ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Effects of Fundamental Frequency and Vocal Tract Resonance on Sentence Recognition in Noise

Jing Yang,[a] Xianhui Wang,[b,c] Victoria Costa,[b] and Li Xu[b]

[a] Communication Sciences & Disorders Program, University of Wisconsin–Milwaukee [b] Department of Hearing, Speech and Language Sciences, Ohio University, Athens [c] Center for Hearing Research, Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology—Head and Neck Surgery, University of California, Irvine

ARTICLE INFO

Article History:
Received October 29, 2024
Revision received February 17, 2025
Accepted March 6, 2025

Editor-in-Chief: Rachael Frush Holt
Editor: Chang Liu

https://doi.org/10.1044/2025_JSLHR-24-00758

ABSTRACT

Purpose: This study examined the effects of change in a talker's sex-related acoustic properties (fundamental frequency [*F*0] and vocal tract resonance [VTR]) on speech recognition in noise.
Method: The stimuli were Hearing in Noise Test sentences, with the *F*0 and VTR of the original male talker manipulated into four conditions: low *F*0 and low VTR ($L_{F0}L_{VTR}$; i.e., the original recordings), low *F*0 and high VTR ($L_{F0}H_{VTR}$), high *F*0 and high VTR ($H_{F0}H_{VTR}$), and high *F*0 and low VTR ($H_{F0}L_{VTR}$). The listeners were 42 English-speaking, normal-hearing adults (21–31 years old). The sentences mixed with speech spectrum–shaped noise at various signal-to-noise ratios (i.e., −10, −5, 0, and +5 dB) were presented to the listeners for recognition.
Results: The results revealed no significant differences between the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ conditions in sentence recognition performance and the estimated speech reception thresholds (SRTs). However, in the $H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$ conditions, the recognition performance was reduced, and the listeners showed significantly higher SRTs relative to those in the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ conditions.
Conclusion: These findings indicate that male and female voices with matched *F*0 and VTR (e.g., $L_{F0}L_{VTR}$ and $H_{F0}H_{VTR}$) yield equivalent speech recognition in noise, whereas voices with mismatched *F*0 and VTR may reduce intelligibility in noisy environments.
Supplemental Material: https://doi.org/10.23641/asha.29052305

Speech intelligibility can be affected by various aspects of talker characteristics, including a talker's speaking rate (Krause & Braida, 2002; Tanaka et al., 2011), speaking style/mode (Bradlow & Bent, 2002; Krause & Braida, 2002; Uchanski et al., 1996), age (Hazan et al., 2018; Smiljanic & Gilbert, 2017), dialect/accent (Munro & Derwing, 1995; Yang et al., 2023), sex/gender (Bradlow et al., 1996; Hazan & Markham, 2004; McCloy et al., 2015; Yoho et al., 2019), and so forth. Among these factors, talker sex/gender is an important source of talker variation that affects voice quality and speech characteristics in both segmental and suprasegmental aspects (Bradlow et al., 1996; Klatt & Klatt, 1990; Munson & Babel, 2019). In this study, we follow the widely acknowledged definitions of *sex* as a biological attribute assigned at birth and *gender* as a social or cultural construct. Gendered speech involves multifaceted variations that encompass talkers' physiological differences, linguistic behaviors, and social positions (Munson & Babel, 2019; Tripp & Munson, 2022). Sex differences in speech anatomy and physiology play an important role in characterizing gendered speech. The anatomically grounded sex differences in speech are salient for listeners to identify (Bachorowski & Owren, 1999; Jacewicz et al., 2023; Weirich & Simpson, 2018). However, there has been no consensus regarding whether and how male and female voices differ in speech intelligibility.

In 1953, Silverstein et al. examined intelligibility differences between male and female speakers over standard military communication equipment in high-level noise. They found that male speakers showed higher intelligibility in the pretraining test. After speech training focusing on loudness and clear pronunciation, the female and male

Correspondence to Li Xu: xul@ohio.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

speakers showed no significant intelligibility differences in the posttraining test. Stevens et al. (2005) compared the intelligibility of male and female voice text-to-speech synthesis and reported that the male voice was more intelligible than the female voice. Robinson (2011) compared the perception results of vowel identification and word recognition by pediatric hearing aid users. The speech materials were recorded from adult males, adult females, 10- to 12-year-old girls, and 5- to 7-year-old girls. The results showed significantly higher recognition accuracy for the adult male talkers than for the three groups of female talkers. McCloy et al. (2015) also found significantly higher intelligibility for male talkers than for female talkers, after the speech features of intensity, pitch, and vowel space were accounted for.

Contrary to the male-talker advantage in speech intelligibility, many other studies revealed higher intelligibility for female talkers. Bradlow et al. (1996) examined intelligibility of 2,000 sentences produced by 20 talkers presented in quiet condition. Each talker was transcribed by 10 listeners. The intelligibility scores showed that female talkers, as a group, were more intelligible than male talkers. Markham and Hazan (2004) compared the intelligibility of speech samples recorded from adult males, adult females, teenage boys, and teenage girls presented with 20-talker babbles at +6 dB signal-to-noise ratio (SNR). The listeners included adults, older children, and younger children. The results demonstrated that women talkers were significantly more intelligible than the other three talker groups for all listener groups. Ferguson (2004) examined the increase in vowel intelligibility with the change in speaking style in both quiet and noise conditions in male and female talkers. The author found that female talkers had significantly higher vowel intelligibility and a larger clear speech effect relative to male talkers in clear speech. In another study by Kwon (2010), 20 talkers, including 10 men and 10 women, were recorded reading a passage that was judged by three trained speech pathologists for intelligibility scores on a 10-point scale. The rating scores revealed significantly higher intelligibility for women talkers in comparison to men talkers. In a more recent study, Yoho et al. (2019) compared the perception and intelligibility rating of IEEE sentences produced by five male talkers and five female talkers. The sentence stimuli were presented to two groups of listeners. One group (n = 50) was instructed to repeat the sentence stimuli to assess the percentage of correctly repeated words. The other group (n = 118) was recruited through crowdsourcing for intelligibility rating on a 7-point scale. The results showed that regardless of the methodological difference in data collection and intelligibility measurement, the stimuli from female talkers were generally more intelligible relative to those from male talkers. Oh et al. (2022, 2023) examined the benefit of multisensory integration, including visual and tactile cues for speech recognition in noise. In Oh et al.'s (2022) study, the speech stimuli were Harvard sentences produced by a male and a female native English speaker. In another Oh et al. (2023) study, the speech stimuli were drawn from the Coordinate Response Measure speech corpus spoken by four male and four female talkers. In both studies, the authors found a greater female-talker benefit. That is, listeners showed a greater improvement when multimodal cues were temporally integrated as compared to auditory-only stimuli, for female talkers relative to male talkers.

Researchers sought to identify possible global and fine-grained acoustic–phonetic features—such as pitch variation, vowel space area size, clarity of articulation, speaking rate, and total energy in the 1- to 3-kHz range—that might explain group-level differences in speech intelligibility between male and female voices (Bradlow et al., 1996; Byrd, 1994; Hazan & Markham, 2004; Yoho et al., 2019). Yet, no consistent acoustic–phonetic correlates could be identified to account for the potential sex-related differences in speech intelligibility. In fact, many researchers found no significant difference in speech intelligibility between male and female talkers. Tielen (1989) compared the consonant–vowel–consonant and phoneme intelligibility between male and female speakers, which showed no significant difference between the two sexes. Nixon, Morris, et al. (1998) compared the intelligibility of male and female speech in high levels of aircraft cockpit noise ranging from 95 to 115 dB overall SPL. The authors found no significant differences except at the highest level of cockpit noise in which the female voice was less intelligible than the male voice. However, when the speech stimuli were changed to vocoded and recognized by automatic speech recognition systems, no intelligibility difference between the vocoded male and female speech was found (Nixon, Anderson, et al., 1998).

