

# Effects of fundamental frequency and vocal tract resonance on speech recognition in noise by non-native listeners

Xiao Xiao<sup>1</sup>, Jing Yang<sup>2\*</sup> , Michael S. Harris<sup>3</sup> , and Li Xu<sup>4</sup> 

<sup>1</sup>Department of Music, Hunan University of Science and Technology, Xiangtan, Hunan, PR China

<sup>2</sup>Program of Communication Sciences and Disorders, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

<sup>3</sup>Department of Otolaryngology & Communication Sciences, Medical College of Wisconsin-Milwaukee, Milwaukee, WI, USA

<sup>4</sup>Department of Hearing, Speech and Language Sciences, Ohio University, Athens, OH, USA

Received 5 April 2025, Accepted 10 June 2025

**Abstract** – The present study examined the influence of changes in speakers' fundamental frequency ( $f_o$ ) and vocal tract resonance (VTR) on speech recognition in different types of noise by non-native listeners. The goal was to identify whether the  $f_o$ -VTR relationship has a similar effect on non-native listeners as it does on native listeners. Twenty-six adults who were native Mandarin speakers learning English as a second language were presented with English Hearing-in-Noise Test (HINT) sentences in four voice conditions with the original male speaker's  $f_o$  doubled and/or VTR scaled up by a factor of 1.2: (1) low  $f_o$  low VTR ( $L_{f_o}L_{VTR}$ , the original recordings); (2) low  $f_o$  high VTR ( $L_{f_o}H_{VTR}$ ); (3) high  $f_o$  high VTR ( $H_{f_o}H_{VTR}$ ), and (4) high  $f_o$  low VTR ( $H_{f_o}L_{VTR}$ ). The stimuli were presented in speech-shaped noise (SSN) and four-talker babble (FTB) at signal-to-noise ratios of  $-3$ ,  $0$ ,  $+3$  dB. The results showed that the non-native listeners performed more poorly with  $f_o$ -VTR mismatched voices than with  $f_o$ -VTR matched voices and the negative influence of mismatched voice features was mainly manifested in the  $H_{f_o}L_{VTR}$  condition. Compared to SSN, FTB had a greater adverse impact on the non-native listeners' recognition accuracy. Further, the performance difference between matched and mismatched conditions showed distinct patterns across SSN and FTB.

**Keywords.**  $f_o$ , Vocal tract resonance, Sentence recognition, Non-native listeners

## 1 Introduction

According to the source-filter theory, speech production (e.g., vowel production) involves two major mechanisms: vocal fold vibration that generates a vocal source [with the rate of vibration determining the fundamental frequency ( $f_o$ )] and vocal tract resonance (VTR) that filters the vocal source. The variations of these two components are reflected in the differences in speakers' vocal tract size and vocal fold structure that change as a function of the speaker's sex and age. Numerous studies have reported that both acoustic correlates contribute to the identification of a speaker's sex [1–3]. The two correlates are also associated with a speaker's speech intelligibility [4–6].

Assmann and Nearey (2008) [4] applied downward and/or upward shifting to the formant frequencies (e.g., by a factor of 0.8, 1.0, 1.5, etc.) and/or  $f_o$  (by a factor of 0.5, 1.0, 2.0, etc.) of adult male, adult female, and child speakers to create resynthesized /hVd/ words via STRAIGHT vocoder for vowel recognition. The results

showed that the recognition accuracy dropped when one feature was shifted alone. However, when  $f_o$  and spectrum envelope were shifted in the same direction together, recognition accuracy improved, even when the shifting scale factors were in the range outside of natural speech. The authors proposed that listeners tend to have an internalized knowledge about the covariation of  $f_o$  and formant. Mismatched  $f_o$  and formant has an adverse impact on speech intelligibility. In a recent study [7], we modified the  $f_o$  and/or VTR (i.e., formant values) of the male speaker of Hearing-in-Noise Test (HINT) sentences by shifting up  $f_o$  with a factor of 2.0 and/or shifting up VTR with a factor of 1.2. This setting conforms to the ratios of group mean vocal tract length and  $f_o$  between female and male speakers. The stimulus sentences were mixed with speech-shaped noise and were presented to listeners at four signal-to-noise ratios (SNRs at  $-10$ ,  $-5$ ,  $0$ ,  $+5$  dB). While there was no significant difference in the intelligibility between  $H_{f_o}H_{VTR}$  (high  $f_o$  paired with high VTR) and  $L_{f_o}L_{VTR}$  (low  $f_o$  paired with low VTR), the intelligibility declined significantly for  $H_{f_o}L_{VTR}$  (high  $f_o$  paired

\*Corresponding author: [jyang888@uwm.edu](mailto:jyang888@uwm.edu)

with low VTR) and  $L_{f_o}H_{VTR}$  (low  $f_o$  paired with high VTR) at certain SNRs. The intelligibility gap between the matched and mismatched conditions (e.g.,  $H_{f_o}H_{VTR}$  and  $H_{f_o}L_{VTR}$ ) could reach up to 20 percentage points. These results echoed previous findings about listeners' sensitivity to  $f_o$ -VTR relationship. Violating the relationship could negatively influence speech intelligibility in adverse listening conditions.

