Supplementary Methods

Study 1

Face stimuli were generated using FaceGen Modeler (Singular Inversions) to appear somewhat sex-ambiguous. Ten male faces were generated at 60%-Male/40%-Female and 10 female faces at 60%-Female/40%-Male.

Ten male and 10 female speech samples of American dialect and neutral tone were obtained from the International Dialects of English Archive (http://web.ku.edu/~idea). Clips of 2000 ms were extracted from each sample, the content of which was selected to be naturalistic for a first-impression encounter (e.g., “My family’s origins are pretty interesting.”). We used Praat software (http://www.fon.hum.uva.nl/praat) to morph each male voice into a sex-typical (masculine) version (formant shift ratio:1/1.1) and a sex-atypical (feminine) version (formant shift ratio:1.1), and morphed each female voice into a sex-typical (feminine) version (formant shift ratio:1.1) and a sex-atypical (masculine) version (formant shift ratio:1/1.1), consistent with prior work (e.g., Groen et al., 2008). We did not manipulate median pitch because its alteration tends to sound computer-like and artificially synthesized.

Study 2

In Study 2, an additional 20 sex-ambiguous faces (10 male, 10 female) were generated using the same procedure outlined in Study 1. These were added to the 20 sex-ambiguous faces used in Study 1, resulting in 40 faces (20 male, 20 female) available for the task. These 40 faces were randomly paired (without replacement) with the 40 voice stimuli (20 sex-typical, 20 sex-atypical) of Study 1. These 40 face-voice pairs were randomly divided into 20 experimental trials and 20 control trials (each containing 5 sex-typical male, 5 sex-typical female, 5 sex-atypical male, and 5 sex-atypical female pairs). These 40 trials were presented across 5 blocks.

Block 1 contained all 20 experimental trials, which involved categorizing the face-voice pairs as MALE or FEMALE (thus identical to the trials of Study 1). The 20 control trials involved decisions between the correct sex-category and either a FARM or JUNGLE response (rather than the opposite sex-category). Which stimuli were paired with a FARM or JUNGLE response was randomized. These trials were presented across Blocks 2 through 5. Each of these control blocks contained 10 trials, half of which involved the presentation of the human face-voice pairs (which served as the control trials) and the other half of which were filler trials (farm and jungle animals). Filler trials involved the presentation of an animal's face and voice. To create the filler trials, images and sounds of 5 farm animals (e.g., cow) and 5 jungle animals (e.g., monkey) were obtained from public domain websites. Images were cropped as to preserve only the animal's face and sized identically as the human face stimuli, and 2000 ms sound clips were extracted from an animal's vocal stream (e.g., for a cow face, 2000 ms of "mooing").

Block 2 involved MALE vs. FARM decisions, Block 3 involved MALE vs. JUNGLE decisions, Block 4 involved FEMALE vs. FARM decisions, and Block 5 involved FEMALE vs. JUNGLE decisions. The 20 trials involving human face-voice pairs across these 4 blocks comprised the control trials (whereas the other 20 trials served as filler trials). The order of the 5 blocks was randomized, and which category appeared on the left/right was also randomized. Thus, control trials were essentially the same as experimental trials, except that the opposite response alternative was FARM or JUNGLE rather than the opposite sex-category. If the target was human, participants were instructed to categorize the face's sex, and if the target was a nonhuman animal, they were instructed to categorize it as a farm or jungle animal. As in Study 1, participants were instructed to only use the voice if it could help resolve the face's sex, as correct

responses were based on the face. The mouse-tracking procedures were identical to those in Study 1.

Supplementary Results

Categorization accuracy, initiation times, and response times were submitted to repeated-measures ANOVAs with factors of sex-typicality (typical/atypical) and condition (experimental/control).

Categorization accuracy

Zero errors were made in the sex-typical and sex-atypical control trials. A main effect of condition [$F(1, 24) = 59.16$, $p<.0001$] indicated that more errors were made in the experimental condition than the control condition. A main effect of sex-typicality [$F(1, 24) = 17.59$, $p < .0001$] and an interaction [$F(1, 24) = 17.59$, $p < .0001$] indicated that more errors were made for sex-atypical experimental trials (M = 16.4%, SE = 2.7%) than for sex-typical experimental trials (M = 3.8%, SE = 1.0%), whereas no errors (0%) were made in the control trials.

Initiation time

A main effect of condition [$F(1, 24) = 11.53$, $p < .01$] indicated that initiation times were shorter for control trials ($M_{\text{sex-typical}}$ = 255 ms, $SE_{\text{sex-typical}}$ = 20 ms; $M_{\text{sex-atypical}}$ = 245 ms, $SE_{\text{sex-atypical}}$ = 19 ms) than experimental trials ($M_{\text{sex-typical}}$ = 331 ms, $SE_{\text{sex-typical}}$ = 37 ms; $M_{\text{sex-atypical}}$ = 317ms, $SE_{\text{sex-atypical}}$ = 35ms). The main effect of sex-typicality was not significant [$F(1, 24) = 1.53$, $p = .23$], nor was the interaction [$F(1, 24) = 0.04$, $p = .85$].

Response time

A main effect of condition [$F(1, 24) = 73.73$, $p < .0001$] indicated that response times were quicker for control trials than experimental trials, and a main effect of sex-typicality [$F(1,24) = 6.73$, $p < .05$] indicated that response times were quicker for sex-typical trials than

sex-atypical trials. These main effects were qualified by a significant interaction [$F(1, 24) = 5.28$, $p < .05$]. Response times for sex-atypical experimental trials (M = 1493 ms, SE = 55ms) were longer than those for sex-typical experimental trials (M = 1399 ms, SE = 55ms), $t(24) = 3.34$, $p < .01$. However, response times for sex-atypical control trials (M = 1165 ms, SE = 48 ms) were no different than those for sex-typical control trials (M = 1143 ms, SE = 45 ms), $t(24) = 0.83$, $p = .41$.

Distributional analyses of Maximum Deviation (MD)

The continuous-attraction effect of Study 2, as indicated by trajectories' larger MD for sex-atypical experimental trials relative to sex-typical experimental trials, was unimodally distributed. Specifically, MD distributions in the sex-atypical and sex-typical conditions were within the bimodality-free zone: b's = .518 and .400, respectively (Figure 4). Further, a Kolmogorov-Smirnov test indicated that the shapes of these two distributions were statistically indistinguishable (D = .07,p = .62), ruling out the possibility that trajectories for sex-atypical experimental trials possessed latent bimodality. These analyses ensure that the continuous-attraction effect was not spuriously produced by a combination of discrete-like movements.

Main effects of MD

The significant condition × sex-typicality interaction in MD from the ANOVA of Study 2 is described in the manuscript. The main effect of sex-typicality was significant [$F(1, 24) = 8.25$, $p < .01$] as well, with sex-atypical trials showing more MD relative to sex-typical trials. The main effect of condition was also marginally significant [$F(1, 24) = 3.53$, $p = .07$], with experiments trials showing a trend of more MD than control trials.

References

Groen, W. B., van Orsouw, L., Zwiers, M., Swinkels, S., van der Gaag, R. J., & Buitelaar, J. K. (2008). Gender in voice perception in autism. *Journal of Autism and Developmental Disorders, 38*, 1819-1826.