Although the intelligibility difference between female and male voices might be unclear at the group level, variabilities in speech intelligibility among talkers were evident from many studies. Gengel and Kupperman (1980) recorded CID W-22 words produced by three female and three male talkers that were presented in noise at SNRs of −3, +1, and +5 dB to 42 normal-hearing listeners. The six talkers showed different intelligibility scores, but the difference was not related to talker sex. Hazan and Markham (2004) found that the relative intelligibility of individual talkers was highly consistent across listeners of different ages. The previously cited studies that yielded contradictory findings regarding the effects of talker sex on speech intelligibility all included multiple talkers. The talker-specific traits might act as confounding factors to

interfere with talker sex. In the present study, we utilized an alternative approach to strictly control the potential confounding factors that might introduce individual variabilities to speech intelligibility. Instead of including multiple talkers for each sex, we used one single talker but manipulated the two major sex-related acoustic correlates to examine the influence of voice features on speech intelligibility in noise.

According to the source filter theory (Fant, 1960), speech production (e.g., vowel production) involves two major mechanisms: vocal fold vibration that generates the vocal source and the vocal tract that filters the vocal source. Anatomically based sex differences affect both source (fundamental frequency [*F*0]) and filter (vocal tract resonance [VTR]) components (Smith & Patterson, 2005), and both acoustic correlates contribute to the identification of talker sex (Fuller et al., 2014; Gaudrain et al., 2009; Hillenbrand & Clark, 2009; Poon & Ng, 2015; Skuk & Schweinberger, 2014). The rate of vocal fold vibration (*F*0) is determined by properties such as stiffness, mass, and length of the vocal folds, which, according to some studies, played a primary role in talker sex identification (Gelfer & Mikos, 2005; Poon & Ng, 2015; Whiteside, 1998a, 1998b). The VTR also contains talker sex information because male and female talkers differ in the average vocal tract length that determines the formant locations (Bachorowski & Owren, 1999; Gelfer & Bennett, 2013; Smith & Patterson, 2005). Generally speaking, female talkers as a group tend to have thinner and shorter vocal folds and shorter vocal tract length, which generate higher *F*0 and higher VTR, as compared to male talkers as a group. We acknowledge that within-sex variability in vocal fold properties and vocal tract size may lead to low *F*0 and VTR for individual female talkers and high *F*0 and VTR for individual male talkers. Also, a given talker can manipulate vocal folds and vocal tract length for different vocal modes. In the present study, we focused on the physiological-based, sex-related voice features of *F*0 and VTR for overall male and female speakers.

So far, a few studies adopted a similar approach of manipulating *F*0 and/or VTR to assess the influence of a talker's voice features on speech recognition (Assmann & Nearey, 2008; Darwin et al., 2003; Holmes et al., 2018; Vestergaard et al., 2009). For example, Assmann and Nearey (2008) examined the effects of *F*0 and formant frequency on the recognition of frequency-shifted vowels synthesized through vocoders. In their first experiment, the authors created resynthesized /hVd/ words for vowel recognition from adult male speakers by shifting the spectrum envelope (formant frequency) at five levels of 1.0, 1.25, 1.5, 1.75, and 2 and/or, for *F*0, by shifting up at three levels of 1.0, 2.0, and 4.0. In the following experiment, the authors modified the settings for the spectrum envelope and *F*0 by applying both downward and upward shifting (shifting by a factor of 0.6, 0.8, 1.0, 1.5, or 2.0 for formant frequency and by a factor of 0.5, 1.0, or 4.0 for *F*0). In addition to male talkers, the authors added female and child talkers and implemented similar frequency-shifting procedures. The results revealed that the recognition accuracy dropped when the spectrum envelope or *F*0 was shifted independently. However, when the *F*0 and spectrum envelope were shifted in the same direction, the recognition accuracy improved. Finally, the authors extended the shifting scale factors to the range outside of natural speech and observed similar results. Compared to the study by Assmann and Nearey (2008) that tested only vowel recognition and included voice conditions beyond natural speech, the present study examined sentence recognition in noise, with voice features manipulated to reflect average anatomical differences between males and females.

In a more recent study, Holmes et al. (2018) manipulated the acoustic correlates *F*0 and VTR of 22 talkers and tested sentence recognition using a closed-set task. The 22 talkers included seven males and 15 females who formed 11 pairs, including opposite- and same-sex pairs. Each participant produced 480 sentences that were processed with the *F*0 and/or VTR shifted (*F*0 shifted up by 40% and VTR shifted up by 27%), which generated four conditions: *F*0 shifted, VTR shifted, both shifted, and unshifted. Although the main purpose was to investigate the effect of *F*0 and VTR change on the perception of voice familiarity and the intelligibility benefit from familiar voice as compared to unfamiliar voice in the presence of a competing talker at various target-to-masker ratios, the authors examined the influence of the manipulations on the recognition performance for the male and female voices. The analysis revealed decreased speech intelligibility in the three shifted conditions compared to the unshifted condition for both familiar and unfamiliar voices. For the speech intelligibility benefit (intelligibility difference between familiar and unfamiliar voices), there was neither a significant difference between male and female talkers nor a significant interaction effect between voice gender and manipulation. In a follow-up study, Holmes and Johnsrude (2023) presented stimuli in three voice manipulation conditions: unshifted, *F*0 manipulated, and VTR manipulated. In this study, the *F*0 and VTR were manipulated to match individual participants' pitch or formant spacing discrimination thresholds. The results of the speech intelligibility task revealed consistently higher recognition accuracies with familiar voices than with unfamiliar voices. However, the three manipulation conditions showed no significant difference.

The studies discussed above reported inconsistent outcomes regarding whether and how the manipulation of

the two acoustic correlates of voice influences speech intelligibility, highlighting the need for further investigation on this topic. In the present study, we adopted the original recording of the Hearing in Noise Test (HINT) sentences (Nilsson et al., 1994) produced by a male native English speaker and applied algorithms to modify the *F*0 and/or VTR of the original speaker to generate different voice conditions. The stimuli were presented with speech spectrum–shaped noise (SSN) at various SNRs. The research purpose is twofold: (a) to examine whether female and male voices differ in speech intelligibility in noise and (b) to identify which acoustic attributes, or combinations thereof, influence a talker's speech intelligibility. The rationale of using one single speaker and manipulating the voice features to generate modified speech stimuli was to rule out the interference of other talker-specific variables such as speech rate, speaking style, age, dialect/accent, and so forth.

Method

Listeners

The participants included 42 native English-speaking, normal-hearing young adults (20 men, 22 women). The subjects were between the ages of 21 and 31 years ($M \pm SD = 23.3 \pm 2.2$). They were screened for normal hearing ($\leq$ 20 dB HL) at octave frequencies between 250 and 8000 Hz. None of the participants reported having any speech-language or cognitive problems. The use of human subjects was reviewed and approved by the institutional review boards (IRBs) of Ohio University (Protocol Nos. 15X061 and IRB-FY25-110) and the University of Wisconsin–Milwaukee (IRB No. 25.048). All participants completed the consent process before starting the experiment.

