Attention, Perception, & Psychophysics (2019) 81:558–570
https://doi.org/10.3758/s13414-018-1635-3

Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology

Sarah E. Yoho[1] · Stephanie A. Borrie[1] · Tyson S. Barrett[2] · Dane B. Whittaker[1]

Published online: 30 November 2018

Abstract
Talker and listener sex in speech processing has been largely unknown and under-appreciated to this point, with many studies overlooking the possible influences. In the current study, the effects of both talker and listener sex on speech intelligibility were assessed. Different methodological approaches to measuring intelligibility (percent words correct vs. subjective rating scales) and collecting data (laboratory vs. crowdsourcing) were also evaluated. Findings revealed that, regardless of methodology, the spoken productions of female talkers were overall more intelligible than the spoken productions of male talkers; however, substantial variability across talkers was observed. Findings also revealed that when data were collected in the lab, there was an interaction between talker and listener sex. This interaction between listener and talker sex was not observed when subjective ratings were crowdsourced from listener subjects across the USA via Amazon Mechanical Turk, although overall ratings remained similar. This possibly suggests that subjective intelligibility ratings may be vulnerable to bias, and such biases may be reduced by recruiting a more heterogeneous subject pool. Many studies in speech perception do not account for these talker, listener, and methodology effects. However, the present results suggest that researchers should carefully consider these effects when assessing speech intelligibility in different conditions, and when comparing findings across studies that have used different subject demographics and/or methodologies.

Keywords Hearing · Speech perception

Introduction

Despite an increased emphasis on diversity of human subject participants in behavioral and health science research in recent decades (Allmark, 2004; National Institutes of Health, 2017), there remains a stark lack of diversity in many studies that examine speech perception. This may in part be due to the homogenous population from which research participants for these types of studies are recruited – typically convenience samples of undergraduate university students, who vary little in terms of demographics such as age, socioeconomic status, and place of birth. Therefore, many current research studies in the field of speech perception are based on a relatively homogenous sample of listeners. In addition, perhaps due to convention, or simply the utilization of speech recordings that are readily available, speech perception studies are often carried out without the use of diverse talker representation. The impact of this lack of diversity on the development of speech processing models has been largely unknown and under-appreciated to this point. Here we postulate that in terms of homogeneity of inclusion in these studies, there are two primary influences: the selection of the particular talker or talkers as speech stimulus, and the demographic makeup of the listener group. In addition, there may be interactions between these two factors and the particular method or measure of speech intelligibility utilized in the study.

Effects of talker sex

Although the diversity of talkers used as stimulus in speech perception studies may be lacking along many dimensions (e.g., race, age, dialect), one of the most fundamental demographics is talker sex. Females represent roughly 50% of the worldwide population, however, many studies include only male voices as stimulus. For example, in the Speech Intelligibility Index (ANSI, 1997), a commonly used ANSI

✉ Sarah E. Yoho
sarah.leopold@usu.edu

[1] Department of Communicative Disorders and Deaf Education, Utah State University, Logan, UT 84322, USA

[2] Department of Kinesiology and Health Sciences, Utah State University, Logan, UT 84322, USA

standard, which provides tables and calculations for predictions of speech intelligibility, the vast majority of the data were derived with only male voices represented. In addition, many seminal, highly influential studies in the area of speech perception have included only male voices (e.g., Cooke, 2006; Shannon et al., 1995; Wang & Bilger, 1973). This lack of diversity not only under-represents many talker demographics, but may also influence the outcomes of these studies in ways that are not well appreciated. For example, Yoho et al. (2018) found that there are important differences due to talker sex in the distribution of the frequency-importance functions such as those found in the Speech Intelligibility Index, and those differences are not currently documented in the ANSI standard.

It is well known that there are several distinguishing characteristics of male and female voices, and, as such, listeners are generally able to accurately identify talker sex (Coleman, 1971; Lass et al., 1976; Schwartz, 1968). These dissimilarities can be attributed to physical, anatomical differences between males and females, as well as learned articulation behaviors due to societal influences. Perhaps the most prominent voice-related anatomical difference between males and females is the average vocal tract size of males (17–18 cm) versus females (14–14.5 cm; Simpson, 2009). Physical differences such as this result in perceptually salient phonatory and articulatory differences, such as fundamental frequency (F0; i.e., acoustic correlate of pitch), harmonic spacing, vowel space, and voice-onset time (Simpson, 2009).

The manner in which these differences between male and female voices impact the perception of speech is somewhat complex. In fact, there seems to be some conflict even for the most basic question of whether male or female voices are overall more intelligible for listeners. Although a handful of studies have found female talkers to be more intelligible, there is disagreement. Bradlow et al. (1996) found that for listeners responding to sentences in quiet conditions, female talkers were significantly more intelligible than males. Likewise, Markham and Hazan (2004) found females to be more intelligible when listeners were presented with monosyllables in background noise spoken by male and female talkers. However, the effect in both studies was small and there were high and low intelligibility talkers within each sex group. Conversely, McCloy et al. (2015) found male talkers to be more intelligible, and Gengel and Kupperman (1980) found male and female voices to be equivocal in terms of intelligibility.

The question of why one sex may be more intelligible than another remains somewhat unclear. Bradlow et al. (1996) examined the influence of F0 and speaking rate on intelligibility across male and female talkers and found no significant effects, but postulated that the higher F0 range of female speakers may have played a role in the overall higher intelligibility of that group. In addition, it was found that the timing of co-articulation differed between the sexes, with females adding more pauses between syllables and overall demonstrating more precise articulation. Byrd (1994) examined the TIMIT speech corpus (Garofolo, 1988), which contains sentences spoken by 630 different individuals. In this analysis, it was found that male speech is generally more "reduced" or less well-articulated than female speech. The speaking rate of male speech was on faster on average than that of females, and males released sentence-final stops less, had more centralized vowels, and had more voiceless vowels. Ferguson (2004) found that when individuals were instructed to speak "clearly," vowel intelligibility was higher for females than for males, but there was no difference between the sexes when speaking conversationally, so it is conceivable that females are simply more adept at articulating more clearly when the situation requires.

