The Development of Gender-Typical Speech in Children: A Longitudinal Analysis

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Natasha Lackas

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Benjamin Munson, Ph.D.

June 2019

Acknowledgements

Special thanks to my advisor, Dr. Benjamin Munson for supporting and guiding me through this project, and for his support throughout my professional education from start to finish. Thanks to Dr. Stronach and Dr. Schlauch for their support throughout my time at the University of Minnesota, Twin Cities, for serving on my committee, and for being so flexible with their time.

Stimulus preparation, transcription, and data collection was a team effort with help from everyone at the Learning to Talk Lab. Thanks go out to Kiana Koeppe, Gisela Smith, and Katherine Dougherty.

I appreciate the opportunity to work alongside so many talented individuals.

Abstract

Sociophonetic variation has been observed between and within gender categories in adults. Some speech differences between men and women are anatomically based, but there is a growing body of evidence to support the idea that some of the gender differences in the speech of adults are socially and culturally learned and influenced. There is little research on this sociophonetic variation in children, including how it is learned. In this study, naive adult listeners rated the gender typicality of single word productions of children at two time-points, around the age of 2 ½ and 4 ½ years. The children completed standardized tests of vocabulary and articulation at the first time-point to analyze for predictors of increased gender typicality at the last time-point. Results confirmed that sociophonetic variation between boys and girls increased over time. This change was due to the boys being rated as more gender typical at the last time-point as compared to the first time-point. Standardized scores of vocabulary showed no correlation with change in gender typicality. Poorer scores on a test of articulation was correlated with being rated as more gender typical at the final time-point for boys only. These results contribute to our understanding of how and when children learn gender-marking in speech.

Table of Contents

List of Tables

List of Figures

1. Introduction

Our voice and speech style hold much information about us. Some of the variation that conveys information about us relates to our use of higher-level language, like word choice and syntax. For example, a person's choice of the word 'kitty' over the word 'cat' suggests a particular attitude toward objects of the class *felis catus*. Further, phonetic variation can convey a great deal about a speaker. For instance, segmental variation in the sounds /k/, /æ/, and /t/ in the word *cat*, as well as the overall pitch and pitch variation during the production of this word, can allow unfamiliar listeners to piece together information such as a speaker's age, class, gender, and regional affiliation. Much has been studied about the inter-speaker differences that are identified with these attributes in adults, but less research has been done to study how children learn and acquire these communicative styles and habits. This thesis is concerned with how this socially meaningful phonetic variation might develop, and, specifically, whether the differences that convey speaker gender are strictly anatomical and physiological in nature (i.e., related to sex dimorphism of the vocal tract), or whether they are learned behaviors throughout development.

1.1 'Sex' vs. 'gender'

Before reviewing the biological and acoustic differences that have been identified between men and women, it is important to define 'gender' more clearly. Munson and Babel (2019) review studies the distinction between 'sex' and 'gender' in the context of phonetic and sociolinguistic research, and this section summarizes the key points of their argument. In much of the current literature these terms are used somewhat interchangeably. The difference between these terms is sometimes summarized by saying that sex is biological, while gender is sociocultural. The biological aspects of sex might include chromosomal and hormonal composition, and external markers of sex, like genetalia and secondary sex characteristics. Gender is a social construct that refers to people's participation in activities and communities that have historically (including by our evolutionary ancestors) been completed by only one sex (i.e,. child-rearing for women, hunting for men). In a sense, then, gender refers to a set of practices that enforce social divisions between men and women. Gender expression is a complicated aspect of an individual's identity that often interact with social class, age, and other factors. While sex is generally seen as binary, gender has a more fluid presence in which individuals may still be perceived as feminine or masculine but more or less so based on complex interactions of physical and social cues. Individuals who identify as transgender in their adult lives must make conscious adjustments to their habits and social cues to be identified as the gender they identify with. Additionally, words like *femininity* and *masculinity* are "local instantiations of gender ideologies, and vary considerably across communities and through time, and are deeply imbricated with other hierarchies of power" (Eckert, 2014, p. 530). For this reason, we will be exploring gender as a continuum, allowing for the fact that individuals can make judgements that a person's speech is more or less gender typical according to stereotypes, despite their binary sex being clear.

When considering the results of the articles discussed here, it is essential to observe critically the intended definitions of these concepts by the authors. Most of these studies simply report the number of males and females in a study, without reporting how this information was gathered. Typically, this information is gathered by self-report or caregiver report, rather than by chromosomal analysis or inspection of primary or secondary sex characteristics. For our purposes, it is important to keep in mind the difference between the two terms, while accepting that "gender perception" is a term used to describe the assumed sex of an individual as it interacts with other sociophonetic factors. While our use of the term 'gender' throughout this paper does not do justice to the different definitions we have indicated above, it does allow us a term to account for both acoustically perceived biological sex, as well as capturing the subtle variations along a continuum in gender presentation. Please see Munson and Babel for a more in-depth discussion of these issues.

1.2 Variation in adult men and women

Phonetic differences between adult men and women's speech are pervasive. Indeed, the differences between men and women's speech are so robust that a person's gender can be discerned from very short samples of content-neutral speech, such as a short clip of a single vowel (Bachorowski & Owren, 1999). An article by Simpson (2009) describes phonetic differences common between adult males and females saying, "these are divided into two main types: biophysical inevitabilities and learned behaviors, more popularly referred to as nature and nurture" (p. 621). Many of the phonetic differences between men and women's speech are plausibly related to sex dimorphism of the vocal tract. For example, women have smaller, lighter vocal folds than men, which leads them to have higher fundamental frequencies (F0) (Titze, 1989). Based on research completed near Milwaukee, Wisconsin, an average F0 for men is between 100 and 146 Hz, while women typically have a F0 between 188 to 221 Hz (Gelfer & Mikos, 2005). These average ranges have been shown to vary, particularly across cultures and languages. This suggests that even average F0 of the voice is partially a learned behavior. In a study by Van Bezooijen (1995), she showed that individuals from different cultures have different values and expectations of what is important in the stereotypical concepts of men and women and how they should present themselves through pitch. It was shown that Japanese women valued higher pitched voices to present as more feminine, as compared to women from a Dutch culture. This paper confirmed all three of its theses; that higher pitches were commonly associated with smaller bodies, weakness, and dependence, that Japanese women placed more value in portraying these characteristics than Dutch women, and that there was a general preference shown toward higher pitches in women in the Japanese culture and lower pitches for women in the Dutch culture.