Extending beyond this earlier work, in the present study, we examined whether sensitivity to the covariation relationship between  $f_o$  and VTR would be manifested in non-native listeners. The covariation of  $f_o$  and VTR is determined by the sex-related anatomical differences (i.e., on average, male speakers, as a group, tend to have a lower  $f_o$  and lower formant frequencies compared to female speakers) rather than language-specific features. When the same set of English stimuli were presented to non-native listeners, we hypothesized that they should also demonstrate a similar pattern of higher recognition accuracy for stimuli with a matched  $f_o$ -VTR relationship compared to a mismatched relationship, although the overall recognition performance of non-native listeners would likely be lower than the native listeners. As speaker characteristics are largely indexed in their voice features, the results of the present study provide valuable information to further our understanding of the impact of adverse source involving speaker characteristics on non-native speech perception.

In addition to speech-shaped noise, in the current research, we also presented sentence stimuli with multi-talker babble as maskers. We wondered whether listeners' performance in the four voice conditions would show a different pattern between speech-shaped noise and multi-talker babble. Babble speech is a fluctuant speech noise that introduces informational masking in addition to energetic masking [8]. Listeners' ability to segregate target speech from babble noise is interfered with the audible fragments of the babble masker. Previous studies examining non-native perception in noise have shown that non-native listeners consistently performed more poorly than did native listeners and the performance difference between native and non-native listeners was greater in noise conditions than in the quiet condition [9–13]. Compared to speech-shaped noise, multi-talker babble had a greater masking effect on both native and non-native listeners [9, 12], with the extent of the effect depending on the number of talkers in the babble (e.g., [8]).

Unlike speech-shaped noise that does not contain voice cues, speech babble contains speakers' voice information. In speech-on-speech perception (with either one competing talker or multiple talkers), voice difference between the target speaker and babble speaker(s) could serve as an important cue to segregate the target speech from babble noise [14–19]. Darwin et al. (2003) [15] reported that compared with two competing sentences produced by the same speaker, increases in  $f_o$  difference ( $>2$  semitones) or increases in VTR difference (ratio  $> 1.08$ ) alone both improved the recognition performance of the

target speech. Systematic changes in both  $f_o$  and VTR towards a different sex between the target and masker speech introduced the greatest improvement. Flaherty et al. (2021) [19] tested sentence recognition threshold in a two-talker masker by children of different ages. The stimuli were presented in four conditions: target and masker speech with no differences in  $f_o$  and VTR, differed in  $f_o$  only, differed in VTR only, or differed in both  $f_o$  and VTR (high  $f_o$ -high VTR or low  $f_o$ -low VTR). The results revealed that child listeners showed greater improvement in masking release when the target and masker speech differed in both features compared to the conditions in which the target and masker speech differed in one single feature. Both studies evidenced that listeners, including children and adults, benefit more from greater voice distinctiveness between target and masker speech in speech-on-speech recognition. In the present study, the two sex-related acoustic features of the target speech were manipulated but the babble speech were produced by male or female speakers with no feature changed. We hypothesized that listeners would show different patterns of perception performance in the two different types of noise: (a) the target speech in mismatched voice conditions would be more salient compared to that in matched voice conditions in the presence of speech babble and (b) the negative impact of mismatched voice features would be reduced in babble maskers compared to speech-shaped noise.

## 2 Methods

*Listeners.* A total of 26 Mandarin-speaking English learners (24 females, 2 males) participated in this study. All the listeners aged between 20 and 43 years old ( $M = 23.8$  yrs,  $SD = 5.2$  yrs) and self-reported as having normal hearing. The listeners were senior-year college students or graduate students recruited in Central South region of China. All had received English language instruction since elementary or middle school in China, and 20 of the 26 participants majored in English. None of the listeners had experience residing in English-speaking countries. A web-based Vocabulary Level Test (VLT) [20] and Lexical Test for Advanced Learners of English (LexTALE) [21] were given to each listener. VLT measures the degree of mastery of English words in the second, third, fifth, and tenth 1000-word frequency bands (2000, 3000, 5000, 10 000 level) as well as Coxhead's (2000) Academic Word List [22]. LexTALE is a non-speeded visual lexical decision task targeting tester's English vocabulary knowledge. For our participants, the group average LexTALE score was 77.1%, and the group average VLT score was 85.7. The results of the two English language tests were highly correlated ( $r = 0.807$ ). A large-scale study testing LexTALE revealed that the average score for Dutch and Korean advanced English learners was 70.7% [21]. In the present study, 21 of the 26 participants scored higher than 70%. For the VLT, 23 of the 26 participants had an average score higher than 70. Therefore, we contended

that our sample of non-native listeners was proficient in English.

*Perception stimuli.* The perception stimuli were HINT sentences [23] with the voice features of  $f_o$  and VTR manipulated. HINT sentences included 26 original lists each containing 10 sentences. The same set of stimuli was used in [7]. The original HINT sentences were produced by a male speaker. In the current study, we manipulated the  $f_o$  and/or the vowel formants of the original sentences using the built-in ‘‘Change Gender’’ function in PRAAT software [24]. There were four voice conditions: male  $f_o$  paired with male VTR ( $L_{f_o}L_{VTR}$ ) by adopting the original sentences produced by the male speaker; female  $f_o$  paired with female VTR ( $H_{f_o}H_{VTR}$ ) by doubling the  $f_o$  of the original male voice (i.e., from a mean of 110 Hz to a mean of 220 Hz) and scaling up the formants of the male speaker by a factor of 1.2 [3]; female  $f_o$  paired with male VTR ( $H_{f_o}L_{VTR}$ ) by doubling the  $f_o$  of the original male voice but maintaining the formant values of the original male speaker; male  $f_o$  paired with female VTR ( $L_{f_o}H_{VTR}$ ) by maintaining the original male  $f_o$  but scaling up the formants of the male speaker by a factor of 1.2. For all three manipulated conditions, the duration of sentences remained the same as the original sentences. The total number of sentences was 1040 (26 lists  $\times$  10 sentences  $\times$  4 voice conditions). All sentences were root-mean-square (RMS) equalized.