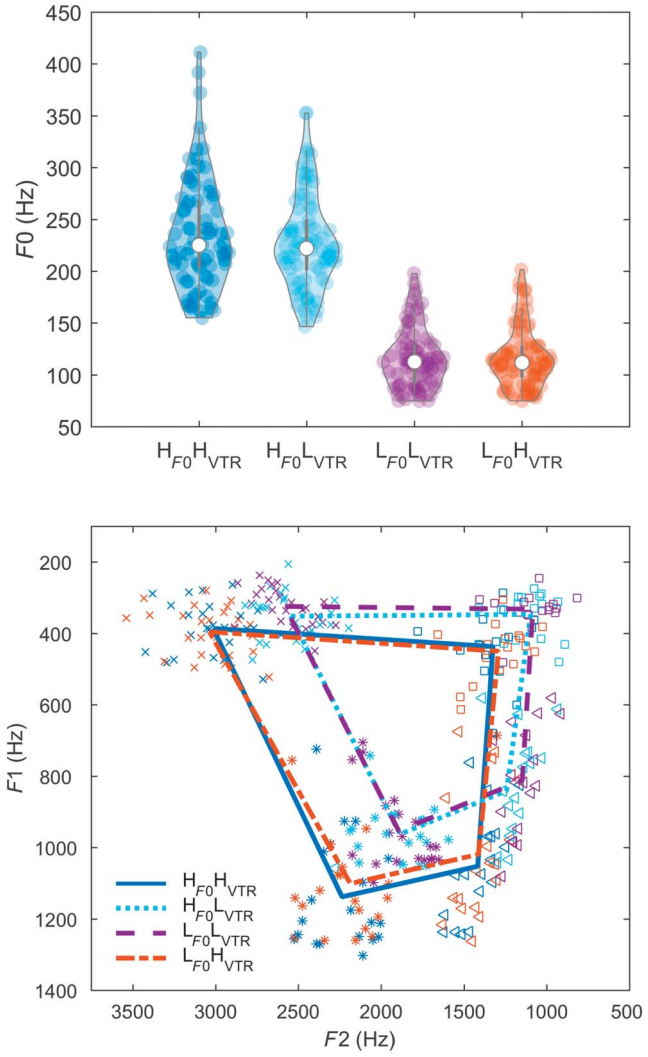
Stimuli and Signal Processing

The perception stimuli were HINT sentences, which consist of 26 lists, each containing 10 sentences. The length of individual sentences varies from four to seven words. The original HINT sentences were produced by a male talker. In the current study, we manipulated the *F*0 and/or the formants of the original sentences to change the perceptual voice gender using a built-in "Change Gender" function in the Praat software (Boersma & Weenink, 2020). One set of stimuli was generated by doubling the *F*0 of the original male voice (i.e., from a mean of 110 Hz to a mean of 220 Hz) and scaling up the formants (the formant spacing for all formants was changed accordingly) of the male speaker by a factor of 1.2 (Poon & Ng, 2015), which resulted in a high *F*0 and high VTR ($H_{F0}H_{VTR}$) condition. These ratios of *F*0 and VTR used in our manipulations were in the physiological range of male and female voices (Poon & Ng, 2015). One set of stimuli was generated by doubling the *F*0 of the original male voice but maintaining the formant values of the original male vowels, which resulted in a high *F*0 and low VTR ($H_{F0}L_{VTR}$) condition. The last set of stimuli was generated by maintaining the original male *F*0 but scaling up the formants of the male speaker by a factor of 1.2, which resulted in a low *F*0 and high VTR ($L_{F0}H_{VTR}$) condition. For all three manipulated conditions, the duration of sentences remained the same as the original sentences. Together with the original sentences (low *F*0 and low VTR [$L_{F0}L_{VTR}$]), there were four sets of sentences (N = 1,040 sentences [26 lists × 10 sentences × 4 voice conditions]). All 1,040 sentences were root-mean-square (RMS) level equalized.

We performed acoustic analysis to extract the *F*0 and formant frequencies of the four sets of sentence materials (including the original set and three sets of manipulated sentences) used in the present study. For each set of 260 HINT sentences, we segmented the syllables that contained one of the four corner vowels (i.e., /i, u, ɑ, æ/) using Cool Edit 2000 (Syntrillium Software). This resulted in 26, 18, 17, and 21 tokens for /i/, /u/, /ɑ/, and /æ/ in each set. For *F*0 extraction, a custom-written MATLAB script based on autocorrelation analysis was used (Xu et al., 2004; Zhou & Xu, 2008). For formant extraction, we used the speech analysis program TF32 (Milenkovic, 2003). The frequencies of the first two formants, F1 and F2, of each token were extracted based on linear predictive coding analysis. Figure 1 shows the results of the acoustic analysis. The $H_{F0}H_{VTR}$ and $H_{F0}L_{VTR}$ conditions had a similar median *F*0 of approximately 220 Hz, whereas the $L_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$ conditions had a similar median *F*0 of approximately 110 Hz. The vowel spaces for the $H_{F0}H_{VTR}$ and $L_{F0}H_{VTR}$ conditions were similar, and so were those for the $L_{F0}L_{VTR}$ and $H_{F0}L_{VTR}$ conditions. Thus, the acoustic analysis confirmed that manipulations of the *F*0 and formants in Praat had achieved the desired results.

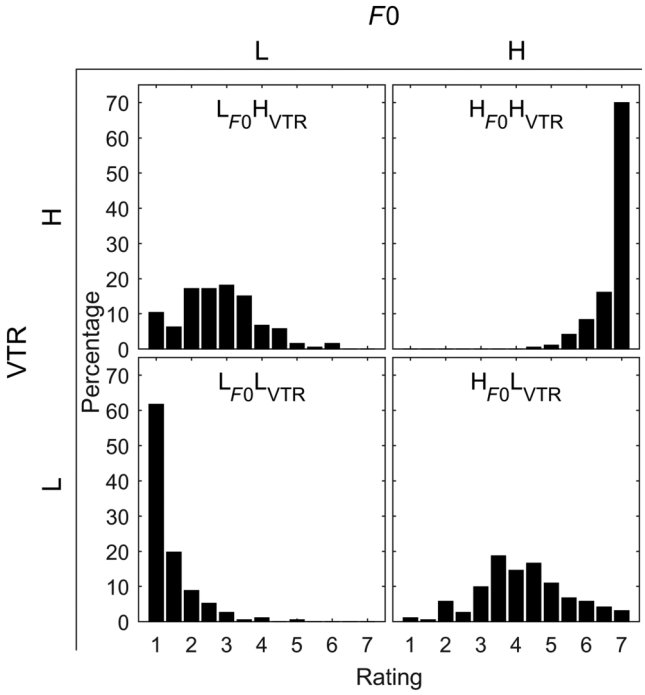
Given that the voice manipulation was based on one single male speaker, we conducted a gender-rating task to examine whether the procedure of manipulating the *F*0 and VTR resulted in anticipated change in voice gender perception. Two sentences were randomly selected from each voice condition, resulting in eight stimulus items. The eight sentences were randomized and presented to 193 college-age adults who were requested to rate the gender of the voice on a scale ranging from 1 (*male*) to 7 (*female*). The mean rating of the two sentences in each voice condition was used to represent the rating result for each rater for that voice condition. As shown in Figure 2, 86.0% of the raters rated 6.5 or higher for the $H_{F0}H_{VTR}$ voice condition, and 81.4% of the raters rated 1.5 or lower

Figure 1. Acoustic analysis of the four corner vowels in each voice condition. Upper panel: fundamental frequency (*F*0) distributions in violin plots in which each symbol represents the mean *F*0 of one vowel token. The white circle and thick vertical lines represent the median and the 25th and 75th percentiles. Lower panel: mean first formant (F1) and second formant (F2) of the four corner vowels /ɑ, æ, i, u/ in each voice condition. The four corner vowels /ɑ, æ, i, u/ are represented by triangle, asterisk, cross, and square symbols, respectively. The mean values of F1 and F2 of all corner vowels are connected to represent the vowel spaces of the four voice conditions. L = low; H = high; VTR = vocal tract resonance.



for the $L_{F0}L_{VTR}$ voice condition. Of the two conditions, $H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$, the ratings were scattered around the middle of the scale. However, $H_{F0}L_{VTR}$ was rated closer to a female voice, whereas $L_{F0}H_{VTR}$ was rated more toward a male voice. The average group rating was 4.2 for $H_{F0}L_{VTR}$ and 2.8 for $L_{F0}H_{VTR}$. Given the large enough sample size (N = 193), a parametric analysis was conducted to compare the rating scores. The rating scores were fitted with a linear mixed-effects model in which the

