresponding button on a 3×5 on-screen button grid. The 15 /hVd/ frames (*hid, heed, head, had, hudd, hard, hod, hoard, hood, hood, who'd , hade, hide, hoyed, hoad and how'd*) were shown on the buttons.

3) VRT test procedure: Different from the vowel identification in quiet test, the hVd stimuli were delivered together with noise in the English VRT test. The SNR for each hVd token was modified dynamically according to the history of subjects' perception responses, following a 2-down 1-up adaptive procedure [30], and the step size was fixed at 2dB. For example, if the current SNR for an "hid" token was -4dB, and the listener gave an incorrect answer, then the SNR for the next "hid" token would be increased to -2dB (1-up). If the listener gave a correct answer, then the SNR for the next "hid" token would be kept at -4dB. If the listener could correctly identify the "hid" token at -4dB again, then the SNR for the third "hid" token would be decreased to -6dB (2-down). The SRT for each vowel was calculated by averaging the SNR values for the last 5 tokens for that vowel, and was refer to as Vowel Reception Threshold (VRT) by analogy with SRT. Previous studies have demonstrated that different phonemes have various speech reception thresholds [20], [31], 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., where SNR values have been still continuously going down for the last few tokens) [22]. In the current study, the initial SNRs were set based on individual vowels rather than a fixed value for all sounds (see TABLE1), according to a pilot study.

4) Training: The 16 training sessions were conducted over 4 consecutive days, each containing 4 blocks. The CVCs were mixed with SSN at 4 different fixed SNRs of 3, 0, -3and -6dB, one for each block. Using different SNR values aimed to promote variability in the availability of speech cues in different noise level, and to simulate everyday noisy environments [16]. Similar to the English VRT test, the 15 alternative forced choice identification paradigm was followed in the training. Subjects had to classify the vowel they heard in each CVC into one of the 15 categories. However, if the subjects gave an incorrect response, they would get feedback immediately, that is, the correct answer would be highlighted and the subject had to click the right answer and listen again to proceed.

TABLE I. INITIAL SNRS FOR ENGLISH VOWELS

Vowel Initial SNR	ɪ	i:	e	æ	ʌ	ɑ:	ɒ	ɔ:
(dB)	-6	-6	-6	-6	-6	-6	-8	-6
Vowel Initial SNR	ʊ	u:	eɪ	aɪ	ɔɪ	əʊ	aʊ	
(dB)	-6	-6	-6	-10	-10	-10	-10	

III. RESULTS

A. Vowel Reception Threshold

Fig. 2 shows the mean Vowel Reception Thresholds over all vowels for control and training groups in pre- and post-tests. Repeated-measures ANOVA confirmed that there was a significant main effect of test (pre-post) [F (1, 18) = 83.50, $p < 0.001$, η_p^2= 0.823]and a significant main effect of group (control-train) [F (1, 18) =10.18, p = 0.005, η_p^2= 0.361]. The interaction between test and group was also significant [F (1, 34) = 21.9, $p < 0.001$, η_p^2= 0.393]. Further simple effect analysis with Bonferroni adjustment revealed that there were significant differences of VRT between the two groups before training (p = 0.039) and after training (p = 0.002). There was a small but significant 4.90dB VRT change for the control group after training (6.42dB for pre-test and 1.52dB for post-test, $p < 0.001$), while the training group's VRT had a significant 10.71dB decrease ($p < 0.001$) from pre-test (1.14dB) to posttests (-9.58dB). Statistical result confirmed that training group had significant larger VRT decrease than control group (p<0.05).