*Procedure.* The perception tests were administered in a sound-treated room through a custom MATLAB program. The speech stimuli were mixed with two types of noise: speech-shaped noise (SSN) and four-talker speech-babble noise (FTB) at three SNRs:  $-3, 0, +3$  dB. These SNRs were selected to avoid ceiling or flooring effects of masking on sentence recognition in non-native listeners. The SSN were generated by matching the long-term average spectrum of a white noise to that of the concatenated 1040 sentences. For each SSN-mixed stimulus file, the speech-plus-noise portion was preceded and followed by a 200 ms SSN. The FTB was generated by mixing speech narratives produced by two native English female speakers and two native English male speakers. Previous studies have shown that listeners’ recognition of vowels, consonants, words, and sentences in the presence of multi-talker babble varied as a function of the number of babble talkers [8, 25, 26]. In general, as the number of babble talkers increases, the effect of informational masking decreases [14, 26]. On the other hand, very few babble talkers result in substantial masking release. In the current study, we set the number of babble talkers as four to ensure the appropriate level of masking. The speech narratives were selected from English audiobook compact discs. For each FTB-mixed stimulus file, the speech-plus-babble portion was preceded by a 1400 ms babble with a 400 ms beep inserted in the middle. The beep sound was used as an alert to caution the listeners about the start of the target speech. The FTB continued for 500 ms after the end of the target sentence. The level of the sentence stimuli was fixed and the desired SNR was achieved by changing the RMS

level of the maskers relative to that of the sentence stimulus. Sentence recognition was measured in 24 conditions (4 voice conditions  $\times$  2 maskers  $\times$  3 SNRs). One sentence list was used for each test condition.

All listeners completed a practice session prior to the actual testing. The practice session included 4 voice conditions  $\times$  2 types of noise maskers at 0 dB SNR. A total of 16 sentences, with 2 sentences in each condition, were used for the practice session. The practice stimuli were selected from two HINT lists that were not used in the actual testing. Feedback was provided in the form of written text of the target sentence shown on the computer screen. In the actual testing, the four voice conditions were randomized first in which the masker type and SNR level were randomized. The 10 sentences in each list were also randomized. Upon listening to each stimulus file, listeners were asked to type what they heard in a textbox shown on the computer screen. To ensure optimal performance, listeners could adjust the volume to their most comfortable level and repeat each sentence stimulus up to three times.

The perception responses were scored by a native English speaker. The accuracy of each sentence list was calculated by dividing the total number of correctly recognized words by the total number of words in all sentences. Given that the listeners were all non-native listeners and the research goal was to test auditory perception rather than grammatical knowledge and processing, language errors (e.g., subject-verb disagreement, misuse of verb tense) due to their native language interference and obvious spelling errors (e.g., ‘‘strewberry’’ for strawberry) were accepted and counted as correct.

*Statistical analysis.* The listeners’ perception accuracy data was analyzed using a Generalized Linear Mixed Model (GLMM). The model was fit using a binomial distribution and a logit link function. The factors of masker type, SNR, and voice condition were defined as fixed effects with treatment coding applied for each factor, encompassing all two-way and three-way interactions. Listeners were defined as the random effect. Subsequently, GLMM was applied to the perception accuracy in SSN and FTB separately to further examine the effect of voice condition in each type of masker. In each model, the factors of SNR and voice condition were set as fixed effects with SNR by voice condition interaction effect included. Treatment coding was applied for each factor. Listeners were set as the random effect.

We calculated the performance difference between the  $f_o$ -VTR matched and mismatched conditions to quantify the intelligibility difference caused by voice features. Given that the non-native listeners performed very poorly in certain conditions (e.g.,  $-3$  dB SNR in FTB), the percentage accuracy data was converted into rationalized arcsine units (RAU, [27]) scores first. Then, we calculated the performance difference in RAU between the two matched and two mismatched conditions:  $L_{f_o}L_{VTR}-L_{f_o}H_{VTR}$ ,  $L_{f_o}L_{VTR}-H_{f_o}L_{VTR}$ ,  $H_{f_o}H_{VTR}-L_{f_o}H_{VTR}$ , and

$H_{f_o}H_{VTR}-H_{f_o}L_{VTR}$ . The matched-mismatched performance difference in RAU was analyzed using Linear Mixed-effects Model (LMM) in which the performance difference type, masker type, and SNR were defined as the fixed effects and listeners were defined as the random effect. To further investigate how the two types of maskers influence the performance difference between matched and mismatched voices in different SNRs, the performance difference data was fitted with LMM for each SNR separately. In each model, the fixed effects included performance difference type, masker type, and the two-way interaction, and the random effect included listeners. Listener age or gender were not included in the GLMM or LMM models.

Given that the non-native listeners, though were all advanced English learners, varied in the vocabulary test scores, we conducted a correlational analysis between the non-native listeners' English language test results and the recognition performance. An average recognition accuracy across all tested conditions was calculated as the overall recognition accuracy for each listener. Bivariate correlation was run between LexTALE and the overall accuracy as well as between VLT and the overall accuracy.