Figure 2. Histograms showing the gender-rating results across four voice conditions ($H_{F0}H_{VTR}$, $H_{F0}L_{VTR}$, $L_{F0}H_{VTR}$, and $L_{F0}L_{VTR}$), as rated by 193 college-age adults. A rating of 1 on the scale represents a male voice, whereas a rating of 7 represents a female voice. L = low; H = high; *F*0 = fundamental frequency; VTR = vocal tract resonance.



voice condition was defined as the fixed effect and listeners were defined as the random effect, with by-subject intercepts included. The results revealed a significant effect of voice condition, $F(3, 582) = 1{,}293.0$, $p < .0001$. The pairwise comparison suggested that all four voice conditions were significantly different from each other. The gender-rating task verified that doubling the *F*0 and scaling up the VTR by a factor of 1.2 reliably altered the listeners' subjective judgment of the voice gender from a male voice to a female voice. Meanwhile, the mismatched conditions ($H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$) resulted in listeners' uncertainty of voice gender perception.

A 5-min-long SSN was generated by matching the long-term average spectrum of white noise to that of the concatenated 1,040 sentences. Each sentence stimulus was mixed with a randomly selected SSN segment that was 400 ms longer than the sentence duration at four SNRs: −10, −5, 0, and +5 dB. The sentence was presented in the middle of the SSN, with 200 ms of noise preceding and following the sentence signal. The noise header and tail were at the same level as the noise in the speech-plus-noise portion. The level of the sentence stimuli was fixed, and the desired SNR was achieved by changing the RMS level of the SSN relative to that of the sentence stimulus.

Procedure

The perception test was conducted in a sound-treated booth. Each listener started with a practice session in which 24 HINT sentences mixed with SSN were presented in two SNRs: 0 and +5 dB (12 sentences for each SNR). At each SNR level, three sentences were tested in each voice condition. Feedback was provided in the form of written text shown on the computer screen. The purpose of the practice session was to familiarize the listeners with the test procedure. None of the HINT sentences used in the practice session were recycled for the test session.

During the actual test session, the four SNRs were randomized first, and the four voice conditions were then randomized at each SNR level. A total of 16 different HINT sentence lists were used for each listener (4 SNRs × 4 voice conditions). Listeners could adjust the volume to the most comfortable level, which was typically in the range of 60–65 dBA as measured with a sound-level meter (Brüel & Kjær Type 2231). After listening to a sentence, the listeners typed what they heard in the text box on a computer screen. To ensure optimal performance from individual participants, each sentence can be repeatedly played up to three times.

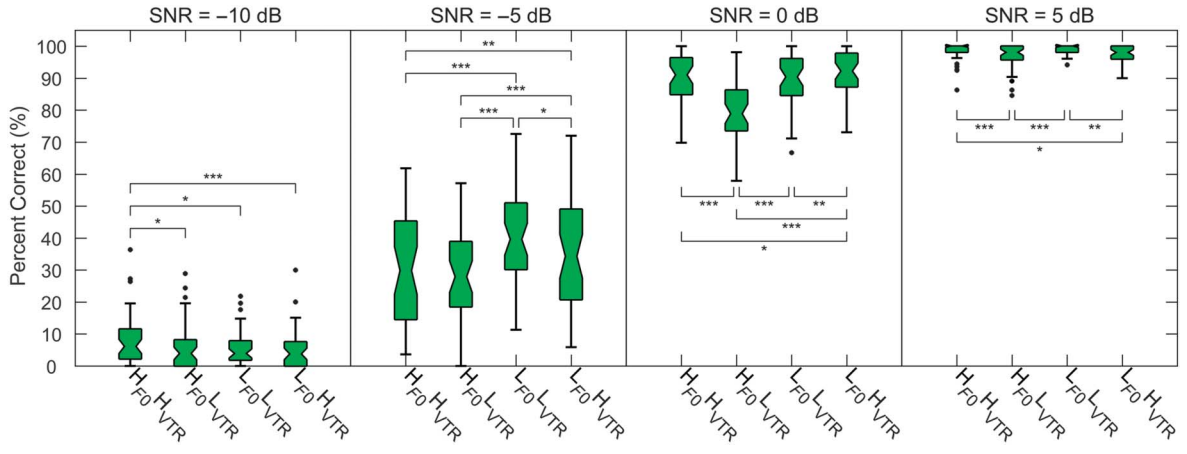
The perception responses were scored by a native English speaker. The accuracy of each sentence list was calculated by dividing the total number of correctly recognized words by the total number of words in all sentences. Strict scoring rules were applied. Spelling errors, homonyms, and grammar errors (such as verb tense use, plural marking, and subject–verb agreement), if present, were counted as wrong.

Results

Figure 3 displays the perception score in percent correct of HINT sentences in the four voice conditions at each SNR level. The perception accuracy was very low at −10 dB SNR (group average < 10% accurate) and very high at +5 dB SNR (group average > 95% accurate). At −5 dB SNR, the group average was approximately 34% correct across all four voice conditions. Meanwhile, the listeners demonstrated considerable individual variability in perception accuracy. When the SNR increased to 0 dB, the listeners performed reasonably well (> 85% correct) except for the $H_{F0}L_{VTR}$ condition. Of the four voice conditions, the listeners showed measurably lower accuracies for $H_{F0}L_{VTR}$ stimuli at 0 and +5 dB SNRs. At −5 dB SNR, the listeners performed slightly better with $L_{F0}L_{VTR}$ stimuli than in the other three conditions.

Percent-correct data were fitted with a generalized linear mixed model in which voice features (four conditions: $H_{F0}H_{VTR}$, $H_{F0}L_{VTR}$, $L_{F0}H_{VTR}$, and $L_{F0}L_{VTR}$) and SNRs (four levels: −10, −5, 0, and +5 dB) were defined as fixed effects and subjects were defined as the random effect. A series of models were implemented, which yielded the best fit model that included by-subject intercepts and by-subject random slopes for voice and SNR. The results showed significant effects of SNR, $F(3, 656) = 999.8$, $p < .001$; voice, $F(3, 656) = 14.6$, $p < .001$; and the

Figure 3. Box plots showing the perception accuracy of Hearing in Noise Test sentences in four voice conditions ($H_{F0}H_{VTR}$, $H_{F0}L_{VTR}$, $L_{F0}H_{VTR}$, and $L_{F0}L_{VTR}$) in four signal-to-noise ratios (SNRs) of −10, −5, 0, and +5 dB. 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. The brackets above or below the box plots indicate statistical significance between pairs of voice conditions, with single (*), double (**), and triple (***) asterisks representing $p < .05$, $p < .01$, and $p < .001$, respectively. L = low; H = high; *F*0 = fundamental frequency; VTR = vocal tract resonance.



SNR × Voice interaction, $F(9, 656) = 13.6$, $p < .001$. The analysis of pairwise contrasts revealed that all SNRs were significantly different from each other. For the factor of voice, there was no significant difference between the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ or between $H_{F0}H_{VTR}$ and $L_{F0}H_{VTR}$. All other comparisons were significant. As for the SNR × Voice interaction effect (see Figure 3), at −10 dB SNR, the $H_{F0}H_{VTR}$ condition was significantly different from the other three voice conditions. At −5 dB SNR, all voice condition pairs were significantly different except for $H_{F0}H_{VTR}$ versus $H_{F0}L_{VTR}$. At 0 dB SNR, there was no significant difference between $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$, but all other comparisons were significant. At +5 dB SNR, there was no difference between $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ or between $H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$, but all other comparisons were significant (all ps < .05).