Effects of listener sex

Another potentially important factor that has been even less explored in the literature is the possible influence of *listener* sex in speech perception. It is imperative to understand this potential influence, particularly as speech perception studies commonly fail to balance the distribution of male and female listeners, oftentimes skewing quite heavily towards one sex (e.g., Borrie & Schäfer, 2017; Healy et al., 2013; Klasner & Yorkston, 2005). In addition, many studies simply do not report the distribution of listener sex (e.g., Fogerty, 2011; Van Engen & Bradlow, 2007). While there has been limited investigation into the direct influence of listener sex in understanding speech, there are some data, including from psychoacoustic and neurophysiological studies, to indicate that differences may exist between the male and female listeners on auditory tasks. For example, differences are known to exist in the anatomy of the primary auditory cortex between males and females (Rademacher et al., 2001), and functional differences have been observed in verbal working memory (Bleecker et al., 1988). Several studies have shown differences in auditory brainstem responses as a function of sex, with males displaying longer latencies and smaller amplitudes (e.g., Dehan & Jerger, 1990; Don et al., 1993).

There are also several documented differences between the sexes for psychoacoustical tasks. For even arguably the most basic auditory task, audiometric thresholds for simple sinusoids, female listeners on average display lower thresholds than males by approximately 2–3 dB at some frequencies (McFadden, 1998). In addition, females display stronger click-evoked otoacoustic emissions, narrower auditory filters, and even slightly better gap detection (temporal) thresholds, whereas males display better localization, more sensitive detection of basic signals in complex maskers, and higher sensitivity in a profile-analysis task (a measure of informational masking) (McFadden, 1998; McFadden et al., 2018). Males

have also been shown to display better fundamental frequency contour discrimination (McRoberts et al., 1992).

Although it is clear that there are differences in the processing of non-speech auditory tasks between male and female listeners, the relationship between more basic auditory tasks and speech perception can be quite complex, and very little is currently known about the impact of listener sex on speech perception. One study by Ellis et al. (1996) examined male and female listener perceptions of speech samples produced by male and female talkers. Male and female listeners were presented with recordings of one male and one female talker. Using a magnitude estimation scale procedure, listeners rated the recordings, and results indicated no differences between male and female listeners' ratings. However, when asked to give an overall impression of intelligibility of the two talkers, male listeners indicated that the female voice was easier to understand, and female listeners indicated that the male voice was easier to understand. Rogers et al. (2003) found that both male and female listeners prefer a similar signal-to-noise ratio when listening to speech in noise. Finally, Markham and Hazan (2004) found no difference in intelligibility in terms of percent words correct as a function of listener sex.

Effects of methodology

Some of the uncertainty surrounding the influence of talker and listener sex in speech perception may be due to the differing methodologies employed. For instance, in the studies that have examined these effects either directly or indirectly, differences exist in stimulus type (vowels, monosyllables, sentences, running speech) and listening environment (quiet, babble, speech-shaped noise, etc.) (e.g., Bradlow et al. 1996; Ferguson 2004; Hazan & Markham, 2004). Another potentially important methodological difference is the particular outcome measure utilized. Many of the studies discussed so far performed a direct objective measurement of speech intelligibility – a measure of percent words correct, either for sentences or for isolated words. However, other studies, such as that by Ellis et al. (1996), who looked at listener sex differences, involve more subjective measurements such as listener impressions of speech. In another example of using subjective measurement, Kwon (2010) examined listeners' perceptions of speech intelligibility for both male and female talkers. Three trained speech-language pathologists rated the intelligibility of ten male and ten female talkers using a 10-point Likert scale. In accord with some of the more objective (e.g., percent word correct) measures, results indicated that perceptions of intelligibility of female talkers were significantly higher than perceptions of male talkers. However, correlations between these ratings and acoustic analyses of talkers' voices (i.e., F0, F0 range, formant frequency, formant range, vowel working space area, vowel dispersion) were low. Likewise, Ferguson and Morgan (2018) found that listeners rated female voices as clearer than male voices when the talkers were utilizing "clear speech." Unfortunately, it is difficult to make direct comparisons between these different objective and subjective methodologies due to the many other differences that exist across the studies.

Purpose of the current study

The purpose of the current study was to clarify these effects by performing a systematic and joint examination of these three important variables: talker sex, listener sex, and methodology, on speech perception. To do so, we utilized sentence-level stimuli presented in a cafeteria noise background to represent a reasonably realistic listening environment. Talkers of each sex were carefully chosen to ensure that their average F0 fell within the average F0 range for their sex, and to control for overall speaking rate, which is known to influence listener judgments of speakers (e.g., Brown, 1990; Smith et al., 1975). In addition, participants were selected such that there was ample representation of both male and female listeners. Comparisons were made between objective speech intelligibility (percent words correct) and subjective listener impressions of speech intelligibility to examine methodology effects. Lastly, an additional experiment was conducted in which crowdsourcing via Amazon Mechanical Turk (MTurk) was utilized to enable the recruitment and inclusion of a more heterogenous listener group. Thus, the aims of the current study were: (1) to determine the effects of both talker and listener sex on the intelligibility of speech in the presence of a background noise typical of those encountered in everyday life, and (2) to determine if the particular methodology employed (percent words correct vs. subjective ratings; laboratory-based vs. crowdsourcing) influences these effects.

Experiment 1: Laboratory data collection

Method

Participants

Two groups of listeners participated in the first experiment. One group consisted of 25 male participants between the ages of 18 and 33 years (M = 22.9 years). The other group consisted of 25 female participants between the ages of 19 and 35 years (M = 22.8 years). All but one of the participants were Caucasian. All had pure-tone audiometric thresholds on day of test at or below 20 dB HL at 1,000, 2,000, and 4,000 Hz (ASHA, 1997). None had previous exposure to the sentence materials utilized in this study. Listeners were recruited from the student population of Utah State University and surrounding community of Logan, Utah, and received course credit or monetary incentive for participation.

Stimuli and procedure

Test stimuli consisted of 100 sentences from the Harvard IEEE corpus (IEEE, 1969). Each sentence contains five key words for scoring. In-house recordings from ten different talkers judged to have Standard American English were used (five male, all recordings 22 kHz, 16-bit). The talkers were given no specific instructions on how to speak. Using a customized feature extraction script and acoustic analysis software (Praat, Boersma, & Weenink, 2017), a series of F0 measures, including average F0, and speaking rate, reported in syllables per second (sps) were calculated for each talker. The average F0 of each talker was measured to ensure all selected talkers fell within the pitch norms of standard "male" and "female" voices, and the average speaking rate for each talker was measured to ensure the selected talkers presented with similar speaking rates. Results of the acoustic analysis revealed that the average F0 of the female talkers fell within the range of 185–260 Hz (M = 240.51 Hz), and the average F0 of the male talkers fell within the range of 90–146 Hz (M = 111.06 Hz) . These ranges are well within what is consider normal for each sex (Titze, 1989). Acoustic analysis of speaking rate revealed that all talkers fell within the range of 3.30–3.80 sps, with females having an average of 3.56 sps and males having an average of 3.58 sps. Additional F0 measures are shown in Table 1. Of note, the female talkers presented with greater F0 variation and range (max F0 − min F0) compared with the males talkers, which reflects sex norms (Goy et al., 2013).