Also, commonly considered an important biological difference between men and women is the difference in the supralaryngeal vocal tract. Men have longer vocal tracts than women do, which leads them to produce continuant sounds with lower resonant frequencies than women. The average length of the adult female vocal tract, from the vocal folds to the lips, is 14-14.5 cm, and 17-18 cm for men (Simpson, 2009). This difference in vocal tract length has implications concerning the source-filter theory, where the sound energy produced at the vocal folds is shaped by the size and contour of the vocal tract and articulators. "Due to the resonant frequencies (formants) of a particular vocal tract configuration, certain frequency components (overtones, harmonics) of the tone produced at the glottis are strengthened" (Simpson, 2009, p.626).

Differences in vocal quality has also been noted between men and women, with men utilizing more creaky voice than women do, and women tending to have breathier vocal qualities. While this difference may appear sociophonetic, Simpson, 2009 states:

> Titze's (1989) computation model of vocal fold vibration predicts that male vocal folds will completely close the glottis, briefly shutting off the airflow during each cycle. In contrast, the vertically thinner female vocal folds in the model never make complete closure during each opening-closing cycle creating a constant airflow, that is, breathy voice. (p. 623).

Pitch range has also been examined as a possible difference between men and women. However, studies utilizing semitones instead of Hz, help avoid the non-linearity of sound perception with increasing F0. Semitones are used as a logarithmic measure instead, showing that men and women do not exhibit any significant differences in pitch range (Simpson, 2009).

It is of interest to note the findings of research with transgender individuals indicating how they develop a new set of behaviors, with a relatively fixed vocal anatomy, to align their gender perception with their gender identity. According to a meta-analysis of literature exploring the gender perception of transgendered individuals conducted by Leung, Oates, and Chang (2018) F0 was identified as the most salient factor contributing 41% of the variance in gender perception for transgendered adults. Other factors found to be statistically significant include resonance, loudness, articulation, and intonation. Although F0 may contribute the most compared to other factors, it makes up for less than half of the total, given all cues. "This interpretation of the results of the meta-analysis also supports current understanding that female voice is not simply a higher-pitched male voice" (Leung et al., 2018, p. 288).

Some of the phonetic differences between men and women are not plausibly related to sex dimorphism in the vocal tract. Stuart-Smith (2007) examined differences in /s/ between men and women of different ages (older and younger) and social classes (working-class and middle-class). Differences in /s/ between men and women were found for all age by social class combinations. However, the difference was smallest for the younger, working-class individuals. This appeared to be due to the younger, working-class women producing /s/ in way that lowered its peak frequency, thus minimizing how different it was from the /s/ of younger, working-class men, and making the /s/ conspicuously different from that of younger, middle-class women, and of older women. Stuart-Smith's finding illustrates the fact that some gender differences might be a learned, socially and culturally conventional ways of speaking.

Indeed, a growing consensus is that many of the phonetic differences between men and women represent a combination of sexual dimorphism of the speech-production mechanism, and learned phonetic habits. Many studies have tried to delimit the relative contributions of these two sets of influences on sex differences in a phonetic variable. For example, Fuchs and Toda (2010) attempted to determine the nature of sex-specific variation in /s/ by examining whether sex-specific differences in palatal shape could predict variation in /s/. They found a mixture of effects, in that though there was some evidence to support a biological explanation, even when the analysis controlled for the differences in palatal dimensions, persistent differences in the male-female articulation of /s/ were found.

1.3 Variation in children's speech and gender perception

Some of the differences discussed above emerge at puberty, when men's larynxes descend and their vocal folds thicken. A long-standing assumption has been that prior to puberty, there are no anatomical differences in the vocal tracts of boys and girls (Vorperian et al., 2005). Consequently, any differences between boys and girls' speech prior to puberty should represent learned behaviors. When comparing gender perception accuracy that analyzed both isolated vowels and connected speech of children, accuracy was higher for connected speech (Amir, Engel, Shabtai, & Amir, 2012). This indicates that factors other than F0 and formant frequencies contribute to gender perception in children.

The status of sex differences in vocal-tract anatomy prior to puberty is one of active debate. Early research has developed evidence that sexual dimorphism is not present until after puberty. One study by Barbier et. al (2015) found no evidence of pre-pubertal sexual dimorphism saying that sexual dimorphism is present in the pharyngeal cavity at age 12, and present in regards to vocal tract length at age 15. Another article by Vorperian et al. (2005), similarly found a lack of evidence of pre-pubertal sexual dimorphism until about 7 years of age.

Vorperian and colleagues (2011) reports some differences between boys' and girls' vocal tracts prior to puberty. However, these are not stable across the large age-range that they studied, but rather wax and wane in the pre-puberty years. Moreover, Barbier et al. (2015) were unable to replicate Vorperian et al.'s finding of gender differences in vocal tracts prior to puberty. Barbier et al. (2015) argued that the early sex differences found in Vorperian were an artifact of imbalanced sampling of sexes and ages in their study.