### 3 Results

Figure 1 displays the recognition accuracy of the four voice conditions in SSN and FTB at the three SNRs. For both masker types, the recognition accuracy increased as the SNR increased. Of the two masker types at the same SNR, the recognition accuracy was lower in FTB than in SSN. Among the four types of voice conditions, the accuracy of  $H_{f_o}L_{VTR}$  was consistently lower than the other conditions in both masker types across all SNRs. The performance of  $H_{f_o}H_{VTR}$  and  $L_{f_o}L_{VTR}$  that had matched  $f_o$  and VTR was similar at all three SNRs for both masker types. At certain SNRs, such as +3 dB in SSN and 0 dB in FTB, the accuracy of  $L_{f_o}H_{VTR}$  was lower than the two matched conditions.

Model comparison for the GLMM on the perception accuracy data revealed that the best-fit model included by-subject random intercept and random slopes for all three factors. The GLMM results showed significant effects of masker type ( $F(1, 600) = 248.6, p < 0.001$ ), SNR ( $F(2, 600) = 609.3, p < 0.001$ ) and voice condition ( $F(3, 600) = 22.5, p < 0.001$ ). All two-way interactions and the three-way interaction were significant ( $p < 0.05$ ). The pairwise comparison of the four voice conditions revealed that the perception accuracy of  $H_{f_o}L_{VTR}$  was significantly lower than the other three conditions. In addition, the perception accuracy of  $L_{f_o}H_{VTR}$  was significantly lower than  $H_{f_o}H_{VTR}$ .

In the GLMM for each masker type, the by-subject random intercept and random slopes for SNR and voice condition were included in the models. The results revealed that for both SSN and FTB, there were significant effects of SNR and voice condition and significant SNR by voice condition interaction (all  $p < 0.001$ ).

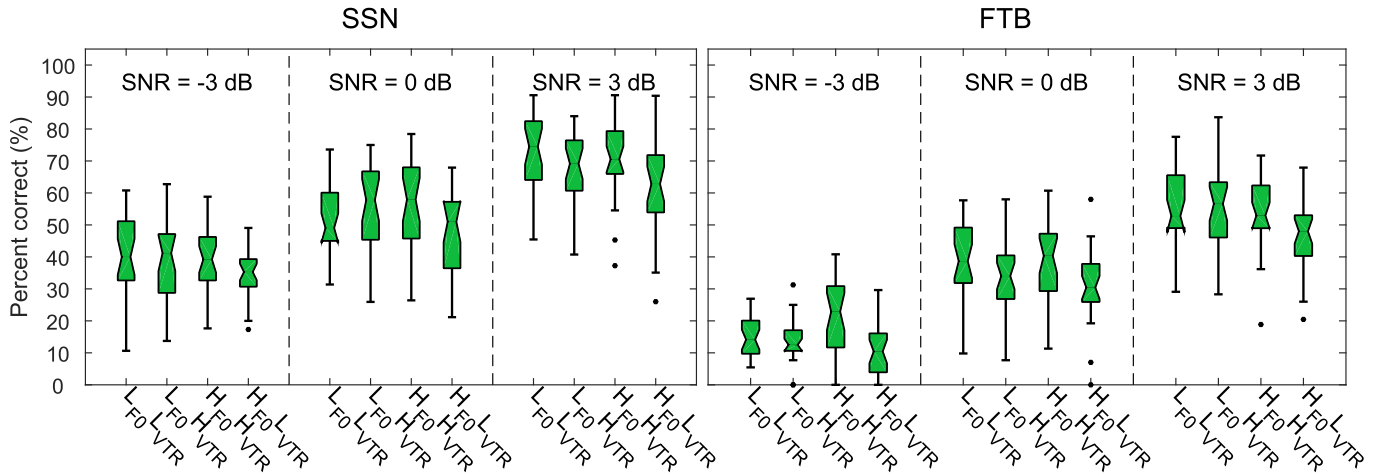
The pairwise comparison of voice condition revealed that the accuracy of  $H_{f_o}L_{VTR}$  was significantly lower than the other three conditions in both noise types. There was no significant difference among the  $H_{f_o}H_{VTR}$ ,  $L_{f_o}L_{VTR}$ , and  $L_{f_o}H_{VTR}$  ( $p > 0.05$ ) in SSN. By contrast, the accuracy of  $H_{f_o}H_{VTR}$  was significantly higher than the other three conditions ( $p < 0.05$ ) in FTB.