Signal processing was performed with Adobe Audition software. The stimuli were equated based on total RMS and concatenated such that ten sentences from each talker were presented contiguously before moving on to the next talker. A minimum of 200 ms of silence preceded and followed each sentence prior to mixing with noise to avoid issues related to masking overshoot (Bacon, 1990). The concatenated sentences were then mixed with cafeteria noise from an Auditec CD (St. Louis, MO, USA; www.auditec.com) at a signal-to-noise ratio of -2 dB. This SNR was chosen based on pilot testing to ensure that the intelligibility levels for each of the ten talkers would not be at ceiling or floor levels. The cafeteria noise consisted of three overdubbed recordings made in a hospital employee cafeteria. Cafeteria noise was chosen over other types of noise (Gaussian, multi-talker babble) due to its relatively more ecologically-valid nature. Participants were pseudo-randomly assigned to one of five randomizations of talker-to-sentence correspondence, for a total of five listeners of each sex in each order of talker-to-sentence correspondence. Each of the five randomizations of talker-to-sentence correspondence was mixed with the cafeteria noise separately. Stimuli were presented diotically through Sennheiser HD 280 supra-aural headphones via a personal computer running a custom E-prime interface (E-prime version 3) and equipped with a Presonus Studio 26 digital-to-analog converter. The average RMS level of the continuous speech and noise mixtures was set to playback at 65 dBA, and calibration was performed using a Larson Davis sound-level meter and flat-plate coupler (Models 824 and AEC 101).

Table 1 Fundamental frequency (Hz) values for females and males

Measure	Females	Males
Min	159.57	85.09
Max	418.85	223.42
Mean	240.51	111.06
SD	45.03	32.31

Participants were seated in a double-walled, sound-attenuated booth. For each sentence, the participants were instructed to repeat as much of each sentence as possible. Participants were given as much time as needed before the next sentence was played by the experimenter. On each trial, the participants were also instructed to report back their impression of the talker's voice. For the subjective impression, the participants were asked "To what extent do you agree with this statement? The talker is easily understood." They were asked to rate each sentence on a 7-point scale, with 1 representing "strongly disagree" and 7 representing "strongly agree." The participants were told to focus on the talker's voice, not on the effect of the noise. The experimenter sat in the booth with each participant and typed the responses on the custom E-prime interface. The experiment took approximately 1 h for each participant to complete.

Percent words correct scoring

For the percent words correct (PWC) measures, participant responses were scored for keywords correct by two independent research assistants. Keywords were counted as correct if they were repeated precisely, or if a syllable or phoneme was added (e.g., "distracted" for "distract"). An analysis of inter-rater reliability indicated that the two scorers were in agreement 98% of the time.

Results of Experiment 1

For Experiment 1, overall average PWC and rating scores for male and female talkers, as well as average PWC and rating scores broken down by listener sex for each talker group, are shown in Table 2. Individual PWC and rating scores for each of the ten talkers are shown in Table 3. Average PWC and rating scores for male and female talkers, as a function of listener sex, are shown in Figs. 1 and 2, respectively. In Fig. 1, two aspects are clear: (1) as a group, female talkers were more intelligible than male talkers, (2) although male listeners performed slightly lower than female listeners in objective

Table 2 Overall intelligibility for percent words correct (PWC) and ratings. Standard deviations are given in parentheses

	PWC	Rating
Overall Talker Group		
Female Talkers	66.6% (0.32)	4.28 (1.57)
Male Talkers	50.8% (0.37)	3.50 (1.72)
Talker Group by Listener Group		
Female Talkers (Female Listeners)	68.1% (0.36)	4.47 (1.57)
Male Talkers (Female Listeners)	52.0% (0.09)	3.62 (1.76)
Female Talkers (Male Listeners)	65.3% (0.33)	4.08 (1.55)
Male Talkers (Male Listeners)	49.3% (0.37)	3.38 (1.67)

measures of intelligibility, this potential difference is minor and does not interact with the effect of talker sex. Figure 2 highlights a similar pattern to Fig. 1 except that there appears to be more of a differential effect of talker sex depending on the listener sex.

The specific effects of listener sex and talker sex on intelligibility in terms of PWC and intelligibility rating (as highlighted by Figs. 1 and 2) were evaluated via linear mixed effects modeling (i.e., multilevel modeling), testing an interaction between talker sex and listener sex on each of PWC scores and intelligibility ratings (see Table 2 for means and standard deviations for both measures used in the models). The random effects structure was selected to accommodate the nested nature of the experimental design using the "maximal" random effects structure (Barr, Levy, Scheepers, & Tily, 2013). As such, the random effects structure allowed the effect of the talker sex and the effect of the listener sex to vary by sentence, and allowed the intercepts to vary by both the talker and the listener. Although intelligibility rating is a Likert-scale type measure, assumptions of the model were checked with no problematic distributions of the residuals for either the model predicting PWC or intelligibility ratings.

Table 3 Average intelligibility in percent words correct (PWC) and ratings for individual talkers

	PWC	Rating
Female 1	65%	4.14
Female 2	72%	4.44
Female 3	50%	3.76
Female 4	76%	4.64
Female 5	69%	4.34
Male 1	32%	2.58
Male 2	44%	3.25
Male 3	67%	4.2
Male 4	76%	4.58
Male 5	36%	2.99

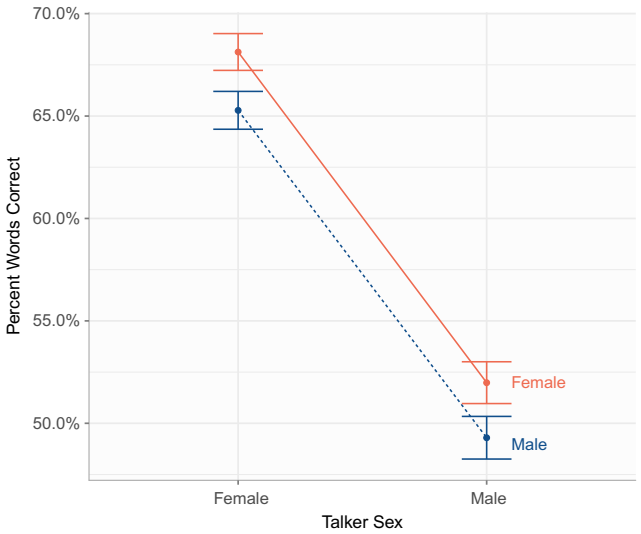

Fig. 1 Average percent words correct for talker sex by listener sex. Error bars represent +/- one standard error