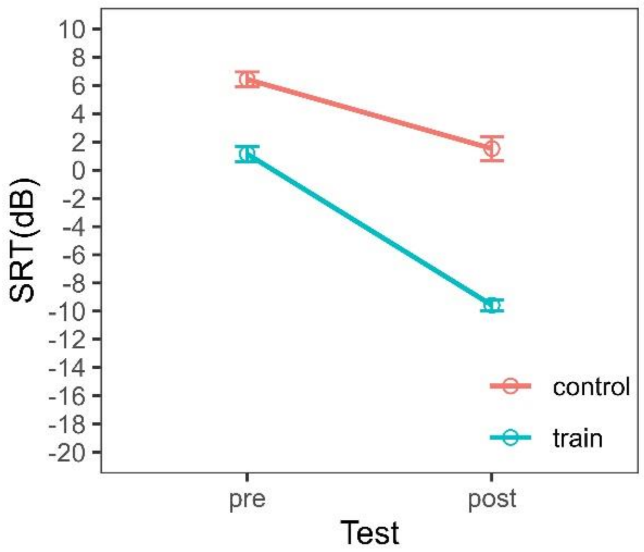

Fig.2. Vowel Reception Threshold before and after training

B. Identification accuracy in quiet

The mean identification accuracy across all 15 vowels for both control and training groups in pre- and post-test are shown in Fig. 3. Repeated-measures ANOVA indicated that there was a significant test effect [F (1, 18) = 63.97, $p < 0.001$, η_p^2= 0.780], a significant group effect [F (1, 18) = 13.26, p = 0.002, η_p^2= 0.424] and a significant test×group interaction [F (1, 18) = 4.87, $p < 0.41$, η_p^2= 0.213]. Further simple effect analysis with Bonferroni adjustment revealed that there was a significant difference between the control group (42%) and the training group (56%) before training ($p < 0.05$) and after training (55% for control and 79% for training, p<0.001). After training, the training group's performance significantly improved ($p < 0.001$) and a significant change was also found for the control group ($p < 0.001$). Similar to the VRT results, statistical analysis confirmed that training group's improvement was significantly larger than the control group (23% vs 13%, p<0.01).

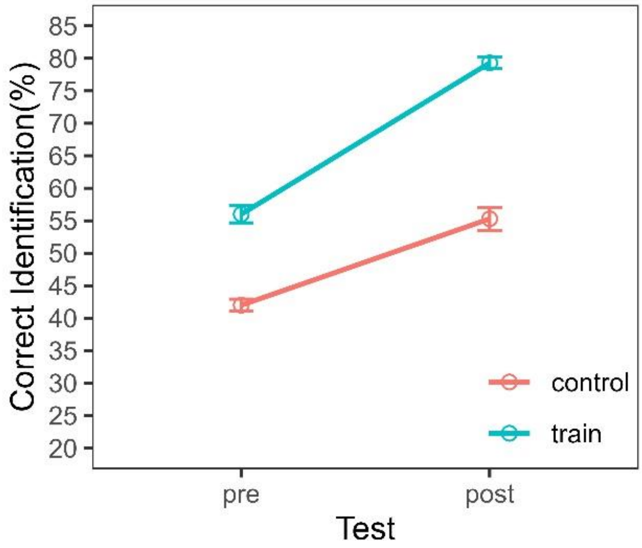

Fig.3. Identification accuracy before and after training

C. Individual Vowels

1) Individual vowel's VRT changes. Individual vowel's VRT changes from pre- to post-test for the training group are shown in Fig. 4. It can be seen that all of the vowels demonstrated a VRT decrease after training. Paired-samples t-test indicated that 11 of them (/aɪ, aʊ, eɪ, ɔɪ, əʊ, i:, ɒ, ɔ:, ʊ, u:, ʌ /) reached significance ($p < 0.05$), while the other 4 (/ɑ:, æ, e, ɪ /) were not significant ($p > 0.05$). There were 7 vowels (/əʊ, eɪ, ʌ, ɔ:, ɒ, aʊ, i: /) that demonstrated large VRT decrease (over 10dB) after training. The largest VRT change came from the diphthong / əʊ /, with a huge 26.44dB decrease after training. There were only 3 vowels (/ɔɪ, e, ɑ :/) showed relatively small VRT decrease. In summary, these results indicated that training was generally effective in improving Chinese learners' individual English vowel' perception in noise