Based on the recognition performance at the four SNR levels, we implemented logistic curve fitting to model the psychometric function of sentence recognition performance as a function of SNR in each voice condition. When performing logistic curve fitting on the data, the percentage of correct responses (p) was first transformed using a logit function: $y = \log[p/(1 - p)]$. The regression intercept (b) and slope (m) in the regression line $y = m \times \text{SNR} + b$ were determined using the least squares method. The curve-fitted (predicted) sentence recognition performance (P) on the psychometric function can be obtained using $P = e^{(m \times \text{SNR} + b)}/[1 + e^{(m \times \text{SNR} + b)}]$. We also examined the goodness of fit of all 168 (42 subjects × 4 voice conditions) psychometric functions. The mean and median R^2 values were .891 and .932, respectively, indicating satisfactory curve fitting. Supplemental Material S1 contains more details of the goodness-of-fit analysis of the logistic curve fitting. Figure 4 (left four panels) shows the logistic fitting of all subjects in the four voice conditions. The speech reception threshold (SRT; i.e., the SNR at 50% correct) is commonly used in audiology clinics to measure speech recognition ability in noise. The advantage of including SRT analysis in the present study is that it is a single number that can represent speech recognition ability in noise independent of the selection of SNRs. Here, the SRT for each participant was calculated based on logistic fitting. That is, when $P = .5$, SRT = $-b/m$. The group average SRTs were −4.62, −3.11, −4.33, and −3.75 dB for the $H_{F0}H_{VTR}$, $H_{F0}L_{VTR}$, $L_{F0}L_{VTR}$, and $L_{F0}H_{VTR}$ conditions, respectively. Figure 4 (rightmost panel) shows the SRT distribution in the four voice conditions. The SRTs were analyzed using a linear mixed-effects model in which the voice condition was set as a fixed effect and subjects were defined as the random effect, with by-subject intercepts included in the model. The results revealed a significant voice effect, $F(3, 129) = 11.67$, $p < .001$. The pairwise comparisons between voice conditions revealed significant differences between $H_{F0}H_{VTR}$ and $H_{F0}L_{VTR}$, $H_{F0}H_{VTR}$ and $L_{F0}H_{VTR}$, and $L_{F0}L_{VTR}$ and $H_{F0}L_{VTR}$ (all ps < .05). All other comparisons showed no significant difference.

Figure 5 shows the group average logistic fitting for the psychometric function of sentence recognition as a function of SNR in each voice condition. The group average fitting lines for $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ were almost identical. The curves for the $L_{F0}H_{VTR}$ and $H_{F0}L_{VTR}$ conditions differed from those of the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ conditions. As observed at the individual levels (see Figure 4), the group results also demonstrated that, of the four voice conditions, $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ had lower

Figure 4. Logistic fitting for the psychometric function of the sentence recognition data in the four voice conditions (left four panels). Each psychometric function represents data from one subject. The calculated speech reception thresholds (SRTs) for all subjects in the four voice conditions are shown in the rightmost panel. The box represents the 25th and 75th percentiles of the SRTs. The notch represents the median, and the whiskers represent the range of the data. The brackets above the box plots indicate statistical significance between pairs of voice conditions, with single (*) and triple (***) asterisks representing $p < .05$ and $p < .001$, respectively. L = low; H = high; $F0$ = fundamental frequency; VTR = vocal tract resonance; SNR = signal-to-noise ratio.

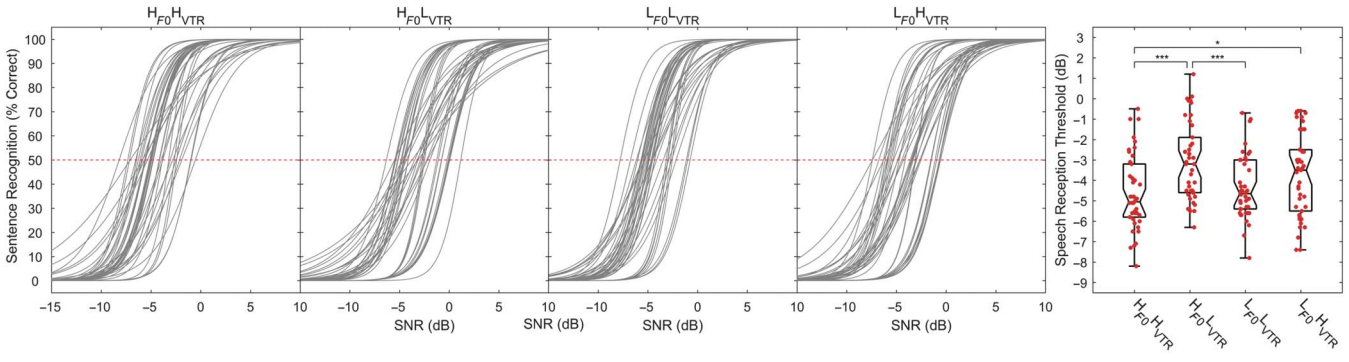
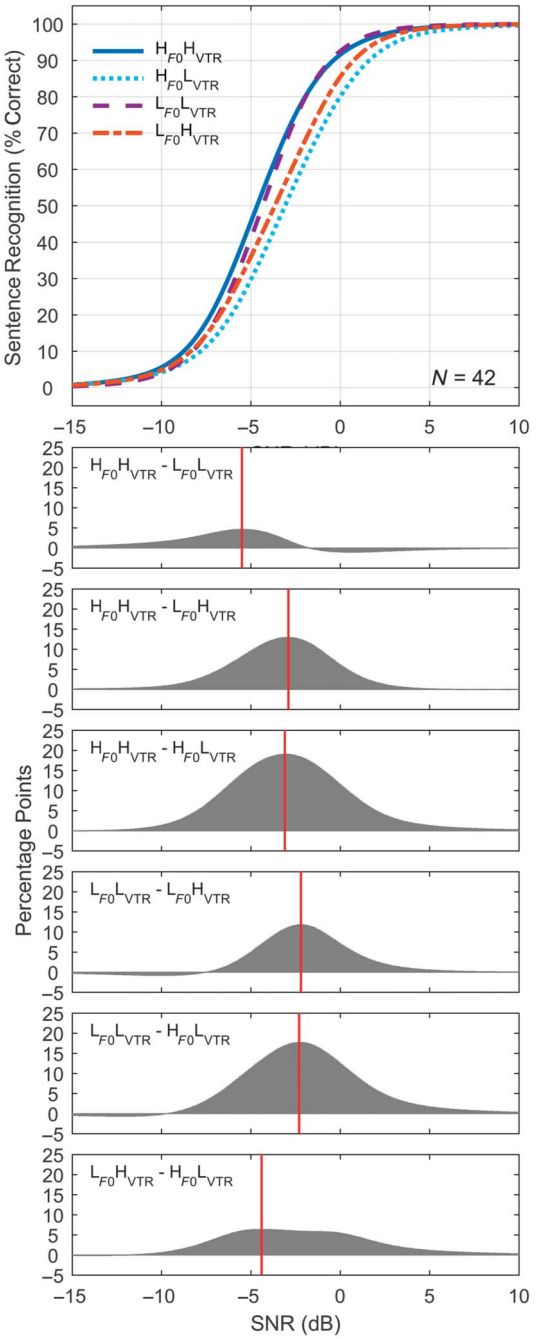


Figure 5. Group average logistic fitting of sentence recognition accuracy as a function of signal-to-noise ratio (SNR) in each voice condition (upper panel). The distributions of the differences between any pair of voice conditions are shown in the lower six panels. The vertical line indicates the SNR at which the maximum difference was located. L = low; H = high; *F*0 = fundamental frequency; VTR = vocal tract resonance.



average SRTs and $H_{F0}L_{VTR}$ had the highest SRTs. In particular, for a given performance accuracy, listeners required a higher SNR when listening to $L_{F0}H_{VTR}$ or $H_{F0}L_{VTR}$ stimuli as compared to $L_{F0}L_{VTR}$ or $H_{F0}H_{VTR}$ stimuli. Compared to $L_{F0}H_{VTR}$, the fitting line for $H_{F0}L_{VTR}$ deviated from $L_{F0}L_{VTR}$ or $H_{F0}H_{VTR}$ to a greater extent. Figure 5 (lower six panels) shows the distribution of the differences in sentence recognition scores between pairs of voice conditions. Large differences can be visualized between $H_{F0}H_{VTR}$ and $H_{F0}L_{VTR}$, $L_{F0}L_{VTR}$ and $H_{F0}L_{VTR}$, and $H_{F0}H_{VTR}$ and $L_{F0}H_{VTR}$ conditions. For instance, the accuracy difference could be greater than 20 percentage points between the $H_{F0}H_{VTR}$ and $H_{F0}L_{VTR}$ conditions.