Figure 2 displays the performance difference between the  $f_o$ -VTR matched and mismatched conditions. In SSN, due to the consistently lower accuracy in  $H_{f_o}L_{VTR}$  than the other three conditions, the performance difference of  $L_{f_o}L_{VTR}-H_{f_o}L_{VTR}$  and  $H_{f_o}H_{VTR}-H_{f_o}L_{VTR}$  was greater than the performance difference of  $L_{f_o}L_{VTR}-L_{f_o}H_{VTR}$ , and  $H_{f_o}H_{VTR}-L_{f_o}H_{VTR}$  at all three SNRs. It is noteworthy that at 0 dB SNR, the performance of  $H_{f_o}H_{VTR}$  was higher than  $L_{f_o}L_{VTR}$ , which resulted in a greater difference of  $H_{f_o}H_{VTR}-H_{f_o}L_{VTR}$  than  $L_{f_o}L_{VTR}-H_{f_o}L_{VTR}$  as well as a greater difference in  $H_{f_o}H_{VTR}-L_{f_o}H_{VTR}$  than  $L_{f_o}L_{VTR}-L_{f_o}H_{VTR}$ . In FTB, the perception accuracy of  $H_{f_o}H_{VTR}$  was higher than the other voice conditions. In the meantime, the performance of  $H_{f_o}L_{VTR}$  was lower than the other conditions. Therefore, the performance difference was the greatest in  $H_{f_o}H_{VTR}-H_{f_o}L_{VTR}$  and the smallest in  $L_{f_o}L_{VTR}-L_{f_o}H_{VTR}$ . At 0 and +3 dB SNRs, the performance of the two matched conditions  $H_{f_o}H_{VTR}$  and  $L_{f_o}L_{VTR}$  was very similar but the performance of  $H_{f_o}L_{VTR}$  was lower than that of  $L_{f_o}H_{VTR}$ . Therefore, in these two SNRs, the performance difference of  $H_{f_o}H_{VTR}-H_{f_o}L_{VTR}$  and  $L_{f_o}L_{VTR}-H_{f_o}L_{VTR}$  was greater than that of  $H_{f_o}H_{VTR}-L_{f_o}H_{VTR}$  and  $L_{f_o}L_{VTR}-L_{f_o}H_{VTR}$ . The LMM analysis on the matched-mismatched performance difference revealed a significant effect of performance difference type ( $F(3, 575.01) = 11.4, p < 0.001$ ) and masker by SNR interaction ( $F(2, 575.01) = 9.0, p < 0.001$ ). The subsequent LMM for matched-mismatched performance difference between the two masker types at individual SNR revealed significant differences between SSN and FTB at -3 dB ( $F(1, 175) = 11.19, p = 0.001$ ) and +3 dB ( $F(1, 175) = 7.90, p = 0.005$ ) SNR. At -3 dB SNR, the average performance difference between matched and mismatched voice conditions was 3.59 RAUs in SSN and 7.04 RAUs in FTB. By contrast, at +3 dB, the average performance difference was 7.99 RAUs in SSN and 3.61 RAUs in FTB.

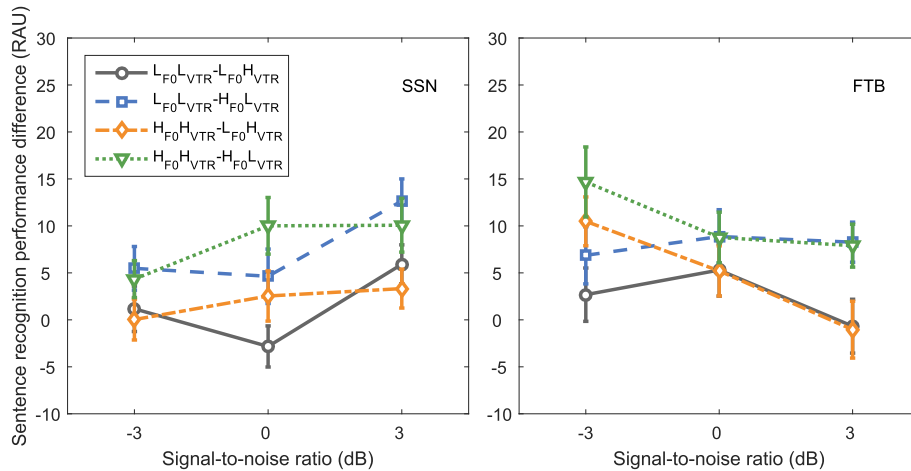
The correlation analysis revealed no significant relationship between the listeners' vocabulary test scores and their overall perception performance.

### 4 Discussion

The purpose of the present study was to examine the influence of two sex-related acoustic correlates,  $f_o$  and VTR, on speech intelligibility for non-native listeners in two types of maskers, SSN and FTB. We tested whether the sensitivity to the covariation of  $f_o$  and VTR reported in native listeners occurred in non-native English listeners. Our results revealed that in both



**Figure 1.** Box plots showing the perception accuracy of HINT sentences in speech-shaped noise (SSN) and four-talker babble (FTB) in four voice conditions ( $L_{f_0}LV_{TR}$ ,  $L_{f_0}H_{VTR}$ ,  $H_{f_0}H_{VTR}$ , and  $H_{f_0}LV_{TR}$ ) at three signal-to-noise ratios (SNRs) of  $-3$ ,  $0$ , and  $+3$  dB. In each condition, the box represents the 25th and 75th percentiles of the performance, the notch represents the median, and the whiskers represent the range. Outliers are plotted with filled symbols.



**Figure 2.** Sentence recognition performance difference (mean and standard error) between the matched ( $H_{f_0}H_{VTR}$  and  $L_{f_0}LV_{TR}$ ) and mismatched ( $H_{f_0}LV_{TR}$  and  $L_{f_0}H_{VTR}$ ) conditions at three signal-to-noise ratios of  $-3$ ,  $0$ ,  $+3$  dB in speech-shaped noise (SSN) and four-talker babble (FTB). RAU = rationalized arcsine unit.