Children at a young age are not only trying to develop skills to articulate the sounds of their language, but also the sociophonetic properties of those sounds. It has been observed that children's speech is more variable than that of adults in regards to segmental durations. This variability can be attributed to "immature motor skills, and imperfect mapping of adult language" (Foulkes, & Docherty, 2006, p. 422). Children not only learn to produce these sounds but do so within an ever-changing vocal tract as they develop. This lends to the fact that variability in articulatory control is not the fault of developing motor speech skills alone, but also of developing anatomy (Whiteside & Hodgson, 2000). The magnitude and overall variability of segmental durations was shown to begin to decrease around age 9 and reflect adult productions at age 12 (Lee, Potamianos, & Narayanan, 1999). Boys and girls also demonstrate different patterns of articulatory development that become more pronounced around 6-7 years of age. Between-sex differences for voiceless fricatives were less pronounced for the youngest age group, suggesting both maturation of the vocal tract throughout time as well as learned development of articulatory habits (Fox & Nissen, 2005).

Adult listeners are able to identify a child's gender based off their speech signal alone. In a study by Amir et al. (2012), the authors sought to examine the ability of untrained adult listeners to accurately deduce the gender and age of 120 children aged 8 to 18 years. Results indicated an ability to identify the age in about 37% of cases, but an overall accuracy of 82% for identifying gender. Another study by Perry, Ohde, & Ashmead (2001), looked at the ability of adult listeners to identify gender correctly based off short carrier phrases in which only one vowel was altered from trial to trial. The children ranged from 4-16 years of age and gross physical measurements were taken to relate them to voice F0 and other acoustic measurements. This study found that adults were able to correctly identify speaker gender in children as young as 4 years old. This study also showed that F0 might not provide sufficient information in children younger than 12 years of age for correct gender perception. Another study showed that untrained listeners could identify the sex of 5- and 6-year-old children through recordings of their speech, again reporting that the average F0 of the children's voices was not shown to contribute to listener's performance (Weinberg & Bennett, 1971).

There is evidence of strictly sociophonetic differences in children as well. Munson, Crocker, Pierrehumber, Owen-Anderson, and Zucker (2015) reported that pre-pubescent children may display some early markers of gender identity in their speech and habits. This expression of gender identity suggests that children were learning to alter their speech style in ways that were not dependent upon anatomical structures. Children diagnosed with gender identity disorder often identify as transgender, homosexual, or bi-sexual as adults (Munson et al., 2015). In this study, participants (both clinical professionals and untrained raters) were asked to rate the children's speech as more boy-like or less boy-like as all children participants were male. It was shown that boys with gender identity disorder were identifiable through aspects of the children's speech and voice. A study conducted by (Li et al., 2016) noted that children are able to manipulate their speech styles and language choices to approximate "various sex/gender personas" (p. 60). Spectral properties of productions of /s/ by children 4-16 years old was analyzed, and gender identity information was collected through parent report. Properties of /s/ were shown to be different between sex groups, and variation in /s/ also explained some within-sex variation for the boys. This supports the notion of socially learned methods of /s/ productions.

1.4 The development of sociophonetic variation in children's speech

If children are learning gendered variations in speech, they must select whom it is they are trying to emulate. In an article by Foulkes & Docherty (2005), it was suggested that variation in the ways in which people speak is frequent and persistent. Children are likely to hear a variety of speech styles from different talkers as they grow, learn about the world around them, and develop their own ability to communicate. Furthermore, the quality and the variability to which a child is exposed may depend upon their sex and their age. As children age, their speech styles change in correlation with the changes in speech styles that are being directed to them (Foulkes, 2005). Ladegaard & Bleses (2003) explores two hypotheses of how children acquire gender specific speech styles. It describes the frequency hypothesis and the role-model hypothesis. The frequency hypothesis claims that children acquire speech variations from whoever they hear more frequently (e.g. their mothers or primary caretakers) and from the ways that adults change their speech style depending on if they are talking to a boy or a girl. The role-model hypothesis claims that children try to adopt the speech style variations of those they feel more comfortable with and more closely identify with. This study supports, that while both may come into play to a degree, the role-model hypothesis has more standing. This study also showed children were able to acquire gender specific preferences at an earlier age when exposed to a greater amount of socialization.

The quality of interaction between infants and their caregivers has been shown to influence their likeliness to imitate others. Parents are able to direct their children's attention to stimuli in their environment that are significant to them (Rvachew & Brosseau-Lapre, 2018), thus influencing their children to attend more to gender specific stimuli. As children develop their gender identities, they evaluate a group of people positively as soon as he or she identifies with that group. Experimental studies using novel toys demonstrated that children show more interest and remember more details about toys they believed to be intended for their gender group (Martin & Ruble, 2004). "Children also seem to want to generate or exaggerate male-female differences, even if none exist" (Martin & Ruble, 2004, p. 68.). In fact, there is a tendency for many species of animals to exaggerate sexual dimorphism, particularly of vocalizations, both anatomically and behaviorally (Diehl, Lindblom, Hoemeke, & Fahey, 1996).

Parents' reactions to their children's gender conformity or nonconformity varies individually and culturally. For instance, parents interviewed in New England about their reactions and feelings about gender nonconformity in their preschool children tended to welcome nonconformity among their daughters, but less among their sons. Particularly, heterosexual cis-gender fathers were the most likely to be motivated to endorse their concept of *masculinity* with their sons (Kane, 2006).

1.5 Clinical implications and research questions

The information reviewed concerning children's propensities to exaggerate gender differences, retain gender relevant information, and select whom they will emulate is clinically significant because if children are selectively learning from some adults, then speech/language learning in therapy might be slow if the clinician is the 'wrong adult'. It also provides support for the importance of promoting diversity in a field that has typically been dominated by a relatively homogenous group.