Results indicate that for PWC scores, there is a significant effect of talker sex ($p < .001$; all p-values are derived using the Satterthwaite approximation for the degrees of freedom), but a non-significant effect of listener sex ($p = .217$), and a non-significant interaction between talker and listener sex ($p = .654$; see Fig. 1). For intelligibility rating scores there is an interaction approaching significance between talker and listener sex ($p = .062$; see Fig. 2), indicating that the effect of talker sex may depend on the sex of the

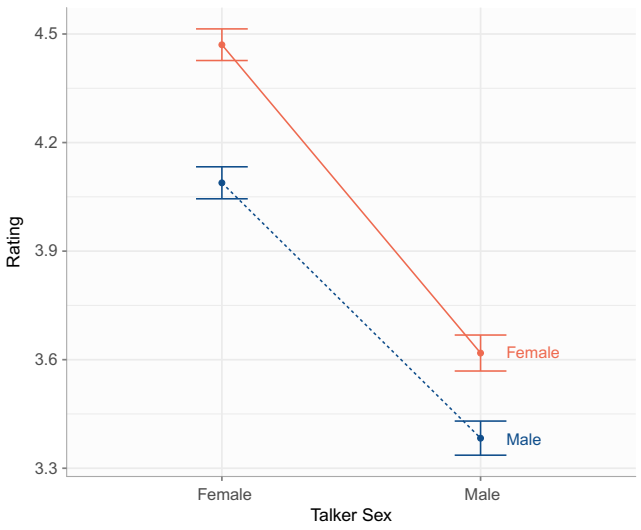

Fig. 2 Average rating for talker sex by listener sex that demonstrates the interaction between the two factors. Error bars represent +/- one standard error

listener. There was also a significant main effect of talker sex ($p < .001$) but not listener sex ($p = .087$) for rating. Finally, the correlation between PWC and intelligibility rating was moderately strong ($r = .658$, $p < .001$). See Table 4 for full model output.

Experiment 2: Crowdsourced data collection

Due to the results indicating a possible interaction between listener and talker sex in the rating scale portion of Experiment 1, a second experiment was carried out to examine these findings more closely – specifically, whether the finding from Experiment 1, that the effect of the listener sex on intelligibility rating depended on the talker sex (i.e., the effect of the listener sex was moderated by the talker sex), was the result of a particular bias of a relatively homogenous listener population (18- to 35-year-old individuals at one university in Northern Utah). As no interaction effect was observed in the PWC portion of Experiment 1, and rating scales are substantially more subjective in nature than PWC measures, only the rating scale portion was included in this follow-up experiment. Experiment 2 involved crowdsourcing the study online, thus, the listener participants in Experiment 2 represented a substantially more heterogenous population.

Method

Participants

Participants were recruited via crowdsourcing through the MTurk platform. Demographic information regarding age, sex, geographic region, and level of education of the participants is shown in Table 5. All participants were considered voluntary workers, protected through MTurk's participation agreement and privacy notice. We used a number of setup options regarding participant prerequisites, limiting participation to individuals with a previous approval rate of ≥ 99%. Any participants who self-reported a native language other than English or a history of speech, language, or hearing impairment were excluded from analysis. In addition, any participants who completed the task in less than 5 min or participants who responded with only one or two values from the rating scale were excluded from analysis. This data collection method was approved by Utah State University Institutional Review Board (IRB).

Table 4 Results of the linear mixed effects models for both Experiment 1 and Experiment 2

	Experiment 1		Experiment 2
	PWC	Intelligibility rating	Intelligibility rating
Model constant (Intercept)	0.589***	3.938***	3.951***
	(0.034)	(0.161)	(0.268)
Talker sex (Male)	0.150***	0.611***	0.641*
	(0.028)	(0.121)	(0.252)
Listener sex (Male)	0.015	0.158	-0.114***
	(0.012)	(0.090)	(0.029)
Interaction (Talker Sex × Listener Sex)	0.002	0.033	0.025
	(.004)	(0.012)	(0.025)
AIC	1429.21	17039.95	21397.94
BIC	1513.94	17124.67	21484.7
Log likelihood	-701.61	-8506.98	-10685.97
Num. obs.	5000	5000	5849
Random effects structure			
By Sentence			
Listener sex: Female	0.01	0.41	0.43
Listener sex: Male	0.01	0.31	0.1
Talker sex: Female	0.01	0.03	0.07
Talker sex: Male	0.02	0.13	0.26
Intercept by listener	0.01	0.39	-0.03
Intercept by talker	0.01	0.25	0.61

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The effects were all dummy coded, providing contrasts that relate to the reference groups (females in both talker and listener sex)

Table 5 Demographic distribution data expressed in percentage scores for listener participants (n=118)

Sex	Percentage
Males	62
Females	38
Age, y	
≥ 50	11
40–49	17
30–39	40
≤ 29	32
Ethnicity	
White	76
Indian	9
Latino	3
Prefer Not To Answer	12
Education	
Master's	8
Bachelor's	49
Attending College	2
High School Graduate	34
GED	4
Haven't Graduated High School	3
Region	
Midwest	19
Northeast	26
Pacific	9
Rocky Mountain	1
Southeast	34
Southwest	11

Stimuli and procedure

The stimuli and signal processing were identical to that of Experiment 1, except that a subset of the sentences was randomly chosen (five from each talker, for a total of 50 sentences). The decision to test with 50, rather than 100, sentences was made to reduce the likelihood of listener fatigue for participants who would be completing the experiment online with no experimenter present. As in Experiment 1, each participant was randomly assigned to one of five versions of talker-to-sentence correspondence. Participants were given a brief description of the task including the requirement of wearing headphones, being in a quiet environment, and information concerning remuneration (US$2), and then directed to a webpage loaded with a listener-perception application hosted on a secure university-based web server. Before beginning the study, individuals were required to read through the IRB-approved consent form. By clicking *Agree*, individuals indicated that they had read and understood the information provided in the consent from and voluntarily agreed to participate. Participants were then required to complete a brief questionnaire regarding demographic information and questions related to inclusion/exclusion criteria.