types of maskers at all three SNRs, the recognition accuracy of the mismatched condition of  $H_{f_0}LV_{TR}$  was significantly lower than the other conditions. The performance of the other mismatched condition,  $L_{f_0}H_{VTR}$ , was lower than the performance of the two matched conditions in SSN at 3 dB and FTB at 0 dB SNR. However, this pattern was not consistently shown in other conditions. The GLMM did not yield a significant difference between  $L_{f_0}H_{VTR}$  and the two matched conditions when the data was collapsed across masker types and SNRs. These results suggested that non-native listeners also showed sensitivity to the covariation pattern of  $f_0$  and VTR. However, importantly, the adverse impact of mismatched voice features was mainly reflected in high  $f_0$  paired with low VTR. Assmann and Nearey (2008) [4] discussed that high  $f_0$  resulted in sparsely distributed harmonics and thus poorly resolved formant peaks, which

might have a greater adverse effect on speech recognition especially when the  $f_0$  was higher than the first formant of vowel sounds. In the study of Yang et al. (2025) [7], native English listeners also demonstrated the lowest recognition accuracy in  $H_{f_0}LV_{TR}$  condition. In the present study, the consistently lower accuracy in  $H_{f_0}LV_{TR}$  condition in both SSN and FTB at all SNRs provided additional evidence about the negative impact of this mismatched condition on speech recognition. This finding has clinical implications for gender diverse populations who may undergo surgical procedures, take medications, and/or receive voice training to alter voice pitch for gender affirmation. While voice pitch can be modified, the vocal tract can hardly be resized or reshaped, potentially resulting in mismatched voice features, as observed in the present study. Given the evidence showing decreased intelligibility with

mismatched voice features in the presence of noises, clinical efforts should be made to develop speech training programs or strategies to help these individuals improve speech intelligibility in adverse communication conditions.

The second central question focused on the differences in speech recognition performance in various voice conditions between SSN and FTB. The results revealed that our non-native listeners performed more poorly in FTB than in SSN at all tested SNRs in all four voice conditions. In Mi et al.'s (2013) [12] study, the authors tested English vowel recognition by English native (EN) listeners, Chinese native listeners recruited in the United States (CNU), and Chinese native listeners with no residency history in English-speaking countries (CNC). The authors found that CNU and CNC listeners performed similarly in SSN but the CNC listeners performed worse than CNU listeners in the 12-talker babble at all SNRs. Similarly, in the study of Tao et al. (2018) [28], which examined English consonant recognition by EN, CNU, and CNC listeners in quiet vs 12-talker babble, CNC listeners again showed poorer performance than CNU and EN listeners in multi-talker babble. The participants of the present study were all recruited in China and none of them had experience residing in English-speaking countries. It is possible that the listeners who were not immersed in English-speaking environment had limited exposure to continuous speech in English produced by multiple speakers, which makes speech babble noise particularly challenging for them [12, 28]. Additionally, prior research has shown that native Chinese listeners, especially those who had no residency history in English-speaking regions, were less able to utilize temporal dips and exhibited reduced masking release compared to native English listeners and native Chinese listeners with relatively long-term English immersion [29–33]. This might further explain the difficulty our Chinese listeners experienced in perceiving speech masked by multi-talker babble.

Further, we calculated the performance difference between the two matched and two mismatched voice conditions to quantify the adverse effect of mismatched voice condition. Our results revealed that the performance difference between the matched and mismatched conditions in FTB was greater than SSN at  $-3$  dB SNR but lower than SSN at  $+3$  dB. This result did not fully support our hypothesis of less negative effect of mismatched voice in speech babbles than in SSN. Close examination of the performance difference data revealed that  $L_{f_o}LV_{TR} - H_{f_o}LV_{TR}$  and  $L_{f_o}LV_{TR} - L_{f_o}HV_{TR}$  in FTB was similar to those in SSN at  $-3$  dB SNR. However, because the performance of  $H_{f_o}HV_{TR}$  was much higher than all other conditions in FTB at  $-3$  dB SNR, the performance difference between  $H_{f_o}HV_{TR}$  and the two mismatched conditions was greater in FTB than in SSN.

We analyzed the  $f_o$  of the four talkers of the babble maskers and the target speech signals. The average  $f_o$  values for the two female babble talkers was 234 Hz and 158 Hz, while those for the two male babble talkers, were

116 Hz and 121 Hz. The target speech signals had an average  $f_o$  of 110 Hz for the original male voice and 220 Hz for the  $H_{f_o}$  condition. While there were two babble talkers with similar  $f_o$  as the target signal in  $L_{f_o}$  condition, only one babble talker had a similar  $f_o$  as the target signal in  $H_{f_o}$  condition. This suggests that the masking effect of the speech babble might be greater for the  $L_{f_o}$  condition than for the  $H_{f_o}$  condition. Therefore, the higher performance of  $H_{f_o}HV_{TR}$  as well as the greater performance difference between the  $H_{f_o}HV_{TR}$  and the two mismatched conditions in FTB compared to SSN should be interpreted with caution. Future studies should aim to more carefully control voice characteristics of the babble talkers.

In the meantime, the relatively higher recognition accuracy in  $H_{f_o}HV_{TR}$  condition also warrants further discussion about the potential intelligibility difference between male and female voices. Previous studies reported inconsistent findings on this issue [34–37]. While those studies recruited multiple male and female speakers for whom researchers could hardly strictly control speaker characteristics that may interfere with speech intelligibility, we manipulated the two major sex-related acoustic correlates of one single speaker and the shifted ratios represented the group-level sex differences between males and females. The results of our previous study revealed no significant difference in speech recognition between  $H_{f_o}HV_{TR}$  and  $L_{f_o}LV_{TR}$  presented with SSN in native English listeners [7]. In the current study, non-native listeners presented no significant difference between these two voice conditions at all SNRs in SSN. However, the two voice conditions showed a significant difference in FTB, which was mainly reflected at  $-3$  dB SNR. Other than the potentially reduced masking effect on  $H_{f_o}$  signals due to the fewer babble talkers with similar  $f_o$ , it remains unclear whether female voice presents more salient features that could help listeners overcome the masking effect in challenging conditions with multi-talker babble. Future studies should aim to better control the  $f_o$  of the babble talkers and test more SNRs.