In this study, one of our research questions was to further investigate how children learned to sound more gender typical. We investigate if scores of standardized measurements of articulation and vocabulary correlate with growth in gender typicality of speech. These scores reflect complex skill development in children. Articulation scores were chosen because of their obvious connection to the gender-linked phonetic variation described above. Here, we investigate the hypothesis that children with more advanced articulatory control might not only score better on tests of articulation but would also be able to more closely emulate the speech of those in the group that they identify with.

Another skill that we investigate is vocabulary. Knowledge about phonological forms is accumulated from instances of hearing and saying a word. We learn from these experiences that the ways in which a word is said can vary to an extent and still be understood as that word. We also learn that we can use previous knowledge about phonetic variability and consistency to understand new words and interpret them as a novel concept or understand them as an already known concept. Phonological development is gradual and it interacts with other aspects of development (like vocabulary growth). This provides a framework for thinking about the gradual learning of a feature like how to articulate gender through speech (Beckman, Munson, & Edwards, 2004).

Having reviewed the research, a number of themes emerge. The first is that there is a paucity of longitudinal work on the development of gender typicality in speech. The second is that, other than Munson et al. (2015), there is no work looking systematically at individual differences in the attainment of gendered speech. This project fills in this gap by examining listener ratings of the gender typicality of single-word productions by 80 children (40 boys, 40 girls) at two developmental time-points, one when the children were 2;4-3;3 months old, and one when they were 4;3-5;5 years old.

Our research questions were as follows:

1. Are boys and girls as young as 2 ½ to 4 ½ years old rated differently on a scale of gender typicality as judged by untrained, unfamiliar listeners?
2. Is there significant growth in a child's ability to sound gender typical between the mean ages of 2 ½ and 4 ½ years old such that ratings of gender typicality are more defined at TP3?
3. Do standardized test scores taken at TP1 predict change in ratings of gender typicality from TP1 to TP3?

2. Methods

2.1 Participants

There were two groups of participants in this study, each of which is described separately. The Child Talkers refers to the 80 children whose productions served as stimuli in the perception study. The Adult Listeners refers to the 30 adults who rated the children's productions in the gender-perception task.

Child Talkers. Productions from 80 children (40 female, 40 male) were used as perception stimuli. These children had participated in a longitudinal study on relationships among speech perception, speech production, and word learning in a variety of groups of children, including typically developing children across a range of socioeconomic status, and children with hearing impairment who use cochlear implants. Children were recruited using a variety of methods, including flyers placed in schools, day-care centers, and other community organizations, advertisements in newspapers and websites, and word of mouth. Parents were asked to identify their children's gender, with the choices being male, female, or 'other.' One parent in the study chose not to report their child's gender, and reported that the child was being raised without an imposed gender identity, in hopes that the child's subsequent gender identity would be self-driven, rather than culturally imposed. That child is not included in this study. For the remaining children, we presume that the parent's report is of their child's assigned gender, and aligned with the child's developing gender identity.

The entire set of children in the larger study (N=164) participated in three testing sessions, one when they were 28-39 months old (time point 1 [TP1]), one when they were 40-52 months old (time point 2 [TP2]), and one when they were 53-66 months old (time point 3 [TP3]). At all three TPs, the children completed experimental measures of speech production, speech perception, and lexical processing, and a battery of standardized and non-standardized tests. A team of trained research assistants administered these measures.

The 80 children in this study were all normal hearing individuals and represented a range of socio-economic status (SES). The children were all native speakers of Mainstream American English (MAE, N=73) or African-American English (AAE, N=7). Dialect was determined through an interview with the children's parents, and by observing the parents interacting with children prior to the start of the first testing session at TP1. The 80 children were selected out of the larger copus of 164 children by each boy being matched to a girl within 3 months. They were also balanced for SES, but not for dialect. The only criteria for selecting the children from the larger sample was that they met the matching criteria. The N of 80 was chosen because it was feasible to complete the thesis project in the time-line we specified in advance. The stimuli used in this study were taken from children at TP1 and TP3. At TP1, children were 28-39 months old (Mean=32.4, SD=3.7). At TP3, children were 53-66 months old (Mean=56.9, SD=3.9). Independent-samples t-tests showed that the difference in age between boys and girls was not statistically significant at either TP1 ($t[76]<1$, $p=0.80$) or TP3 ($t[76]<1$, $p=0.90$).

Adult Listeners. Thirty adults served as listeners in the gender-perception task. They were between the ages of 18 and 40 years. All of the listeners self-identified as native English speakers (defined as having learned English from birth from at least one parent who spoke English natively and who communicated with them in English from birth) with no history of hearing impairment, or language/learning disorder.

The Adult Listeners were recruited through posters that were dispersed around the University of Minnesota, Twin Cities Campus, and from recruitment talks to classes. The Adult Listeners were compensated $10 for completing the task that was estimated to take 40 minutes to an hour. If the participant required longer than an hour to complete the task, they were compensated an additional $2.50 per 15 minutes.

2.2 Stimuli

The stimuli were eight word productions by each of the 80 Child Talkers, 4 each from TP1 and TP3 of the longitudinal study, as described above. The production data, in this study, were taken from a Real Word Repetition (RWR) task in the longitudinal study. The RWR task was designed to elicit a large number of productions of target sounds to assess growth in articulatory control in the preschool years. In the RWR task, children saw pictures of familiar words while hearing an auditory prompt naming that picture. Children were instructed to repeat the familiar word. Two sets of production prompts were used. One was produced by a white woman who was born in Minnesota and who spoke the local regional dialect, MAE. The other was produced in AAE by an African-American woman who was born in Wisconsin and who habitually code-switches between the local regional dialect spoken in her birth city, and African-American English. The use of different dialect prompts was motivated by goals of the original study. The production prompts were played in the home dialect, as assessed by the experiment coordinator. Of the 80 children in this study, 7 received the words in AAE and 73 in MAE.