Discussion

The purpose of the present study was to examine the influence of the voice features of *F*0 and VTR on speech intelligibility in noise. Based on anatomy, talker sex differences are mainly reflected in *F*0 and VTR. In this study, we manipulated the two variables and created four sets of stimuli, of which two sets had *F*0 paired with VTR of the same sex ($H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$) and two sets had *F*0 paired with VTR of the opposite sex ($H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$). The stimuli were presented with SSN at various SNRs. Results revealed that changing the two voice features indeed affected speech intelligibility. However, the difference was not reflected between the two conditions that had *F*0 paired with VTR of the same sex. That is, the female voice with $H_{F0}H_{VTR}$ and the male voice with $L_{F0}L_{VTR}$ showed no significant difference in speech intelligibility in noise. This result conflicts with previous studies that reported higher intelligibility in female speech (Bradlow et al., 1996; Markham & Hazan, 2004; Yoho et al., 2019) or in male speech (McCloy et al., 2015; Robinson, 2011). In those studies that compared speech intelligibility between female and male talkers, rate of speech, clarity of speech production, and other talker-specific characteristics could not be strictly controlled. In the present study, we used one single talker but manipulated the acoustic correlates *F*0 and/or VTR of the original male voice of a standardized hearing test. The gender-rating results (see Figure 2) demonstrated that the modification of *F*0 and VTR successfully altered listeners' perceptual judgment of voice gender. Because the voice conditions were all derived from one speaker, the other talker characteristics that may interfere with intelligibility were well controlled. The lack of difference between the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ stimuli suggested that the reported intelligibility differences between male and female talkers in previous studies might be caused by other acoustic–phonetic properties or by social norms and expectations on speech intelligibility associated with male and female talkers.

Although there was no significant difference between the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ conditions, lower intelligibility

was observed in the speech materials with mismatched *F*0 and VTR, especially in the condition of high *F*0 paired with low VTR. Based on logistic fitting results (see Figure 4), the intelligibility of the $H_{F0}L_{VTR}$ condition could be > 20 percentage points lower than that of the $H_{F0}H_{VTR}$ condition at certain SNRs. The intelligibility difference between the $L_{F0}H_{VTR}$ and $H_{F0}H_{VTR}$ conditions as well as between the $H_{F0}L_{VTR}$ and $L_{F0}L_{VTR}$ conditions could be greater than 15 percentage points. As shown in Figure 3, the SRT for the $H_{F0}L_{VTR}$ condition was significantly higher than those for the other three conditions. The SRT for the $L_{F0}H_{VTR}$ condition was also higher than those for the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ conditions. These findings suggest that speech intelligibility was adversely affected in conditions where *F*0 was paired with VTR from the opposite sex, as compared to the conditions in which *F*0 was matched with VTR from the same sex. Our findings echoed the results reported in previous studies (Assmann & Nearey, 2007, 2008). Assmann and Nearey (2007) studied listeners' vocal preference and vowel recognition for STRAIGHT vocoded stimuli showing downward and upward shifting in *F*0 and/or formant frequency. The results revealed that the listeners preferred voices showing the covariation pattern between *F*0 and formant frequencies (i.e., high *F*0 tended to co-occur with high formant frequencies, whereas low *F*0 tended to co-occur with low formant frequencies). Meanwhile, the listeners recognized vowel stimuli showing the *F*0–formant covariation pattern with higher accuracies.

The choice of speech materials used in speech recognition tasks may influence the results. Assmann and Nearey (2007, 2008) conducted studies using 11 vowel tokens as stimuli. Holmes et al. (2018) and Holmes and Johnsrude (2023) employed a closed-set sentence recognition task (i.e., a matrix test) in which the sentences lacked contextual information. The present study adopted an open-set sentence recognition task using everyday sentences rich in contextual information. Previously, we showed that English vowel and sentence recognition performance under vocoder-processed conditions (Yang et al., 2022) was moderately correlated (r = .622). In that study, sentences with greater contextual information yielded higher recognition scores compared to those with less contextual information. Therefore, if we hold *F*0 and VTR manipulation as well as the SNR constant, we expect that using vowels or a matrix test will result in overall lower performance. However, we have no reason to assume that the pattern of differences across various *F*0–VTR conditions will deviate from what we observed in the present study.

With regard to the specific setting for the *F*0 and VTR manipulation, we applied upward shifting of formant frequencies by a factor of 1.2 and/or upward shifting of *F*0 by a factor of 2.0. This setting reflected the anatomically based sex difference in these two acoustic features. A similar setting was tested by Assmann and Nearey (2008). Their data showed that although the upward shift of formant frequency by a factor of 1.25 or the upward shift of *F*0 by a factor of 2.0 alone reduced vowel intelligibility as compared to the unshifted condition or the condition with the formant frequency and *F*0 shifted upward together, the magnitude of reduction was small. Note that the stimuli in the present study were mixed with SSN at various SNRs. Our data showed that at +5 dB SNR, the speech intelligibility in all four conditions was very high, although the $H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$ conditions showed slightly lower intelligibility relative to the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ conditions (see Figure 3). However, when the SNR decreased to 0 dB or lower, the sentence recognition scores in the conditions with mismatched *F*0 and VTR could be 20 percentage points lower than those in the conditions with matched *F*0 and VTR. These results indicated that the adverse effect of frequency shifting exacerbated in noise conditions. Of the two mismatched conditions, our data revealed an even lower accuracy for the $H_{F0}L_{VTR}$ condition than for the $L_{F0}H_{VTR}$ condition. As shown in Figure 2, the recognition accuracy of the $H_{F0}L_{VTR}$ condition was much lower than that of the other three conditions. According to logistic fitting, the greatest performance difference between the $H_{F0}L_{VTR}$ and $H_{F0}H_{VTR}$ conditions could be greater than 20 percentage points. Results in Assmann and Nearey (2008) also showed worse vowel recognition with the upward-shifted *F*0 paired with the downward-shifted spectral envelope. The authors proposed that raised *F*0 resulted in fewer harmonics. Therefore, the vowel formants were poorly represented. In the meantime, when the spectral envelope was shifted downward, the F1 could be lower than the *F*0.

In Holmes et al. (2018) and Holmes and Johnsrude (2023), the authors found robust familiar-voice benefit in speech intelligibility even when the voice features were manipulated within listeners' discrimination threshold or by a large amount. The authors proposed that the speech intelligibility benefit might be related to more efficient and active cognitive processing with familiar voices than with unfamiliar voices. Although the current study does not involve familiarity with talkers' voices, we observed a detectable adverse impact of *F*0–VTR mismatched voices on speech intelligibility in noise conditions. Because *F*0–VTR matched voices are what people are being exposed to most often in their daily life, the mismatched voices may sound less natural. It is possible that compared to *F*0–VTR matched voices that sound more natural, less natural voices may undergo different normalization and less efficient cognitive processing when greater cognitive resources are demanded in adverse conditions (e.g., with background noise).