Participants were informed that they would be presented with 50 sentences spoken by different talkers in background noise, and a test noise was played so that the participant could adjust the sound output to a comfortable level. The participants were instructed to report their impression of the talker's voice. For the subjective impression, the participants were asked the identical statement as was asked in Experiment 1, "To what extent do you agree with this statement? The talker is easily understood." They were asked to rate each sentence on a 7-point scale, with 1 representing "strongly disagree" and 7 representing "strongly agree," with radio dials representing each of the numerical options. The participants were told to focus on the talker's voice, not on the effect of the noise. The experiment took 7 min on average for participants to complete.

Results of Experiment 2

To begin, consistency of ratings across Experiments 1 and 2 (laboratory-based vs. crowdsourcing) was evaluated via a mixed effects model. Overall subjective ratings of intelligibility did not differ between laboratory-based and crowdsourced data collection methods ($p = .576$) when controlling for the nesting of the data and talker and listener sex (as done previously with the random effect structure). Further, a correlation between the individual ratings for each talker across the two data collection methods was extremely high ($r = .94$). However, differences were noted in the specific patterns across talker and listener sex on ratings.

For Experiment 2, overall average rating scores for male and female talkers, as well as average rating scores broken down by listener sex for each talker group, are shown in Table 6, and individual rating scores for each of the ten talkers are shown in Table 7. Average rating scores for male and female talkers, as a function of listener sex, are shown in Fig. 3. The specific effects of listener sex and talker sex on

Table 6 Overall intelligibility ratings. Standard deviations are given in parentheses

	Rating
Overall talker group	
Female talkers	4.62 (1.69)
Male talkers	3.36 (1.86)
Talker group by listener group	
Female talkers (Female listeners)	4.52 (1.79)
Male talkers (Female listeners)	3.21 (1.90)
Female talkers (Male listeners)	4.68 (1.65)
Male talkers (Male listeners)	3.45 (1.83)

Table 7 Average ratings for individual talkers

Talker	Rating
Female 1	4.64
Female 2	4.88
Female 3	4.58
Female 4	4.63
Female 5	4.37
Male 1	2.21
Male 2	3.06
Male 3	4.38
Male 4	4.61
Male 5	2.56

intelligibility in terms of intelligibility rating were evaluated via linear mixed effects modeling. This linear mixed effects model tested an interaction between talker sex and listener sex on intelligibility rating (see Table 6 for means and standard deviations of intelligibility rating) and main effects of both. As in Experiment 1, the maximal random effect structure was used, which allowed the effect of the talker sex and the effect of the listener sex to vary by sentence and allowed the intercepts to vary by both the talker and the listener. Although intelligibility rating is a Likert-scale type measure, assumptions of the model were checked with no problematic distributions of the residuals.

Unlike Experiment 1, results indicate that for the crowdsourced data the interaction between talker and listener sex on listener rating is not approaching significance ($p = .207$), indicating that the effect of talker sex did not depend on the listener sex. There were significant main effects of listener sex ($p < .001$) and talker sex ($p = .033$). As found before, female talkers were rated as more intelligible than male talkers. Conversely to Experiment 1, male listeners rated the intelligibility for both female and male talkers higher than female listeners. Again, see Table 4 for full model output.

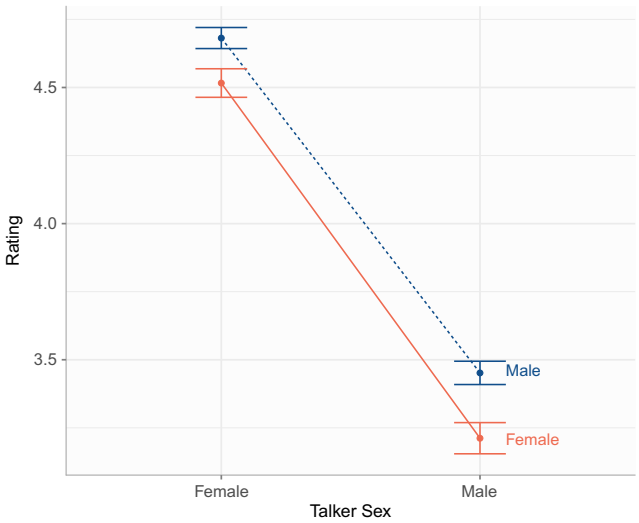

Fig. 3 Average rating for talker sex by listener sex. Error bars represent +/- one standard error

Finally, to further assess the differences across the samples, the two data sets were combined and analyzed with linear mixed effects modeling with the sample label as one of the indicators. Specifically, all two-way and three-way interactions between the sample indicator, the listener sex, and the talker sex were assessed. The three-way interaction was not significant ($p = .775$) nor was the two-way interaction between talker sex and the sample indicator ($p = .326$). However, the two-way interaction between listener sex and the sample indicator was significant ($p = .006$), demonstrating that the effect of the listener sex on the subjective ratings of intelligibility differed across the two samples.

Discussion

The results of Experiment 1 confirm the findings of other studies that have indicated that talker sex does influence intelligibility, with females having higher overall scores (e.g., Bradlow et al., 1996; Ferguson, 2004; Markham & Hazan, 2004). In fact, the observed mean difference between male and female talkers was greater than many of these past studies. In addition, although there is not an effect of listener sex on intelligibility in terms of objective measurement (i.e., PWC), there does appear to be an effect of listener sex in terms of subjective measurement (i.e., rating scales) for Experiment 2. In addition, whereas female listeners gave higher intelligibility ratings in Experiment 1, male listeners gave higher intelligibility ratings in Experiment 2. For Experiment 1, there was an interaction between talker and listener sex, with male listeners, relative to female listeners, rating female talkers as more difficult to understand. However, this interaction did not occur in Experiment 2. This demonstrates an important point – the particular measure employed in a study may play a meaningful role in terms of sex effects in speech intelligibility. This observed difference between PWC and rating scores in Experiment 1 occurred despite the use of the same listeners, and both measures occurring within the same experimental session. In addition, the overall correlation between PWC and rating scores was relatively strong (.57), as has been observed by others (Yorkston & Beukelman, 1978). While this relationship between objective and subjective measures of speech intelligibility exists, it does not suggest that the measures capture precisely the same construct. Nor does it suggest that these measures can be used interchangeably. Indeed, the current findings suggest that the intelligibility measures may become increasingly divergent in certain samples. Subjective rating scales, for example, are more vulnerable to bias, and we advance the idea that such biases may be amplified when using homogenous listener samples.