Words used at TP1 were *cookie, scissors, shovel*, and *garbage*. Words used at TP3 include: *cookie, summer, shovel,* and *rabbit*. *Cookie* and *shovel* were chosen for both TPs. The other words were unique to either TP1 or TP3. Words were chosen because they do not have a strong association with girls or boys, like *doll* or *truck*. They also contain sounds such as /s/ that have been previously shown to cue listener judgments of gender typicality.

The Children Talkers were recorded saying each word a maximum of 2 times, depending on their cooperation during the testing session. One recording was chosen for each word at each TP, after listening to all possible recordings of a word at that TP for each child. The recordings were selected based on the researcher's judgment that the words were produced completely (i.e., no sounds omitted), clearly, and that they were representative of the child's normal mode of production (i.e., words that were produced in 'silly' voices were not included). Further, recordings where the child was speaking over the experimenter, or shouting the word were not selected. The recording was then cropped to contain only the child's production of the word, absent of the eliciting prompts.

Each stimulus word was transcribed and rated for articulatory accuracy. Two undergraduate students, who had both completed coursework in phonetics, were the transcribers. They were asked to listen to each sound in the word and rate it as accurate, inaccurate, or intermediate. The use of the 'intermediate' category is based on a recommendation by Stoel-Gammon (2001) as a solution for coding children's productions that sound mildly distorted but which are not clearly substitution errors. The use of intermediate productions is regularly used in the phonetic transcription protocol for the laboratory, including other transcriptions of these children's productions (e.g., Johnson, Cline, Beckman, Munson, & Edwards, 2015; Logerquist, Kim, Martell, Munson, & Edwards, 2018). After coding each phoneme, the full production was rated as either accurate or inaccurate based on the presence of an error. Productions were not rated as inaccurate if the only deviations involved the production of intermediate sounds. In addition, the transcriber judged the entire word as either:

- Correct (defined as 'likely to be perceived by a naive listener as an adult-like production of the target word),
- Intermediate (defined as 'likely to be perceived by a naive listener as the intended word, albeit with normal developmental errors'), or
- Incorrect (defined as 'likely to be misperceived by a naive listener as something other than the target word').

These transcriber's judgments were made using a custom-written Praat script. An example of the environment used to make these transcriptions is shown in Figure 1:

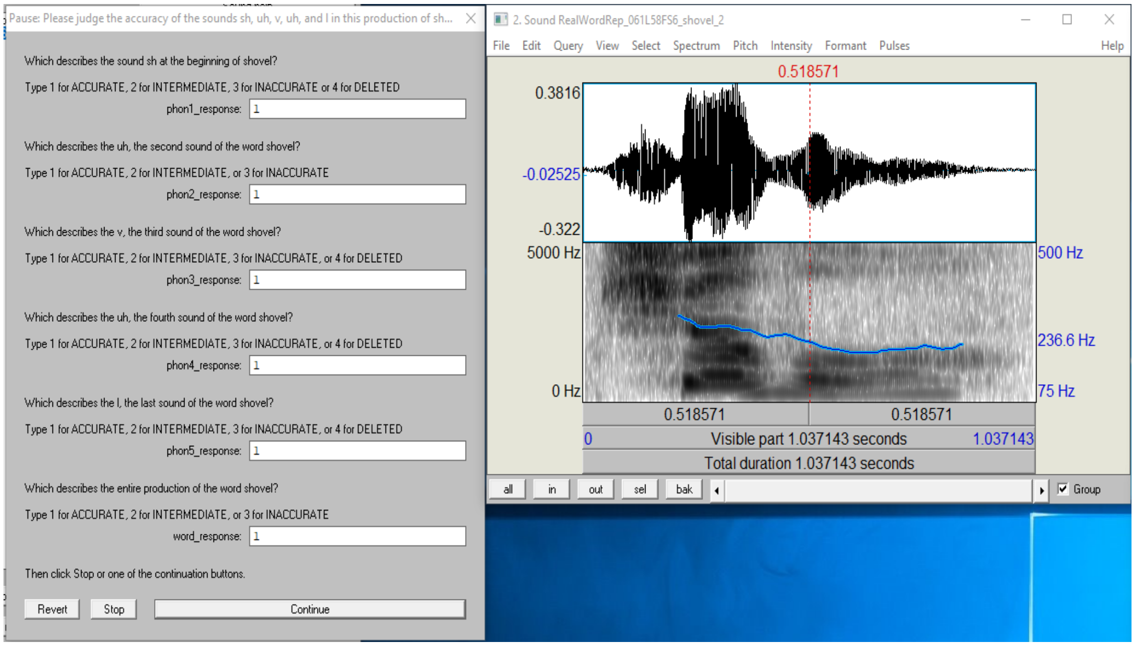

Figure 1. Transcription Window in Praat.

The mean number of phonemes correct per word was calculated, with fully accurate sounds coded as 1, deleted and substituted sounds coded as 0, and intermediate sounds coded as 0.5. The average phoneme accuracy scores at TP1 were 0.71 for girls (SD=0.22) and 0.72 for boys (SD=0.21), and at TP3 were 0.89 for girls (SD=0.14) and 0.90 for boys (SD=0.13). A linear mixed-effects model examined whether there was a significant effect of TP and talker gender on the phoneme accuracy scores. The fixed effects in this model were TP and gender, and the random effects were child and word. Models were built progressively, starting with a base model with only random intercepts for child and word. The next model added a fixed effect for TP (coded using contrast coding), and a random slope for the effect of TP on individual children (modeled as being uncorrelated with the random intercept for children). This model fit the data significantly better than the base mode, $\chi^2_{[df=3]}=110.96$, $p<0.001$, indicating that the effect of TP on accuracy was significant. A model including a fixed effect for gender and a random effect of gender on word did not improve model fit, either alone or in an interaction with TP. This indicates that there was a statistically significant difference in accuracy between boys and girls at either TP.