Another point worth noting was the contribution of *F*0 and VTR to listeners' perception of voice gender. Although both acoustic correlates contain sex-related perceptual cues, many studies revealed the greater contribution of *F*0 in voice gender identification. Gelfer and Mikos (2005) recruited 10 male-to-female transgender persons, 10 biological men, and 10 biological women who produced sustained isolated vowels /i/, /u/, and /ɜ/. For each spoken vowel, two synthesized tokens were generated: one with the *F*0 at 120 Hz and the other with the *F*0 at 240 Hz. The gender identification results from 30 normal-hearing young adults revealed that when male formants were paired with a high *F*0, listeners perceived a female speaker. When female formants were paired with a low *F*0, listeners perceived a male speaker. A similar finding of the predominant role of *F*0 in voice gender identification was reported by Poon and Ng (2015). In that study, the authors created the synthesized vowel /a/ by multiplying the male formants by 10 scale factors from 1 to 1.2 or multiplying the female formants by 10 scale factors from 1 to 0.83. *F*0 was modified in 10 steps from 100 to 250 Hz. The results showed that stimuli with increasing *F*0 were more likely to be perceived as female voices, regardless of the formant values. In the present study, our gender-rating results indicated that between the two mismatched voice conditions, listeners rated the one with the high *F*0 as more toward a female voice even though the VTR was low, whereas the one with the low *F*0 was rated more toward a male voice even though the VTR was high. Our data provided additional support for the importance of *F*0 in voice gender perception. Of the two sex-related acoustic correlates, *F*0 serves as the more dominant cue for voice gender identification.

In general, our data revealed no significant difference between the $H_{F0}H_{VTR}$ and $L_{F0}L_{VTR}$ voice conditions in speech intelligibility. This finding has important implications in the development of speech and hearing tests. In speech and hearing clinical practice, the field currently uses various types of auditory stimuli to assess listeners' perceptual performance, which predominantly involve cisgender male talkers. Yet, there is no solid scientific evidence to support the superiority of cisgender male voices for intelligibility. Our data yielded no significant difference in speech intelligibility associated with anatomically based sex difference when other factors are controlled. Furthermore, our research findings provide valuable information regarding speech communication of gender-diverse populations. For example, for transgender people who undergo surgical procedures, hormone treatment, or voice therapy for gender affirmation, these treatment and training procedures can alter the voice pitch (Chadwick et al., 2022; Cler et al., 2020; Schwarz et al., 2017; Song & Jiang, 2017) but hardly modify vocal resonance because the anatomical structure of the vocal tract is preserved, even though adjusting articulatory positions (e.g., lowering the jaw or opening the mouth to lower the F1) can somehow change vocal resonant features for certain sounds (Leyns et al., 2021). In this case, the voice features of *F*0 and VTR may not show the covariation pattern. Our results suggested that speech intelligibility could be affected in noise conditions with a mismatched *F*0–VTR relationship. This finding highlights the importance of developing speech training and practice programs to address intelligibility in research and clinical practice for gender-diverse populations.

Several limitations of the current study should be noted. First, we had only one male speaker, and the manipulations were all based on this talker. Although the male talker was adopted from a standardized hearing test, the limited number of talkers restricts the generalization of the current findings. For future studies, we should include more talkers. In addition to male talkers and applying upward shifts of the *F*0 and/or VTR to change to a female voice, we should include female talkers to apply downward shifts of the two acoustic variables to change to a male voice and compare the effects between downward and upward frequency shifting. Second, the selection of SNRs at −10 and +5 dB in the present study caused floor and ceiling effects. The listeners demonstrated extremely low or high sentence recognition in these two SNR conditions. According to logistic fitting, the greatest difference among the four talker gender conditions occurred between −5 and 0 dB SNRs. For future studies, we should adjust the SNRs and use finer SNR steps for a more precise comparison of intelligibility differences among different voice conditions. Finally, caution should be exercised when interpreting our results in the context of transgender voices. The acoustic manipulations, as in our $H_{F0}L_{VTR}$ and $L_{F0}H_{VTR}$ conditions, may not be reflective of true transgender voices. Future research is needed to elucidate the potential intelligibility deficiency of transgender voices in noisy listening environments.

Data Availability Statement

The sentence stimuli generated and/or the perceptual data in the current study are available from the corresponding author on reasonable request.

Acknowledgments

The authors would like to acknowledge Ohio University students Andrea DiCiaccio, Ariel Vitartas, Abigail Jaquish, and Brihana Joseph for providing technical assistance in data collection.

References

Assmann, P. F., & Nearey, T. M. (2007). Relationship between fundamental and formant frequencies in voice preference. *The Journal of the Acoustical Society of America, 122*(2), EL35–EL43. https://doi.org/10.1121/1.2719045

Assmann, P. F., & Nearey, T. M. (2008). Identification of frequency-shifted vowels. *The Journal of the Acoustical Society of America, 124*(5), 3203–3212. https://doi.org/10.1121/1.2980456

Bachorowski, J.-A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106*(2), 1054–1063. https://doi.org/10.1121/1.427115

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* (Version 5.4.04) [Computer software]. http://www.praat.org/

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*(1), 272–284. https://doi.org/10.1121/1.1487837

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic–phonetic talker characteristics. *Speech Communication, 20*(3–4), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15*(1–2), 39–54. https://doi.org/10.1016/0167-6393(94)90039-6

Chadwick, K. A., Coleman, R., Andreadis, K., Pitti, M., & Rameau, A. (2022). Outcomes of gender-affirming voice and communication modification for transgender individuals. *The Laryngoscope, 132*(8), 1615–1621. https://doi.org/10.1002/lary.29946

Cler, G. J., McKenna, V. S., Dahl, K. L., & Stepp, C. E. (2020). Longitudinal case study of transgender voice changes under testosterone hormone therapy. *Journal of Voice, 34*(5), 748–762. https://doi.org/10.1016/j.jvoice.2019.03.006

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). 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*(5), 2913–2922. https://doi.org/10.1121/1.1616924

Fant, G. (1960). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations*. De Gruyter Mouton.

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America, 116*(4), 2365–2373. https://doi.org/10.1121/1.1788730

Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q.-J., Free, R. H., & Başkent, D. (2014). Gender categorization is abnormal in cochlear implant users. *Journal of the Association for Research in Otolaryngology, 15*(6), 1037–1048. https://doi.org/10.1007/s10162-014-0483-7

Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. D. (2009). The role of glottal pulse rate and vocal tract length in the perception of speaker identity. In *Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 148–151). https://doi.org/10.21437/Interspeech.2009-54

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender. *Journal of Voice, 27*(5), 556–566. https://doi.org/10.1016/j.jvoice.2012.11.008

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice, 19*(4), 544–554. https://doi.org/10.1016/j.jvoice.2004.10.006

Gengel, R. W., & Kupperman, G. L. (1980). Word discrimination in noise: Effect of different speakers. *Ear and Hearing, 1*(3), 156–160. https://doi.org/10.1097/00003446-198005000-00008

Hazan, V., & Markham, D. (2004). Acoustic–phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America, 116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

Hazan, V., Tuomainen, O., Tu, L., Kim, J., Davis, C., Brungart, D., & Sheffield, B. (2018). How do aging and age-related hearing loss affect the ability to communicate effectively in challenging communicative conditions? *Hearing Research, 369,* 33–41. https://doi.org/10.1016/j.heares.2018.06.009

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f_0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*(5), 1150–1166. https://doi.org/10.3758/APP.71.5.1150

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science, 29*(10), 1575–1583. https://doi.org/10.1177/0956797618779083

Holmes, E., & Johnsrude, I. S. (2023). Intelligibility benefit for familiar voices is not accompanied by better discrimination of fundamental frequency or vocal tract length. *Hearing Research, 429,* Article 108704. https://doi.org/10.1016/j.heares.2023.108704