Finally, divergent from the findings of positive relationship between non-native listeners' language proficiency and perception performance reported in many previous studies [38–40], the correlational analysis of the present study yielded no significant relationship between the vocabulary test scores and the overall recognition accuracy. The test stimuli were HINT sentences that include high frequency words occurring in everyday life. The group of listeners demonstrated a high level of proficiency on the language tests. The group average score for the vocabulary knowledge test (Lex-TALE) was higher than the reported score for advanced English learners. For the vocabulary size test (VLT), the group average was 95 out of 100 for the 2000 and 3000-level words. The high scores of the English tests suggested that the vocabulary difficulty of the sentence stimuli was within the listeners' language ability. The variation of the listeners' recognition performance was less likely caused by their vocabulary

proficiency. Previous studies have shown that working memory capacity played an important role in speech recognition in adverse conditions (e.g., speech in noise or interrupted speech) for both native and non-native listeners [40–43]. It is possible that the non-native listeners' working memory abilities may partially account for the individual differences in their perceptual performance.

In sum, our results revealed that non-native listeners performed more poorly with  $f_o$ -VTR mismatched voice than with  $f_o$ -VTR matched voice. However, the negative influence of mismatched voice features was mainly manifested in the  $H_{f_o}L_{VTR}$  condition. Compared to SSN, FTB had a greater effect on the non-native listeners who had no residency experience in English-speaking countries/regions. Further, the performance difference between matched and mismatched conditions showed different patterns in speech-shaped noise and multi-talker babble. These findings suggest that when developing hearing test stimuli, we should take into account the potential influence of speaker's voice features on listeners' recognition performance. When testing speech-in-noise perception, various noise types should be used to provide a thorough assessment of listeners' perceptual performance and recognition ability. It is noteworthy that the present study only adopted one male speaker and the manipulation of  $f_o$  and VTR were all based on this speaker. While the purpose of this approach was to strictly control the potential confounding factors involving speaker-specific traits that likely introduced individual variabilities to speech intelligibility, applying algorithms to manipulate the voice features of one single speaker and resynthesize speech might restrict the generalizability of the current findings and undermine the ecological validity of the study. The naturalness of the manipulated voice conditions, especially the mismatched conditions, might be compromised, which could potentially affect the testing results. Further, the present study adopted one-step upward shifting for both features; for future studies, we should adopt female speakers and apply downward shifting to test if listeners show similar results to the current research. We should also modify the shifting ratios beyond the range of natural speech as was done in Assmann and Nearey (2008) [4] to test whether the influence of  $f_o$ -VTR covariation on speech intelligibility by non-native listeners is valid when the voice features exceed natural speech range. We also acknowledged that the female listeners recruited in this study outnumbered male listeners. Though efforts were made to recruit an equivalent number of male and female listeners, the perceptual task of the present study was challenging and required high proficiency in English; therefore, most participants were recruited from English or other social science majors that typically have a higher proportion of female students. The influence of listener's sex on speech recognition is an insightful research question that has been addressed in a few recent studies [37, 44]. For future studies, we should recruit a balanced number of listeners in each sex to examine the effect of listener sex on the intelligibility of speech presenting various voice features.

### Conflicts of interest

The authors declare no conflict of interest.

### Data availability statement

The sentence stimuli generated and/or the perceptual data are available on request from the authors.

### References

1. J.M. Hillenbrand, M.J. Clark: The role of  $f_0$  and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics* 71 (2009) 1150–1166.
2. C.D. Fuller, E. Gaudrain, J.N. Clarke, J.J. Galvin, Q.-J. Fu, R.H. Free, D. Baskent: Gender categorization is abnormal in cochlear implant users. *Journal of The Association for Research in Otolaryngology* 15 (2014) 1037–1048.
3. M.S. Poon, M.L. Ng: The role of fundamental frequency and formants in voice gender identification. *Speech, Language and Hearing* 18 (2015) 161–165.
4. P.F. Assmann, T.M. Nearey: Identification of frequency-shifted vowels. *Journal of the Acoustical Society of America* 124 (2008) 3203–3212.
5. M.D. Vestergaard, N.R. Fyson, R.D. Patterson: The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America* 125 (2009) 1114–1124.
6. E. Holmes, Y. Domingo, I.S. Johnsrude: Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science* 29 (2018) 1575–1583.
7. J. Yang, X. Wang, V. Costa, L. Xu: Effects of fundamental frequency and vocal tract resonance on sentence recognition in noise. *Journal of Speech, Language, and Hearing Research* 68 (2025) 3011–3022.
8. X. Wang, L. Xu: English vowel recognition in multi-talker babbles mixed with different numbers of talkers. *JASA Express Letters* 4 (2024) 045202.
9. M.L. Lecumberri, M. Cooke: Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America* 119 (2006) 2445–2454.
10. M. Broersma, O. Scharenborg: Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication* 52 (2010) 980–995.
11. S.H. Jin, 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* 132 (2012) EL391–EL397.
12. L. Mi, S. Tao, W. Wang, Q. Dong, S.H. Jin, C. Liu: English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners. *The Journal of the Acoustical Society of America* 133 (2013) EL391–EL397.
13. L. Zhong, C. Liu, S. Tao: Sentence recognition for native and non-native English listeners in quiet and babble: effects of contextual cues. *The Journal of the Acoustical Society of America* 145 (2019) EL297–EL302.
14. D.S. Brungart: Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109 (2001) 1101–1109.
15. C.J. Darwin, D.S. Brungart, B.D. Simpson: Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The*