In the current study, speaking rate was tightly controlled (averages of 3.56 and 3.58 for male and female talkers, respectively). This is important, as speaking rate has been shown to impact listener judgments of both male and female talkers, such as differences in perceived competence of talkers (e.g., Brown, 1980; Smith et al., 1975). Indeed, listener judgments of talkers appear to be quite complex and a result of interactions across multiple acoustic features (Parker & Borrie, 2018), and biases concerning male and female speech can affect listener perceptions of many talker attributes. For example, expectations regarding traditionally "male" or "female" speech can play a role, as listeners have been shown to rate higher-pitched female voices and lower-pitched male voices as more attractive than lower-pitched female and higher-pitched male voices (Hodges-Simeon et al., 2010; Re et al., 2012). Therefore, gender norms and cultural expectations may have played a role in the results of the subjective rating scale of the current study.

One of the most interesting findings of the current study is this potential interaction, or difference in subjective ratings of male and female talkers by male and female listeners in Experiment 1. This motivated Experiment 2 – specifically the use of a more heterogenous listener sample, collected via online crowdsourcing, on intelligibility ratings of the same talkers in Experiment 1. For the laboratory portion of the study, the way that male listeners rated the male and female talkers somewhat differed from the way that female listeners rated the male and female talkers, and the effect of listener sex on ratings differed across the two experiments. Given that the listener group recruited for the laboratory portion of the study was fairly homogeneous, this finding may in part represent a particular bias of this population. The participants for this experiment were all recruited from the student population at Utah State University, which is made up of 82% students who identify as Caucasian (all but one of our participants were Caucasian). Further, 80% of the students are from either Utah (77%) or the neighboring state of Idaho (3%) (Utah State University Office of Analysis, Assessment and Accreditation, 2018). These participants were all of a similar age, and had a similar level of education (current undergraduate students). In addition, the large majority of students at this university (approximately 70%) are members of the Church of Jesus Christ of Latter-day Saints (Utah State University, 2015). With such a large proportion of students belonging to this religion, cultural differences or differences in the understanding of traditional gender roles may be another possible factor (Beaman, 2001; Sumerau & Cragun, 2014). This may have played a role in the rating scale results of Experiment 1, in which male listeners rated female talkers relatively lower than female listeners rated female talkers. Alternatively, for the listeners recruited via crowdsourcing (and therefore representing a much more diverse group of individuals), this potential bias was not observed. Although it cannot be known for certain whether the homogeneity of listener participants in Experiment 1 was the primary factor in the observed differences between the two experiments, it remains an important point that there may be an influence of the particular demographic makeup of the subject population for speech perception studies, and findings of studies that use a homogenous group of subjects should be interpreted with caution. Crowdsourcing of speech perception studies is becoming increasingly common (e.g., Borrie, Baese-Berk, Van Engen, & Bent, 2017; Yoho & Borrie, 2018), and has been validated as an effective and reliable means of collecting human listener data (Lansford et al., 2016; McAllister Byun, Halpin, & Szeredi, 2015; Stole & Strand, 2016). An analysis of the laboratory-based and crowdsourced data collection in the current study confirmed this, showing that even though the specific pattern of results differed between the laboratory-based and crowdsourced data, the overall rating scores for the different talker and listener groups were not significantly different. This suggests effects due to methodology or inclusion of listener groups of different demographics may not be obvious when assessing overall scores but may have an important impact on the conclusions. Further, there was an extremely high correlation ($r = .94$) between the individual ratings for each talker across the two data collection methods. These comparable results afford important validation for the use of crowdsourcing websites to collect speech perception data. One of the key advantages of crowdsourcing is easy access to a highly heterogenous sample of listener subjects, while still allowing for control over certain demographic criteria.

Although there was a significant effect of talker sex, with females showing overall higher PWC intelligibility by nearly 16 percentage points on average in Experiment 1, it should be noted that there was a large degree of variability across the individual talkers, both in terms of PWC and rating scores (see Table 3). In fact, the degree of variability observed across talker rating, for example, is even slightly higher here than in some previous studies. The talker with the highest rating from Experiment 1 (Female 4) had an overall rating of 4.64 and the talker with the lowest rating (Male 1) had an overall rating of 2.58. That is a wider range than was found by Ferguson and Morgan (2018), which also had listeners rate intelligibility of male and female speech using a 7-point scale. The highest and lowest rated talkers in that study (when speaking conversationally) were on average 5.1 and 3.42, respectively. Although the magnitude of overall PWC difference was larger for this study than for some others (e.g., Bradlow et al., 1996), this finding of "high" and "low" intelligibility talkers within each sex group replicates the previous finding of Bradlow and colleagues. The finding here was despite the fact that talkers were selected to control for overall speaking rate, and to reflect classically "male" or "female" pitches. Of the talkers with the two highest intelligibilities, one was female and one was male (both 76% correct). This variability

reaffirms that the question of whether one sex is more intelligible than another is not an absolute, and that comparisons in speech intelligibility made across speech perception studies using different talkers, even talkers of the same sex, may be more problematic than previously appreciated. This may also partially explain the conflicting findings observed in previous literature concerning which sex may be more intelligible, and the magnitude of such effects (Bradlow et al., 1996; Gengel & Kupperman, 1980; Hazan & Markham, 2004; McCloy et al., 2015). For example, Ellis et al. (1996), which examined subjective impressions of male and female speech, only used one talker of each sex. Given the high variability observed here across talkers, it is possible that the single talkers chosen for that study were not wholly representative of their respective sex overall, and that one individual talker may in fact never be fully representative. A final important consideration of this large degree of variability is that if only one talker is selected as stimulus, as is often the case in both speech perception research and clinical audiological testing, it may skew the results to appear as if intelligibility is significantly higher or significantly lower than it would be on average. For example, one of the most commonly utilized recordings[1] of the CID-W22 monosyllable word list (Hirsh et al., 1952) was found to be the least intelligible talker in the study by Gengel and Kupperman (1980). These findings also call into question the relatively common practice of utilizing monitored live voice in audiologic speech testing (a process by which the audiologist or clinician uses their own voice instead of a standard recording for speech intelligibility testing), as comparisons across testing sessions with different examiners may be highly influenced by differences in talker intelligibility.

Again, an interesting point of note is the difference in overall rating of male and female talkers between Experiments 1 and 2. Whereas in Experiment 1 the interaction between talker and listener sex was approaching significance, this was not the case with the crowdsourced data from Experiment 2. In addition, the listener group with the overall highest intelligibility ratings differed between the two experiments. Although it is possible that these differences may have been a result of the methodologies employed (crowdsourced vs. laboratory), they more likely reflect inclusion of a more heterogenous listener group in Experiment 2. Regardless of the reason, it highlights another consideration when comparing subjective data that has been collected via different means or with different subject populations.