Jacewicz, E., Fox, R. A., & Holt, C. E. (2023). Dialect and gender perception in relation to the intelligibility of low-pass and high-pass filtered spontaneous speech. *The Journal of the Acoustical Society of America, 154*(3), 1667–1683. https://doi.org/10.1121/10.0020906

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820–857. https://doi.org/10.1121/1.398894

Krause, J. C., & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America, 112*(5), 2165–2172. https://doi.org/10.1121/1.1509432

Kwon, H.-B. (2010). Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses. *The Journal of Advanced Prosthodontics, 2*(3), 71–76. https://doi.org/10.4047/jap.2010.2.3.71

Leyns, C., Papeleu, T., Tomassen, P., T'Sjoen, G., & D'haeseleer, E. (2021). Effects of speech therapy for transgender women: A systematic review. *International Journal of Transgender Health, 22*(4), 360–380. https://doi.org/10.1080/26895269.2021.1915224

Markham, D., & Hazan, V. (2004). 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*(4), 725–737. https://doi.org/10.1044/1092-4388(2004/055)

McCloy, D. R., Wright, R. A., & Souza, P. E. (2015). Talker versus dialect effects on speech intelligibility: A symmetrical study. *Language and Speech, 58*(3), 371–386. https://doi.org/10.1177/0023830914559234

Milenkovic, P. (2003). *TF32 software program* [Computer program]. University of Wisconsin–Madison. https://ubeam.engr.wisc.edu/

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-

accented speech. *Language and Speech, 38*(3), 289–306. https://doi.org/10.1177/002383099503800305

Munson, B., & Babel, M. (2019). The phonetics of sex and gender. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge handbook of phonetics* (pp. 499–525). Routledge. https://doi.org/10.4324/9780429056253-19

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). 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), 1085–1099. https://doi.org/10.1121/1.408469

Nixon, C. W., Anderson, T. R., Morris, L., McCavitt, A., McKinley, R., Yeager, D., & McDaniel, M. (1998). Female voice communications in high level aircraft cockpit noises—Part II: Vocoder and automatic speech recognition systems. *Aviation, Space, and Environmental Medicine, 69*(11), 1087–1094.

Nixon, C. W., Morris, L. J., McCavitt, A. R., McKinley, R. L., Anderson, T. R., McDaniel, M. P., & Yeager, D. G. (1998). Female voice communications in high levels of aircraft cockpit noises—Part I: Spectra, levels, and microphones. *Aviation, Space, and Environmental Medicine, 69*(7), 675–683.

Oh, Y., Kalpin, N., Hunter, J., & Schwalm, M. (2023). The impact of temporally coherent visual and vibrotactile cues on speech recognition in noise. *JASA Express Letters, 3*(2), Article 025203. https://doi.org/10.1121/10.0017326

Oh, Y., Schwalm, M., & Kalpin, N. (2022). Multisensory benefits for speech recognition in noisy environments. *Frontiers in Neuroscience, 16,* Article 1031424. https://doi.org/10.3389/fnins.2022.1031424

Poon, M. S. F., & Ng, M. L. (2015). The role of fundamental frequency and formants in voice gender identification. *Speech, Language and Hearing, 18*(3), 161–165. https://doi.org/10.1179/2050572814Y.0000000058

Robinson, E. J. (2011). *The effect of talker age and gender on speech perception of pediatric hearing aid users* [Thesis, Washington University School of Medicine]. Independent Studies and Capstones. https://digitalcommons.wustl.edu/pacs_capstones/635

Schwarz, K., Fontanari, A. M. V., Schneider, M. A., Borba Soll, B. M., da Silva, D. C., Spritzer, P. M., Kazumi Yamaguti Dorfman, M. E., Kuhl, G., Costa, A. B., Cielo, C. A., Villas Bôas, A. P., & Rodrigues Lobato, M. I. (2017). Laryngeal surgical treatment in transgender women: A systematic review and meta-analysis. *The Laryngoscope, 127*(11), 2596–2603. https://doi.org/10.1002/lary.26692

Silverstein, B., Bilger, R. C., Hanley, T. D., & Steer, M. D. (1953). The relative intelligibility of male and female talkers. *Journal of Educational Psychology, 44*(7), 418–428. https://doi.org/10.1037/h0054345

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research, 57*(1), 285–296. https://doi.org/10.1044/1092-4388(2013/12-0314)

Smiljanic, R., & Gilbert, R. C. (2017). Intelligibility of noise-adapted and clear speech in child, young adult, and older adult talkers. *Journal of Speech, Language, and Hearing Research, 60*(11), 3069–3080. https://doi.org/10.1044/2017_JSLHR-S-16-0165

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America, 118*(5), 3177–3186. https://doi.org/10.1121/1.2047107

Song, T. E., & Jiang, N. (2017). Transgender phonosurgery: A systematic review and meta-analysis. *Otolaryngology—Head and Neck Surgery, 156*(5), 803–808. https://doi.org/10.1177/0194599817697050

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). Online experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech & Language, 19*(2), 129–146. https://doi.org/10.1016/j.csl.2004.03.003

Tanaka, A., Sakamoto, S., & Suzuki, Y. (2011). Effects of pause duration and speech rate on sentence intelligibility in younger and older adult listeners. *Acoustical Science and Technology, 32*(6), 264–267. https://doi.org/10.1250/ast.32.264

Tielen, M. T. J. (1989). Intelligibility of male and female voices under a few noise conditions. In *Proceedings of the First European Conference on Speech Communication and Technology* (pp. 2127–2130). https://doi.org/10.21437/Eurospeech.1989-216

Tripp, A., & Munson, B. (2022). Perceiving gender while perceiving language: Integrating psycholinguistics and gender theory. *WIREs Cognitive Science, 13*(2), Article e1583. https://doi.org/10.1002/wcs.1583

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research, 39*(3), 494–509. https://doi.org/10.1044/jshr.3903.494

Vestergaard, M. D., Fyson, N. R. C., & Patterson, R. D. (2009). The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America, 125*(2), 1114–1124. https://doi.org/10.1121/1.3050321

Weirich, M., & Simpson, A. P. (2018). Gender identity is indexed and perceived in speech. *PLOS ONE, 13*(12), Article e0209226. https://doi.org/10.1371/journal.pone.0209226

Whiteside, S. P. (1998a). Identification of a speaker's sex: A fricative study. *Perceptual and Motor Skills, 86*(2), 587–591. https://doi.org/10.2466/pms.1998.86.2.587

Whiteside, S. P. (1998b). Identification of a speaker's sex: A study of vowels. *Perceptual and Motor Skills, 86*(2), 579–584. https://doi.org/10.2466/pms.1998.86.2.579

Xu, L., Li, Y., Hao, J., Chen, X., Xue, S. A., & Han, D. (2004). Tone production in Mandarin-speaking children with cochlear implants: A preliminary study. *Acta Oto-Laryngologica, 124*(4), 363–367. https://doi.org/10.1080/00016480410016351

Yang, J., Barrett, J., Yin, Z., & Xu, L. (2023). Recognition of foreign-accented vocoded speech by native English listeners. *Acta Acustica, 7,* Article 43. https://doi.org/10.1051/aacus/2023038

Yang, J., Wagner, A., Zhang, Y., & Xu, L. (2022). Recognition of vocoded speech in English by Mandarin-speaking English-learners. *Speech Communication, 136,* 63–75. https://doi.org/10.1016/j.specom.2021.11.008

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception, & Psychophysics, 81*(2), 558–570. https://doi.org/10.3758/s13414-018-1635-3

Zhou, N., & Xu, L. (2008). Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants. *The Journal of the Acoustical Society of America, 123*(3), 1653–1664. https://doi.org/10.1121/1.2832623