- Journal of the Acoustical Society of America 114 (2003) 2913–2922.
16. H.E. Cullington, F.G. Zeng: Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *The Journal of the Acoustical Society of America* 123 (2008) 450–461.
  17. G. Kidd, C.R. Mason, J. Swaminathan, E. Roverud, K.K. Clayton, V. Best: Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America* 140 (2016) 132–144.
  18. J. Zhang, X. Wang, N.Y. Wang, X. Fu, T. Gan, J.J. Galvin III, S. Willis, K. Xu, M. Thomas, Q.J. Fu: Tonal language speakers are better able to segregate competing speech according to talker sex differences. *Journal of Speech, Language, and Hearing Research* 63 (2020) 2801–2810.
  19. M.M. Flaherty, E. Buss, L.J. Leibold: Independent and combined effects of fundamental frequency and vocal tract length differences for school-age children’s sentence recognition in a two-talker masker. *Journal of Speech, Language, and Hearing Research* 64, 1 (2021) 206–217.
  20. I.S.P. Nation: *Teaching and Learning Vocabulary*. Newbury House, New York, 1990.
  21. K. Lemhöfer, M. Broersma: Introducing LexTALE: a quick and valid lexical test for advanced learners of English. *Behavior Research Methods* 44 (2012) 325–343.
  22. A. Coxhead: A new academic word list. *TESOL Quarterly* 34 (2000) 213–238.
  23. M. Nilsson, S.D. Soli, J.A. Sullivan: Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America* 95, 2 (1994) 1085–1099.
  24. P. Boersma, D. Weenink: Praat: doing phonetics by computer (Version 5.4.04) [Computer software]. 2020 Retrieved from <http://www.praat.org/>.
  25. S.A. Simpson, M. Cooke: Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America* 118 (2005) 2775–2778.
  26. S. Rosen, P. Souza, C. Ekelund, A.A. Majeed: Listening to speech in a background of other talkers: effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America* 133 (2013) 2431–2443.
  27. G.A. Studebaker: A “rationalized” arcsine transform. *Journal of Speech, Language, and Hearing Research* 28 (1985) 455–462.
  28. S. Tao, Y. Chen, W. Wang, Q. Dong, S.H. Jin, C. Liu: English consonant identification in multi-talker babble: effects of Chinese-native listeners’ English experience. *Language and Speech* 62, 3 (2019) 531–545.
  29. K.J. Van Engen: Similarity and familiarity: second language sentence recognition in first-and second-language multi-talker babble. *Speech Communication* 52, 11, 12 (2010) 943–953.
  30. A. Stuart, J. Zhang, 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* 21, 04 (2010) 239–248.
  31. M. Li, W. Wang, S. Tao, Q. Dong, J. Guan, C. Liu: Mandarin Chinese vowel-plus-tone identification in noise: effects of language experience. *Hearing Research* 331 (2016) 109–118.
  32. J. Guan, X. Cao, C. Liu: Second language experience facilitates sentence recognition in temporally-modulated noise for non-native listeners. *Frontiers in Psychology* 12 (2021) 631060.
  33. H. Yin, C. Liu: Effect of listeners’ language backgrounds on pure tone detection in temporally modulated noise. *JASA Express Letters* 1 (2021) 094402.
  34. A.R. Bradlow, G.M. Torretta, D.B. Pisoni: Intelligibility of normal speech I: global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20 (1996) 255–272.
  35. D. Markham, V. Hazan: The effect of talker-and listener-related factors on intelligibility for a real-word, open-set perception test. *Journal of Speech Language and Hearing Research* 47 (2004) 725–737.
  36. D.R. McCloy, R.A. Wright, P.E. Souza: Talker versus dialect effects on speech intelligibility: a symmetrical study. *Language and Speech* 58 (2015) 371–386.
  37. S.E. Yoho, S.A. Borrie, T.S. Barrett, D.B. Whittaker: Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception, & Psychophysics* 81 (2019) 558–570.
  38. R.L. Bundgaard-Nielsen, C.T. Best, M.D. Tyler: Vocabulary size matters: the assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics* 32 (2011) 51–67.
  39. L. Kilman, A. Zekveld, M. Hällgren, J. Rönnerberg: The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology* 5 (2014) 651.
  40. J. Yang, N. Nagaraj, B. Magimairaj: Audiovisual perception of interrupted speech by non-native listeners. *Attention, Perception, & Psychophysics* 86 (2024) 1763–1776.
  41. C. Foo, M. Rudner, J. Rönnerberg, T. Lunner: Recognition of speech in noise with new hearing instrument compression release settings requires explicit cognitive storage and processing capacity. *Journal of the American Academy of Audiology* 18 (2007) 618–631.
  42. T. Lunner, E. Sundewall-Thorén: Interactions between cognition, compression, and listening condition: effects on speech-in-noise performance in a two-channel hearing aid. *Journal of the American Academy of Audiology* 18, 7 (2007) 604–617.
  43. M. Rudner, J. Rönnerberg, T. Lunner: Working memory supports listening in noise for persons with hearing impairment. *Journal of the American Academy of Audiology* 22, 3 (2011) 156–167.
  44. W.K. Yumba: Influences of listener gender and working memory capacity on speech recognition in noise for hearing aid users. *Speech, Language and Hearing* 25, 2 (2022) 112–124.