An important question remains – why do female talkers on average display an overall higher intelligibility than male talkers? It is possible that intrinsic acoustic characteristics play a role, and evidence exists to support the notion that talker-specific characteristics that make an individual more or less intelligible than another play a role despite whether the talker is speaking in a native language or in an accented non-native language (Bradlow et al., 2018). Bradlow et al. (1996), however, found no relationship between F0 and overall intelligibility, and the speaking rates of the selected talkers in the current study were very similar to each other. The acoustic data from the current study on fundamental frequency reveals that females, on average, displayed greater pitch variation and range than males. Studies have shown that pitch variation and range contribute to speech intelligibility both in quiet and in noise (e.g., Bunton et al., 2001; Laures & Weismer, 1999; Miller et al., 2010). The results of the current study may be in accord with the data from Byrd (1994) and Ferguson (2004), which indicate that females tend to more carefully articulate when recording speech materials, and that the "clear" speech of females is perceived to have significantly higher clarity than that of male speech (Ferguson & Morgan, 2018). One note of considerable interest is that the male speaker with the highest intelligibility in the current study (76%) was an advanced student in a graduate audiology program who had been trained to speak clearly to listeners with hearing loss. This finding does not entirely support the result from Ferguson (2004), which found that talkers who had more experience speaking to listeners with hearing loss were no more intelligible; however, it is possible that speakers who have been specifically trained to do so are at more of an advantage than simply those (primarily much older) individuals who have more passive experience due to having family members or companions with hearing loss.

Limitations and future directions

In addition to differences in subject demographics between Experiments 1 and 2, there were also other important methodological differences that could have played a role in the findings. The data collection method (online vs. laboratory) differed between the two. Several studies in the area of speech perception have found high agreement between data collected in these two different modalities (Lansford et al., 2016; McAllister Byun, Halpin, & Szeredi, 2015; Slote & Strand, 2016), and our own analysis found good agreement in overall ratings between the two and a strong correlation between the ratings of individual talkers. In the current study, the participants for the crowdsourced experiment self-reported information, such as native language, ethnicity, and history of speech or language impairments. Although it is possible that some participants misrepresented these facts, this information is often self-reported for laboratory-based studies as well. Lastly, for Experiment 2, each talker only produced 25 sentences across all listeners instead of the 50 sentences across all listeners in Experiment 1, reducing the overall representation of each talker.

[1] This recording is the voice of Dr. Ira Hirsh

The aim of the current study was to broadly determine whether there were listener or talker sex effects in speech intelligibility, and therefore several demographic details of the listeners, such as religion, gender identification, or place of origin, were not collected. Although the purpose of the current study was not to explore issues specifically related to listener demographics such as these, the differing rating scale results between Experiments 1 and 2 provide strong motivation for a subsequent study to examine these factors, particularly given the specific gender and other cultural norms in Northern Utah.

The inclusion of only ten talkers (five of each sex) in the current study allowed for productions of several sentences from each talker to be presented to each listener, and therefore talker-to-talker variability could be examined. Although ten talkers allows for more diverse representation of each sex than a smaller number of talkers used in some previous studies (e.g., Ellis et al., 1996), it is still a limited set that restricts major generalization of the current findings. In particular, the finding of such high talker-to-talker variability in the set of talkers utilized here provides motivation for a future, larger-scale study utilizing a substantially higher number of talkers. Such larger scale studies should also include a comprehensive acoustic analysis of the speech stimuli, including an evaluation of rhythmic, articulatory, and phonatory dimensions of speech. Lastly, an evaluation of whether these findings hold for languages other than American English, or if the findings of the subjective rating scale remain when the speech is presented in a language that is unfamiliar to the listeners, as many sociolinguistic factors may play a role.

Finally, the specific selection of the random effects structure in the linear mixed effects models has benefits and limitations that should be acknowledged. First, there appear to be benefits of maximizing the random effects structure (Barr et al., 2013), with evidence suggesting that it causes a more conservative type I error rate (i.e., reduces the risk of a false-positive conclusion). However, in this the power of the significance tests are also reduced (i.e., increasing type II error). Given that the study was exploring interactive effects between listener and talker sex, the more hindering problem may be power as it requires far more statistical power to detect interactions than it does to detect main effects. Therefore, the more powered analyses were used herein. In addition, the further analyses comparing the two samples' ratings corroborate that there are important differences across the samples. However, additional studies should be carried out to better elucidate the impact that nuances of the sample subject group have on results in the field, especially for subjective measures.

Conclusions

It appears that in accord with many of the results of previous literature, female talkers are more intelligible overall in terms of percent words correct. However, there is a large degree of variability across talkers, and any individual talker may be substantially more or less intelligible than their sex group as a whole. Additional investigations are needed to determine whether this talker sex effect generalizes over a more diverse representation of talkers. In addition, the subjective rating of intelligibility of male and female speech may differ as a function of the subject group employed. Although there does not appear to be an effect of listener sex in terms of an objective measure of speech intelligibility (PWC) for the conditions tested here, there is an effect of listener sex on the more subjective rating scales for the crowdsourced data, indicating again that results may vary based on the particular intelligibility measure employed. In addition, although it cannot be known for certain what role bias played in the current results, biases may exist within a particular homogenous listener group that may skew results, providing additional evidence for the fact that convenience samples are not representative of the population as a whole, and diversity and heterogeneity amongst listeners is often ideal. Crowdsourcing offers an important and effective medium through which to recruit these heterogenous groups. These listener, talker, and methodology effects should be considered in models of speech perception and when making comparisons across studies.

Acknowledgments This paper was written with partial support from the National Institute of Deafness and Other Communication Disorders, National Institutes of Health Grant No. R21 DC 016084 (awarded to S.A.B.). We gratefully acknowledge our research assistants Nicole Thiede and Monica Muncy for data analysis and manuscript preparation assistance.

References

Allmark, P. (2004). Should research samples reflect the diversity of the population? *Journal of Medical Ethics, 30*, 185–189.

American National Standard Institute (1997). ANSI S3.5 (R2007). American National Standard Methods for the Calculation of the Speech Intelligibility Index (American National Standards Inst., New York).

American Speech-Language-Hearing Association. (1997). Guidelines for audiologic screening.

Bacon, S. P. (1990). Effect of masker level on overshoot. *The Journal of the Acoustical Society of America*, *88*(2), 698-702.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3).

Beaman, L. G. (2001). Molly mormons, mormon feminists and moderates: Religious diversity and the latter day saints church. *Sociology of Religion*, *62*(1), 65-86.

Bleecker, M. L., Bolla-Wilson, K., Agnew, J., & Meyers, D. A. (1988). Age-related sex differences in verbal memory. *Journal of Clinical Psychology, 44*(3), 403-411.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0. 36.