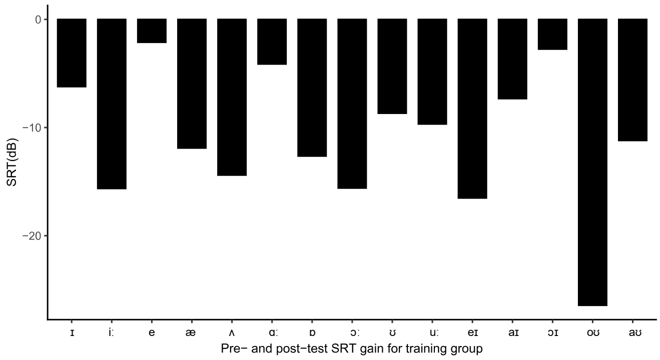

Fig.4. Individual vowels' VRT changes from pre-test to post-test

2) Individual vowel's identification accuracy changes. Individual vowel's identification accuracy changes from pre- to post-test for the training group are shown in Fig. 5. It can also be seen that all of the vowels showed an identification increase after training. Paired-samples t-test indicated that most of them (/aɪ, aʊ, eɪ, ɔɪ, əʊ, i:, ɪ, ɒ, ʊ, u:, ʌ, æ, e /) reached significance (p<0.05), while only 2 (/ɑ :, ɔ: /) were not significant (p>0.05). There were 6 vowels (/əʊ, eɪ, ʌ, ɔː, aʊ, iː /) that demonstrated large identification accuracy increase (over 20%) after training. The largest identification changes also came from the diphthong / əʊ /, with a huge 58% increase after training. Only 3 vowels (/ɔɪ, ɑː, e/) showed small increase. Interestingly, statistical analysis revealed that there was a significant negative correlation between individual vowel's VRT decrease and identification improvement ($r = -0.887$, $p < 0.001$), indicating a general tendency that, for Chinese learners, the more their performance improved in noise, the more their performance improved in quiet.

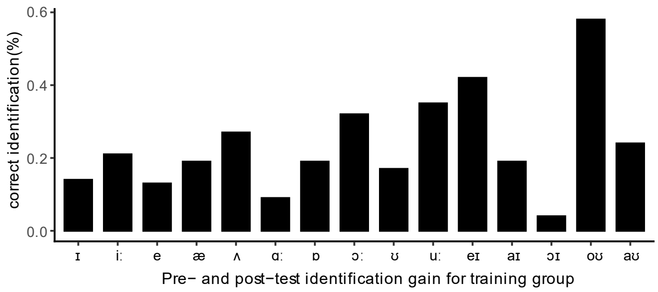

Fig.5. Individual vowels' correct identification changes from pre-test to post-test

3) Individual vowel SNR evolution in VRT tests. Fig. 6 showed the evolution of SNR presentation for each vowel during pre- and post-test for the training and control group respectively. Vowel reception thresholds were estimated by computing the across-listener mean threshold at each point in the presentation sequence, and then taking the mean over the final five SNR presentations as the VRT value. For the training group, SNR fell from pre- to post-test, and most vowels' SNR became stable after training, which indicating an improvement in identifying vowels in noise. While for the control group, a few vowels' SNR showed a slight fall, however, almost no vowel's SNR became stable in the post-test.

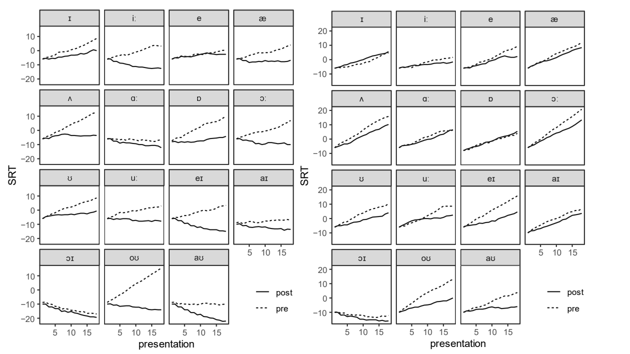

Fig.6. Per-vowel SNR evolution in pre and post-test for training group (left) and control group (right)

IV. Discussion

The primary purpose of this study was to investigate the effectiveness of phonetic training with noise on native Chinese EFL learners' identification of English vowels. After just 16 sessions of intensive training, Chinese learners' overall English VRT was significantly decreased, suggesting a significant improvement of vowel perception ability in noisy environment. This result is in line with several previous studies' findings that indicate perceptual training with noise is generally an effective method to improve non-native perception in adverse acoustic conditions [16], [17], [18], [19]. It should be noted that there was also a VRT improvement from pre- to post-test for the control group in the present study. However, the much larger improvement for the training group was solid evidence that the training was effective. The VRT decrease for the control group might be due to task repetition effect or adaptation to the noise masker, reflecting the procedural learning suggested in [22] and [30].