The whole-word accuracy judgments patterned similarly to the phoneme-accuracy judgements. The proportion of words judged to be adult-like at TP1 was 0.25 for girls and for boys, at TP3 was 0.64 for both girls and boys. The proportion of words judged to be intermediate (i.e., ‘likely to be perceived by a naive listener as the intended word, albeit with normal developmental errors’) at TP1 was 0.27 for girls and 0.32 for boys, and at TP3 was 0.19 for girls and 0.22 for boys. The proportion of words judged to be inaccurate at TP1 was 0.48 for girls and 0.42 for boys, and at TP3 was 0.17 for girls and 0.14 for boys.

In sum, the stimulus accuracy was very similar for boys and girls. Hence, any differences in ratings between boys and girls are unlikely to be due to differences in the accuracy of the stimuli, at least at a gross level.

2.3 Standardized Tests

Each Child Talker completed a series of standardized assessments at TP1, three of which are included in this thesis. The *Goldman Fristoe Test of Articulation-2* (GFTA-2, Goldman & Fristoe, 2000) is a test of articulation. In this test the child is shown colored picture drawings of common objects and asked to name the object. The word productions are then scored for accuracy and analyzed for error patterns. The *Expressive Vocabulary Test-2* (EVT-2, Williams, 2007) is a test of expressive vocabulary. The child is shown pictures and must either name them or provide a synonym for a name that is provided. The *Peabody Picture Vocabulary Test-4* (PPVT-4, Dunn & Dunn, 2007) is a test of receptive vocabulary. The child is shown a set of colored picture drawings and asked to point to the picture that represents a spoken word.

A series of t-tests examined whether the standardized test scores differed between boys and girls. The mean EVT score for boys was 117 (SD=11) and for girls was 117 (SD=14). The mean PPVT score for boys was 105 (SD=18) and for girls was 104 (SD=18). These differences were not statistically significant (t's<1, p's>0.05). The mean GFTA-2 standard score for girls was 89 (SD=13) and for boys was 96 (SD=11).

This difference was statistically significant in a t-test, $t[76]=2.5$, $p=0.015$. It is notable that this difference is not in the direction that has been found in normative studies of speech acquisition, where girls are often found to produce more-accurate speech than boys. We return to this point in the discussion.

2.4 Procedures

The Children Talkers were divided into 3 sets of 13 pairs. There was an additional 1 pair of children that was presented in each set, so each set had 14 pairs of children total. The three sets were balanced so their average age at TP1, test scores, and maternal education level were comparable. Difference in mean age varied between sets by no more than 2 months. Mean PPVT scores varied by no more than 4 points. Mean GFTA scores varied by no more than 2 points. Finally, mean maternal education level varied by no more than 1 unit, where units were granted in an incremental measure for each level of education achieved, ranked as follows: GED < high school diploma < trade school < community college < some college < college graduate < graduate/professional school graduate. We conducted a series of ANOVAs to ensure that the three stimulus sets did not differ in GFTA-2, PPVT-4, EVT-2, and chronological age at TP1. None of these ANOVAs found significant differences among the sets. Table 1 shows exact average for each set.

Set	Age (months)	Maternal Education level (1=low, 7=high)	Proportion of the set that spoke AAE	GFTA-2 standard score	PPVT-4 standard score	EVT standard score
Set 1	32.96	5.92	0.19	94.23	107.72	112.19
Set 2	33.00	5.58	0.03	92.35	113.26	116.80
Set 3	31.42	6.08	0.00	93.23	112.05	119.73

Table 1. Selected characteristics of the Child Talkers, separated by stimulus set

The experiment randomized whether people heard TP1 or TP3 first. Each participant rated gender typicality for one set of children, which included the control pair. They were exposed to 2 trial stimuli using words that were not used in the experiment to ensure they understood the task, and that they knew to use the entire scale when making judgements, rather than just selecting the extreme ends. A continuous rating scale was selected as they have been shown to result in more fine-tuned and less biased judgements of phonetic detail than categorical ratings scales (Munson, Schellinger, & Edwards, 2017).

2.5 Analysis

A series of different statistical procedures were used to analyze the data. A linear mixed-effects model in which gender ratings (i.e., the location on the visual analog scale where the listener clicked, transformed to be a proportion of the length of the line, from 0 to 1) were the dependent measure, and gender (coded using contrast coding, male=-1, female=1) and TP (coded using contrast coding, TP3=1, TP1=-1) and their interaction were predictors was used to determine whether boys and girls elicited statistically equivalent ratings at the two TPs, or larger at TP3 than TP1. There were random intercepts for child, word, and listener. Following the recommendations of Barr et. al (2013) a maximum random-effects structure was used, with the exception of the effect of gender on child, as every child had only one gender, and the effect of TP on word, as not all words were repeated across TPs. The full model was as below:

Rating~sex*time-point+(sex*time-point||word)+(sex*time-point||listener)

To analyze predictors of the growth in gender typicality from TP1 to TP3, we first calculated the mean rating that each child elicited, averaged across words and listeners. We then converted the scores for all of the children so that they represented the distance from the end of the line that corresponded to their gender. We refer to this measure as 'gender-marking accuracy'. This served as the dependent measure in a hierarchical multiple regression, in which the predictors were the standardized test scores listed earlier in this section.

