Perception of Voice Gender in Cochlear Implant Simulations of Children's Speech (2476831)

Daniel Guest, Michelle Kapolowicz, Shaikat Hossain, Vahid Montazeri, Peter Assmann
School of Behavioral and Brain Sciences, University of Texas at Dallas

Background

We have been investigating the perception of voice gender in children's speech by normal-hearing listeners, utilizing a database of recordings of children's speech from 208 speakers from the Texas area (Assmann et al., 2015). The present study extends this research by processing the stimuli with a cochlear implant (CI) simulation. CI simulations mimic the signal processing that occurs in a CI and allow for the use of normal-hearing listeners to answer questions about the effects of CIs on speech perception. This has numerous advantages – testing normal-hearing listeners is often easier and more affordable than testing CI users and avoids possible complicating factors frequently associated with CI users such as pre-lingual deafness/hearing loss. Therefore, while data collected from normal-hearing listeners attending to CI simulations cannot replace data collected from actual CI users, it can serve as an effective and efficient way of "piloting" possible studies of interest before extending them to CI users.

The present study seeks to use information from CI simulation results in two ways. First, it will be used to determine to what extent perception of voice gender in children's speech is possible in conditions of reduced spectrotemporal information (such as in a CI simulation or a CI). Second, it will be used to generate hypotheses about what is to be expected when the experimental design is extended to CI users – that is, how do listeners handle gender recognition in children's speech when some of the key cues in the task are distorted or unavailable?

Voice gender identification depends heavily on correct identification of the speaker's fundamental frequency (F0) and formant frequencies (FFs). These cues are important for understanding speech in quiet, and they also contribute significantly to our ability to enjoy music, determine the emotional content of speech, and distinguish a voice in difficult conditions. CI users often are less able to accurately identify a speaker's F0 and FFs, which can lead to deficits in many of these abilities. Results from the present study may offer a better understanding of how these cues may be perceived and utilized by CI users and therefore how to minimize these deficits in the future.

Objectives

1. Demonstrate effects of the CI simulation on gender and age perception
2. Highlight effects of the voice gender cue swaps (F0 and/or FF) at each age level

Methods

Stimuli — 420 syllables of children's speech
14 Age Levels (5 – 18 years) x 2 Genders (Male and Female) x 5 Speakers x 3 Vowels (/hid/, /hɑd/, and /hud/)

Voice gender cue swap
Implemented STRAIGHT vocoder (Kawahara, 2006) in MATLAB to swap fundamental and/or formant frequencies to opposite sex averages at each age level. Each subject heard one of **four conditions**:

Unswapped – neither F0 nor FFs altered
F0 Swapped – F0 swapped, FFs unaltered
FF Swapped – FFs swapped, F0 unaltered
F0FF Swapped – both F0 and FFs swapped

Experiment #1
Stimuli are NOT processed with a CI simulation
→ Stimuli were presented monaurally to normal hearing listeners via headphones with Tucker-Davis System 3 and RP2.1 hardware
→ Listeners were asked to identify the age and sex of the speaker of a given stimulus
→ **Unprocessed** (n = 80)

Experiment #2
Stimuli processed with a CI simulation
Implemented as tone vocoder according to Fu et al. (2005)
→ Stimuli were presented monaurally to normal hearing listeners via headphones with Tucker-Davis System 3 and RP2.1 hardware
→ Listeners were asked to identify the age and sex of the speaker of a given stimulus
→ **CISIM** (n = 37)

Results

For all graphs presented, unprocessed and CISIM results will be presented side-by-side, with unprocessed on the left and CISIM on the right.

Q: How does gender perception vary across age in each condition for MALE voices?

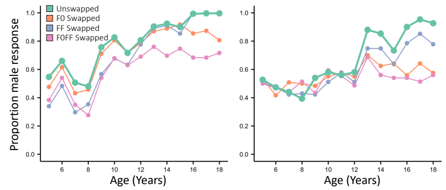

Figure 1: Proportion Male Response, (Male Voices)

Proportion male response at each age level, averaged across listeners and speakers. Color indicates voice gender cue condition. While swapping either F0 or FF has an effect at most age levels, swapping F0 and FF together generally has the largest effect.

Q: How does gender perception vary across age in each condition for FEMALE voices?

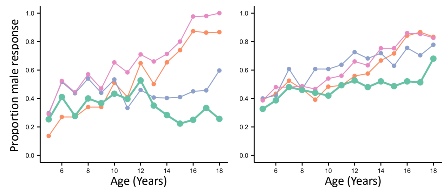

Figure 2: Proportion Male Response, (Female Voices)

Proportion male response at each age level, averaged across listeners and speakers. Color indicates voice gender cue condition. In the unprocessed data, voice gender cue swaps seem to have a larger effect on female voices than they do on male voices.

Q: How variable is gender perception in the unswapped condition?