Previous studies demonstrated that training under adverse conditions can not only benefit non-native perception in adverse environments, but can also improve learners' perception performance in quiet condition [16], [17], [18], [19]. In the current study, Chinese learners' overall English vowel identification accuracy improved significantly after training, providing new evidence that phonetic training in noise is an effective way to improve non-native perception in all conditions. In a previous training in noise study, an improvement of around 48 percentage points was observed for native Chinese listeners on English back vowel perception, after 6 sessions of training with about 180 training tokens for each vowel in total [19]. In the current study, a 23.3 percentage points improvement was achieved for native Chinese subjects using 160 tokens for each vowel all together in 16 training sessions. One possible reason may account for the discrepancy of the training effects between the two studies. That is, the current study had the pre-training score already above 50%, whereas the pre-training score in [19] was only 30%. The relative higher score in the current study might leave less room for improvement. Moreover, the training effect might be affected by a number of factors such as speech materials used (e.g., isolated vowels, vowels in CVC syllables or words], perception tasks (e.g., phonetic identification or discrimination of phonetic contrasts), listening environments (quiet or noise), and training frequency and duration (e.g., few days vs several weeks) [19].

Previous studies derived different results on whether training in noise is more effective than training in quiet. [19] found that training in noise was more effective than training in quiet for vowel perception in both quiet and noise conditions. However, [16] demonstrated that training in quiet led to better consonant identification performance in quiet while training in noise showed some advantages for noise conditions. The authors of [16] argued that training in quiet can maintain all the spectral and temporal information, while training in noise might lose some of them. However, listeners undergoing noise-based training might compensate for this by learning the noise-robust cues, leading to better performance in noise conditions. The current study didn't directly tackle the issue of whether training in noise is more effective than training in quiet, however, the fact that there was a significant high correlation between VRT decrease and identification improvement indicated that Chinese learners could possibly benefit from the noise-based training and apply the robust perceptual cues they learned to both normal and adverse environments.

A dynamic SRT procedure was applied in the current study to investigate Chinese learners' non-native vowel perception in noise. All the vowels showed VRT decrease and SNR convergence after training. On top of that significant correlation was found between VRT decrease and identification improvement, which was similar to a previous study that applied SRT procedure in consonant perception

[23]. These results suggest that SRT can be a useful metric in quantifying the ability changes of non-native sound perception in noise, for both vowels and consonants.

V. Conclusion

High variability phonetic training with noise is a effective way to improve learners' non-native vowel perception in both normal and adverse listening conditions. The VRT can be a useful metric in quantifying changes of ability for non-native vowel perception in noise.

VI. Acknowledgment

This study was supported by the grant Scientific and Technological Innovation Team of Jiangsu University of Science and Technology (2020).

References

[1] M. L. G. Lecumberri, M. Cooke,Effect of masker type on native and non-native consonant perception in noise[J]. The Journal of the Acoustical Society of America, 2006, 119(4): 2445-2454.

[2] M. Cooke, M. L. Garcia Lecumberri, J. Barker,The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception[J]. The Journal of the Acoustical Society of America, 2008, 123(1): 414-427.

[3] A. Cutler, M. L. Garcia Lecumberri, Cooke M, Consonant identification in noise by native and non-native listeners: Effects of local context[J]. The Journal of the Acoustical Society of America, 2008, 124(2): 1264-1268.

[4] M. L. G. Lecumberri, M. Cooke, A. Cutler, Non-native speech perception in adverse conditions: A review[J]. Speech communication, 2010, 52(11-12): 864-886.

[5] J. E. Flege, Second language speech learning: Theory, findings, and problems[J]. Speech perception and linguistic experience: Issues in cross-language research, 1995, 92: 233-277.

[6] J. S. Logan, S. E. Lively, D. B. Pisoni, Training Japanese listeners to identify English/r/and/l: A first report[J]. The Journal of the Acoustical Society of America, 1991, 89(2): 874-886.

[7] A. R. Bradlow, J. Hansen, M. Zampini, Training non-native language sound patterns[J]. Phonology and second language acquisition, 2008: 287-308.