3. Results

Our first research question was whether the boys and girls elicited different ratings of gender typicality, and whether the magnitude of these differences was the same at TP1 and TP3. To examine this, we used Linear Mixed-Effects Models (LMER, Baayen et al., 2008), implemented in R using the lme4 package (Bates et al., 2015). The LmerTest package was used to assess statistical significance of individual factors. The dependent measure in this LMER was gender rating, as described in the methods section, where 0 indicated a rating of 'definitely male' and 1 indicated 'definitely female'. The model had random intercepts for listeners and items (i.e., stimuli of specific words produced by specific talkers). Model building progressed iteratively. The base model included only random intercepts for listeners and items. The next, more-complex, model included a fixed effect for TP coded using contrast coding (TP3=1, TP1=-1), and a random slope that estimated the effect of TP on individual listeners. This model did not fit the data better than the base model, $\chi^2_{[df=2]}=2.362$, $p=0.307$. The next model included a fixed effect for talker gender, coded using contrast coding (female=1, male=-1), and a random slope that estimated the effect of gender on individual listeners. That model fit the data significantly better than the base model, $\chi^2_{[df=2]}=80.540$, $p<0.001$. Finally, we build a model that included an interaction between TP and talker gender, and a random slope that estimated the interaction between TP and talker gender for individual listeners. That model fit the data better than a model with gender only, $\chi^2_{[df=4]}=87.388$, $p<0.001$. The coefficients for fixed effects in the full model are shown in the table below:

Factor	Estimate	SEM	df	t-value	p-value
(Intercept)	0.533964	0.009515	72.8886	56.119	<0.001
Time Point (1=TP3, -1=TP1)	-0.010921	0.006441	613.9696	-1.696	0.091
Gender (1=F, -1=M)	0.051402	0.007495	142.4609	6.858	<0.001
Time Point x Gender	0.016461	0.006441	613.9698	2.556	0.011

Table 2. Result of the statistical tests predicting gender ratings by TP, gender, and their interaction

The interaction between gender and TP is shown in Figure 2. The circles represent the average ratings for individual talkers, averaged across listeners and words. As Figure 2 shows, most of the ratings were around the midpoint of the visual-analog scale. That is, none of the ratings suggested that the listeners perceived the children to be vividly male or female. The cluster of the ratings around the scale's mid-point might show the usefulness of using continuous ratings instead of binary sex judgements, in that more fine-tune comparisons were able to be made. Separate models were run on subsets of the data to examine the nature of the interaction between TP and gender. The first two of these models examined the effect of gender on ratings at TP1 and TP3 separately. The effect of gender on rating was significant for both TPs. However, the size of the coefficient for gender was nearly twice as high for the TP3 model as for the TP1 model (0.067 vs. 0.035). The third and fourth post-hoc models examined the effect of TP separately for boys and girls. The effect of TP on girls' ratings was not significant. In contrast, the effect of TP on the ratings of boys was significant. In sum, the results showed that the speech of boys and girls as young as 28 months of age elicited ratings of gender typicality that were different from one another, albeit only subtly different. This difference was larger at TP3, and the increase appears to be due entirely to the older boys being rated as more boy-like.

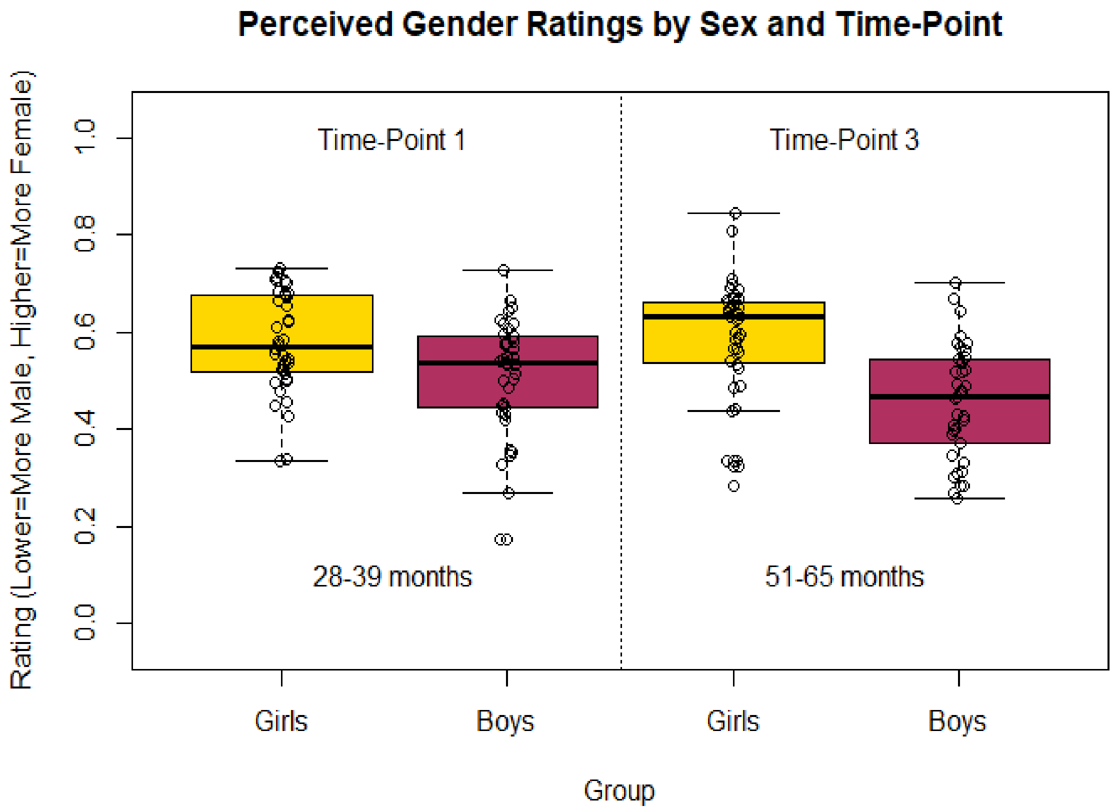

Figure 2. The effect of time-point on gender perception ratings.