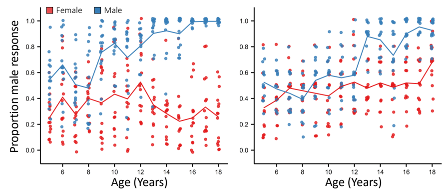

Figure 3: Proportion Male Response, Per Stimulus (Unswapped)

Proportion male response at each age level, averaged across listeners. Each point represents one stimulus. Color indicates speaker sex. Jitter is applied to the points for visibility, and superimposed lines represent the average across stimuli.

Q: How variable is age perception in the unswapped condition?

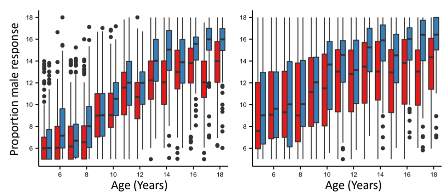

Figure 4: Age Perceived (Unswapped)

Boxplots for age perceived at each age level, averaged across listeners and speakers. Color indicates speaker sex. Note the decreased slope and greater variance of the boxplots in the CISIM, as well as the tendency in both unprocessed and CISIM data to underestimate the age of older female speakers.

Model Description

In order to quantify the trends in the above graphs, generalized linear models were fit separately to the Unprocessed and CISIM data according to the following formula:

$$\mathrm{P(male\ response)} = \mathrm{logit}^{-1}(\theta)$$
$$\theta = \alpha + \beta_1 S + \beta_2 A + \beta_3 C + \beta_4 AS + \beta_5 AC + \beta_6 SC + \beta_7 ASC$$

where S = sex of the speaker, A = age of the speaker, and C = voice gender cue condition. The results of the model, fitted according to the maximum likelihood method, are presented in the following section.

Model Predictions

Both models (1) provide an improvement over null models, (2) contain only factors of statistical significance according to likelihood ratio chi-squared tests, and (3) show good concordance statistics.

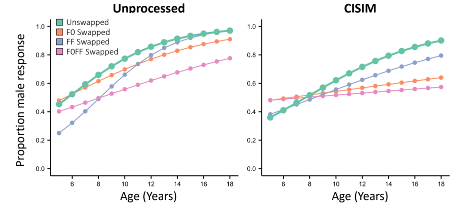

Figure 5: Proportion Male Response, Predicted (Male Voices)

Predicted proportion male response at each age level, averaged across listeners and speakers. Color indicates voice gender cue condition.

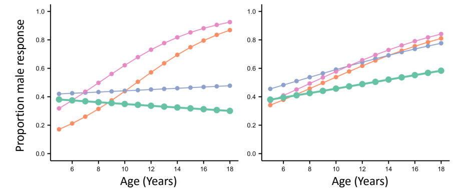

Figure 6: Proportion Male Response, Predicted (Female Voices)

Predicted proportion male response at each age level, averaged across listeners and speakers. Color indicates voice gender cue condition.

Discussion

- Effects of CI simulation on gender and age perception
  - The CI simulation resulted in lower overall performance in the unswapped condition
    - Pairwise comparisons from the model indicate that in the unprocessed data the odds of being heard as male were 10 times larger for male stimuli than for female stimuli, but only 2.5 times larger in the CISIM
    - Proportion male response across age for female stimuli was not significantly different from .50 in the CISIM
  - Age estimation error was approximately 50% larger on average in the CISIM
  - As seen in the slope of the regression lines (Figure 6), *all* older stimuli were more likely to be heard as male in the CISIM, while in the unprocessed data older female stimuli were more likely to be heard as female
- Effects of voice gender cue swaps
  - Pairwise comparisons on the model indicate that in nearly all contexts swapping F0 and FFs together had a larger effect than swapping F0 or FFs alone, as would be expected from Hillenbrand and Clark (2009)
    - For older voices, swapping F0 alone generally had a larger effect than swapping FFs alone
    - For younger voices, swapping FFs alone generally had a larger effect than swapping F0 alone
  - In the unprocessed data, swapping had a larger effect on older female stimuli than on older male stimuli
    - For example, in the unprocessed data the odds that swapping both F0 and FFs would result in a swap in gender perception for stimuli from 16-year-old speakers were about 2.4 times larger for female stimuli

Future Direction

1. Extend the experiment to include CI users
2. Develop a model based on acoustic features of speech signals (as well as listener variability)
3. Identify and determine the cause of common errors in identification of gender in children's speech

References

1. Assmann, P.F., Kapolowicz, M.R., Massey, D.A., Barreda, S., and Nearey, T.M. (2015). Links between the perception of speaker age and sex in children's voices. *The Journal of the Acoustical Society of America*, 138, 1811-1811. http://dx.doi.org/10.1121/1.4933751
2. Fu, Q., Chinchilla, S., Nogaki, G., & Galvin, J. (2005). Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *The Journal Of The Acoustical Society Of America*, *118*(3), 1711. http://dx.doi.org/10.1121/1.1985024
3. Hillenbrand, J.M., and Clark, M.J. (2009). The role of F0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, and Psychophysics*, 61, 1150-1116. http://dx.doi.org/10.3758/app.71.5.1150
4. Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. & Tech.*, *27*(6), 349-353. http://dx.doi.org/10.1250/ast.27.349