Perception of speaker sex in re-synthesized children's voices

Peter Assmann[1] Michelle Kapolowicz[1] David Massey[1] Santiago Barreda[2] Terrance Nearey[3]

[1]University of Texas at Dallas [2]University of Arizona [3]University of Alberta

Background

Fundamental frequency (F0) and average formant frequencies (FF) provide important cues for the perception of speaker sex. Previous experiments with vocoded adult voices[1] have indicated that upward scaling of F0 and FFs increases the probability that a voice will be perceived as female while downward scaling increases the probability that the voice will be perceived as male. The present study extends these manipulations to children's voices.

Stimuli

Syllable stimuli

• **Age**	5-18 years (14 age levels)
• **Sex**	Equal numbers of male & female speakers
• **Vowel**	/hid/, /hɑd/, and /hud/
• **Talker**	5 speakers per age group, drawn from a vowel database of 208 speakers[2]

Synthesis method

Each syllable was processed using the STRAIGHT vocoder[2] to scale F0 and FFs to the *opposite sex average* based on acoustic measurements for each age level.

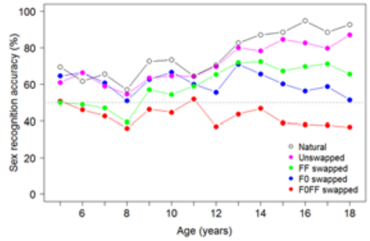
Synthesis condition

- **Unswapped**: neither F0 nor FFs altered
- **FF swapped**: mean F0 swapped, FFs unchanged
- **F0 swapped**: mean FFs swapped, F0 unchanged
- **F0FF swapped**: both FFs and F0 swapped

Participants and procedure

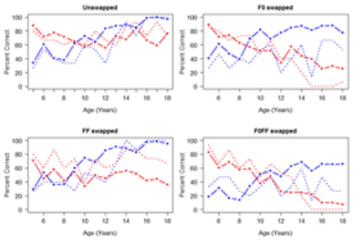
- Adult listeners (different sets of 5 males and 5 females in each synthesis condition) volunteered or enrolled for partial course credit.
- Each participant heard 420 stimuli, with syllables randomly interspersed; stimuli presented monaurally using headphones with Tucker-Davis System 3 and RP2.1 hardware.
- Listeners used a 2-alternative button box to indicate speaker sex and rated their confidence on a 5-point scale.

Condition x Age interaction



- Similar accuracy for synthesized syllables (purple circles) and natural syllables (black circles) from our previous study[3].
- Recognition accuracy improves gradually with age starting around the onset of puberty.
- Swapping *either* F0 or FFs produced a decline in accuracy, with a larger effect of F0.
- Swapping *both* F0 and FFs produced a further decline. However, this manipulation did not invert the function around the 50% line as would be predicted if F0 and FFs were the only cues used by listeners.

Flexible Listening Strategies?

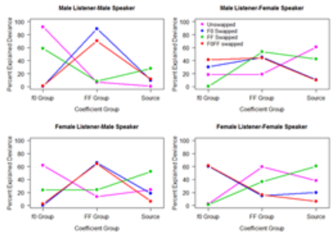


Acoustic variables

F0 group
- Log average fundamental frequency (Hz)
- linear spline with a single knot corresponding to 175 Hz

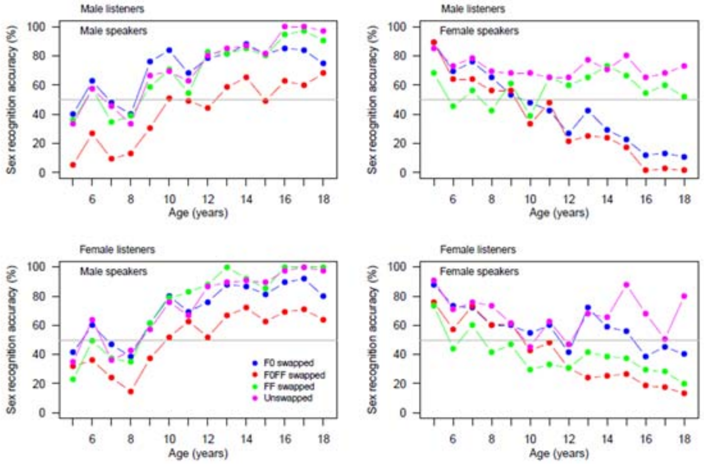
FF group
- Geometric mean of F1 F2 F3 (Hz)
- vowel

Source group
- Corrected magnitude difference between harmonics 1 and 2 (dB)
- Corrected magnitude difference between harmonic 1 and F3 peak (dB)
- Cepstral pitch prominence (dB)
- Harmonics-to-noise ratio below 500 Hz (dB)



- Why does accuracy remain high in swapped conditions for male speakers when important cues for speaker sex recognition have been removed? One possibility is that listeners are altering the weights they assign to F0 and FFs for stimuli with conflicting cues for speaker sex.
- The figure above shows predicted and observed accuracy using a logistic regression model from our previous experiment[3] with natural syllables, using the acoustic measures shown in the table on the right as predictors.
- Consistent with this possibility, we see a relatively good fit for the Unswapped condition but a poor fit for the Swapped conditions.
- The plots above show the percent deviance explained by each coefficient group. Within each panel, differences reflect changes in listening strategies, even when speaker and listener sex are isolated. Differences between panels highlight further variations as a function of speaker and listener sex.
- For example, male speakers (left panels), in conditions where F0 is swapped the F0-related variables make a smaller contribution compared to FF-related variables.

Age x Condition x Speaker Sex x Listener Sex interaction



- For unswapped stimuli, accuracy increased with age for male speakers (purple circles, left panels) but was flatter and/or more variable for female speakers (purple circles, right panels).
- Overall, male speakers were more resistant to the effects of F0 and FF swapping than female speakers. One possibility is that male speakers with raised F0 or FFs could plausibly be interpreted as younger males, while F0 or FFs in the older male range are more likely to be heard as male.
- Male and female listeners showed similar patterns of accuracy for male speakers as a function of F0 and FF. However, F0 swapping produced a larger decline than FF swapping for males listening to female voices, while females listening to female voices assigned greater weight to FFs.

Summary and conclusions

- Consistent with predictions, listeners showed reduced accuracy in conditions where F0 and FF were swapped.
- However, contrary to findings for adult speakers[1], swapping both F0 and FF did not consistently induce a change in perceived speaker sex.
- Accuracy was generally higher in swapped conditions for male speakers compared to female speakers, possibly related to differences in the distribution of F0 or FFs in male and female speakers.
- There were also differences in listening strategies for the swapped conditions, which interacted with speaker sex and differed somewhat as a function of listener sex.

References

[1] Hillenbrand, J. M. and Clark, M. J. (2009). The role of F0 and formant frequencies in distinguishing the voices of men and women. *Perception and Psychophysics* 71(5): 1150-1166

[2] Assmann, P.F., Nearey T.M. & Bharadwaj, S. (2008). "Analysis and classification of a vowel database," Canadian Acoustics 36, 148-149.

[3] Assmann P.F., Barreda S. and Nearey T.M. (2013). Modeling the perception of speaker age and sex in children's voices. Journal of the Acoustical Society of America, 134, 4237 (A).

Acknowledgments

Work supported by the National Science Foundation Grant #1124479. Thanks to Daniel Hubbard and Shaikat Hossain for assistance in data collection and analysis.