Borrie, S. A., Baese-Berk, M., Van Engen, K., & Bent, T. (2017). A relationship between processing speech in noise and dysarthric speech. *The Journal of the Acoustical Society of America, 141*(6), 4660-4667.

Borrie, S.A. and Schäfer, M.C.M. (2017). Effects of lexical and somatosensory feedback on long-term improvements in intelligibility of dysarthric speech. *Journal of Speech, Language, and Hearing Research, 60,* 2151-2158.

Bradlow, A. R., Blasingame, M., & Lee, K. (2018). Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 9*(1).

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.

Brown B.L. (1980). Effects of speech rate on personality attributions and competency evaluations. In: Giles, H., Robinson, W. P., Smith, P. (Eds.) *Language: Social psychological perspectives* (pp. 293–300).

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics*, *15*(3), 181-193.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15*(1-2), 39-54.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech, Language, and Hearing Research*, *14*(3), 565-577.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America, 119*(3), 1562-1573.

Dehan, C. P., & Jerger, J. (1990). Analysis of gender differences in the auditory brainstem response. *The Laryngoscope, 100*(1), 18-24.

Don, M., Ponton, C. W., Eggermont, J. J., & Masuda, A. (1993). Gender differences in cochlear response time: An explanation for gender amplitude differences in the unmasked auditory brain-stem response. *The Journal of the Acoustical Society of America, 94*(4), 2135-2148.

Ellis, L., Fucci, D., Reynolds, L., & Benjamin, B. (1996). Effects of gender on listeners' judgments of speech intelligibility. *Perceptual and Motor Skills, 83*(3), 771-775.

Ferguson, S. (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.

Ferguson, S. H., & Morgan, S. D. (2018). Talker differences in clear and conversational speech: Perceived sentence clarity for young adults with normal hearing and older adults with hearing loss. *Journal of Speech, Language, and Hearing Research, 61*(1), 159-173.

Fogerty, D. (2011). Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure. *The Journal of the Acoustical Society of America, 129*(2), 977-988.

Garofolo, J. S. (1988). DARPA TIMIT acoustic-phonetic speech database. *National Institute of Standards and Technology (NIST), 15,* 29-50.

Gengel, R. W., & Kupperman, G. L. (1980). Word discrimination in noise: Effect of different speakers. *Ear and Hearing, 1*(3), 156-160.

Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., & van Lieshout, P. (2013). Normative voice data for younger and older adults. Journal of Voice, 27(5), 545-555.

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.

Healy, E. W., Yoho, S. E., & Apoux, F. (2013). Band importance for sentences and words reexamined. *The Journal of the Acoustical Society of America, 133*(1), 463-473.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17,* 321–337.

Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2010). Different vocal parameters predict perceptions of dominance and attractiveness. Human Nature, 21(4), 406-427.

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17,* 225–246

Klasner, E. R., & Yorkston, K. M. (2005). Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. *Journal of Medical Speech-Language Pathology*, *13*(2), 127-140.

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.

Lansford, K. L., Borrie, S. A., & Bystricky, L. (2016). Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. *American Journal of Speech-Language Pathology, 25*(2), 233-239.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America*, *59*(3), 675-678.

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. Journal of Speech, Language, and Hearing Research, 42(5), 1148-1156.

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.

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53,* 70–83.

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.

McFadden, D. (1998). Sex differences in the auditory system. *Developmental Neuropsychology, 14*(2-3), 261-298.

McFadden, D., Pasanen, E. G., Maloney, M. M., Leshikar, E. M., & Pho, M. H. (2018). Differences in common psychoacoustical tasks by sex, menstrual cycle, and race. *The Journal of the Acoustical Society of America, 143*(4), 2338-2354.

McRoberts, G. W., & Sanders, B. (1992). Sex differences in performance and hemispheric organization for a nonverbal auditory task. *Perception & Psychophysics, 51*(2), 118-122.

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America*, *128*(1), 435-443.

National Institutes of Health. (2017). NIH guidelines on the inclusion of women and minorities as subjects in clinical research. NIH Grants Policy October, 2017; AII-33. Available at https://grants.nih.gov/grants/policy/nihgps/nihgps.pdf. Accessed 17 June 2018.

Parker, M. A. & Borrie, S. A. (2018). Judgements of intelligibility and likeability of young adult female speakers of American English: The influence of vocal fry and the surrounding acoustic-prosodic context. *Journal of Voice*, 32, 538-545.

Rademacher, J., Morosan, P., Schleicher, A., Freund, H. J., & Zilles, K. (2001). Human primary auditory cortex in women and men. *Neuroreport, 12*(8), 1561-1565.

Re, D. E., O'Connor, J. J., Bennett, P. J., & Feinberg, D. R. (2012). Preferences for very low and very high voice pitch in humans.

PLoS One, 7(3), e32719. https://doi.org/10.1371/journal.pone.0032719

Rogers, D. S., Harkrider, A. W., Burchfield, S. B., & Nabelek, A. K. (2003). The influence of listener's gender on the acceptance of background noise. *Journal of the American Academy of Audiology, 14*(7), 372-382.

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *The Journal of the Acoustical Society of America*, *43*(5), 1178-1179.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621-640.

Slote, J., and Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods, 48,* 553-566.

Smith, B. L., Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Effects of speech rate on personality perception. Language and speech, 18(2), 145-152.

Sumerau, J. E., & Cragun, R. T. (2014). The hallmarks of righteous women: Gendered background expectations in the Church of Jesus Christ of Latter-Day Saints. *Sociology of Religion*, *76*(1), 49-71.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85,* 1699–1707.

Utah State University (2015). *Interfaith diversity experiences & attitudes longitudinal survey.* Retrived from https://interfaith.usu.edu/files/Utah%20State%20University.pdf

Utah State University Office of Analysis, Assessment and Accreditation. (2018). *Utah State University Fall 2017 Enrollment Analysis.* Retrieved from http://www.usu.edu/aaa/enroll_infographic.cfm

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native-and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America, 121*(1), 519-526.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America, 54*(5), 1248-1266.

Yoho, S. E., & Borrie, S. A. (2018). Combining degradations: The effect of background noise on intelligibility of disordered speech. *The Journal of the Acoustical Society of America, 143*(1), 281-286.

Yoho, S. E., Healy, E. W., Youngdahl, C. L., Barrett, T. S., & Apoux, F. (2018). Speech-material and talker effects in speech band importance. *The Journal of the Acoustical Society of America, 143*(3), 1417-1426.

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders, 11*(6), 499-512.