[8] J. Gong, M. L. G. Lecumberri, M. Cooke, Ab initio perceptual learning of foreign language sounds: Spanish consonant acquisition by Chinese learners[J]. System, 2017, 66: 142-155.

[9] S. G. Lambacher, W. L. Martens, K. Kakehi, et al, The effects of identification training on the identification and production of American English vowels by native speakers of Japanese[J]. Applied Psycholinguistics, 2005, 26(2): 227-247.

[10] K. Nishi, D. Kewley-Port, Training Japanese listeners to perceive American English vowels: Influence of training sets[J]. 2007.

[11] J. Gong, Z. Yang, X. Ji, et al, Investigating the effectiveness of auditory training on Chinese listeners' perception of English vowels[C]//Proc. ICPhS XIX. 2019: 2253-2257.

[12] J. C. Mora, M. Ortega, I. Mora-Plaza, et al, Training the pronunciation of L2 vowels under different conditions: the use of non-lexical materials and masking noise[J]. Phonetica, 2022, 79(1): 1-43.

[13] Y. Wang, A. Jongman, Sereno J A, Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training[J]. The Journal of the Acoustical Society of America, 2003, 113(2): 1033-1043.

[14] P. Iverson, M. Pinet, B. G. Evans, Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels[J]. Applied Psycholinguistics, 2012, 33(1): 145-160.

[15] Y. Shinohara, P. Iverson, High variability identification and discrimination training for Japanese speakers learning English/r/–/l[J]. Journal of Phonetics, 2018, 66: 242-251.

[16] M. Cooke, M. L. Garcia Lecumberri, Effects of exposure to noise during perceptual training of non-native language sounds[J]. The Journal of the Acoustical Society of America, 2018, 143(5): 2602-2610.

[17] M. Cooke, M. L. García Lecumberri, Non-native consonant acquisition in noise: Effects of exposure/test similarity[J]. The Journal of the Acoustical Society of America, 2019, 146(1): 297-306.

[18] C. X. R. Leong, J. M. Price, N. J. Pitchford, et al, High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained[J]. PloS one, 2018, 13(10): e0204888.

[19] L. Mi, S. Tao, W. Wang, et al, Training non-native vowel perception: In quiet or noise[J]. The Journal of the Acoustical Society of America, 2021, 149(6): 4607-4619.

[20] D. Woods, E. W. Yund, T. J. Herron, et al, Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise[J]. The Journal of the Acoustical Society of America, 2010, 127(3): 1609-1623.

[21] A. Stuart, J. Zhang, S. Swink, Reception thresholds for sentences in quiet and noise for monolingual English and bilingual Mandarin-English listeners[J]. Journal of the American Academy of Audiology, 2010, 21(04): 239-248.

[22] J. Gong, M. L. G. Lecumberri, M. Cooke, Can intensive exposure to foreign language sounds affect the perception of native sounds?[C]//INTERSPEECH. 2016: 883-887.

[23] J. Gong, Y. Yu, W. Bellamy, et al, Effect of Perceptual Training with Noise on Chinese Learners' English Consonant Reception Thresholds[C]//2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2021: 1087-1091.

[24] L. Mi, S. Tao, W. Wang, et al, English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners[J]. The Journal of the Acoustical Society of America, 2013, 133(5): EL391-EL397.

[25] W. Hu, S. Tao, M. Li, et al, Distinctiveness and assimilation in vowel perception in a second language[J]. Journal of Speech, Language, and Hearing Research, 2019, 62(12): 4534-4543.

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

[27] M. Cooke, M. L. G. Lecumberri, O. Scharenborg, et al, Language-independent processing in speech perception: Identification of English intervocalic consonants by speakers of eight European languages[J]. Speech Communication, 2010, 52(11-12): 954-967.

[28] J. Gong, Y. Yu, W. Belllamy, et al, Comparing native chinese listeners' speech reception thresholds for mandarin and english consonants[C]//2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019: 2038-2041.

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

[30] H. Levitt, Transformed up‐down methods in psychoacoustics[J]. The Journal of the Acoustical society of America, 1971, 49(2B): 467-477.

[31] D. L. Woods, Z. Doss, T. J. Herron, et al, Speech perception in older hearing impaired listeners: Benefits of perceptual training[J]. PloS one, 2015, 10(3): e0113965.