To visualize these data differently, Figure 3 re-scales the scores for boys and girls in reference to the ‘boy’ and ‘girl’ ends of the scale, respectively, so that the ratings reflect the accuracy with which children’s gender was identified. In Figure 3, the values for individual children’s TP1 and TP3 ratings are shown, as is the y=x line. Symbols that fall above the line represent individuals whose TP3 ratings were more accurate (i.e., close to the end of the line that corresponds to their gender) than those at TP1. Those below the line are cases where the TP3 ratings are less accurate than TP1 ratings. In general, the figure shows that the ratings at the two TPs are well correlated, indicating that no child’s ratings changed substantially from TP1 to TP3. This figure shows that the majority of the boys’ ratings fell above the y=x line, indicating that they were identified more accurately at TP3 than at TP1.

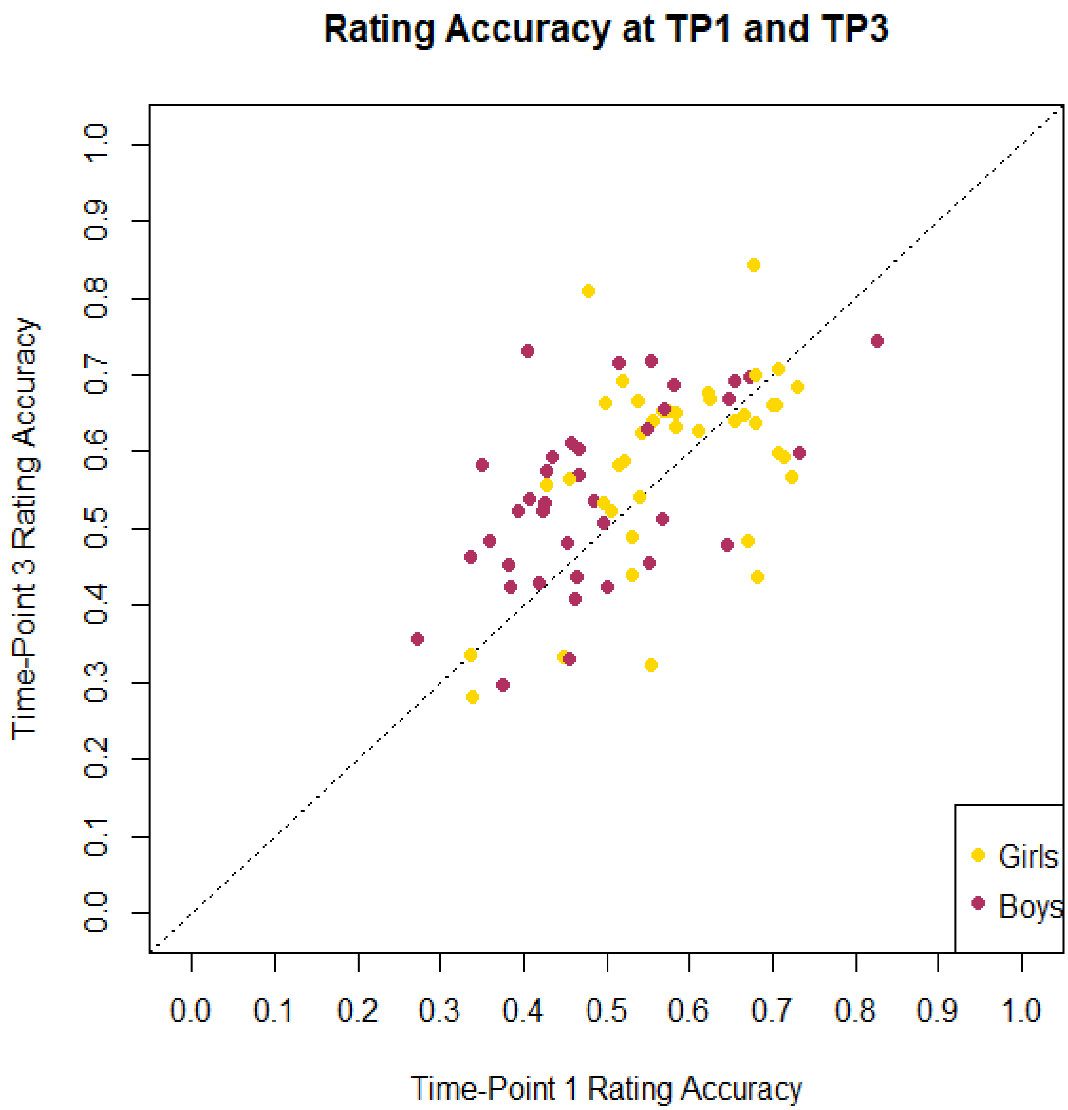

Figure 3. Rating accuracy as a factor of TP, showing that boys were rated more accurately at TP3.

The final analysis examined predictors of the accuracy with which TP3 gender was rated. To examine this, the average rating accuracy for each child at TP3 (as shown in Figure 3) was used as the dependent measure. The dependent measures were the PPVT, EVT, and GFTA scores from TP1. In that regression, the only significant predictor was the GFTA score. The β- coefficient in that regression was -0.0031223, indicating that higher GFTA scores at TP1 were associated with less-accurate gender identification at TP3. This relationship is shown in Figure 4.

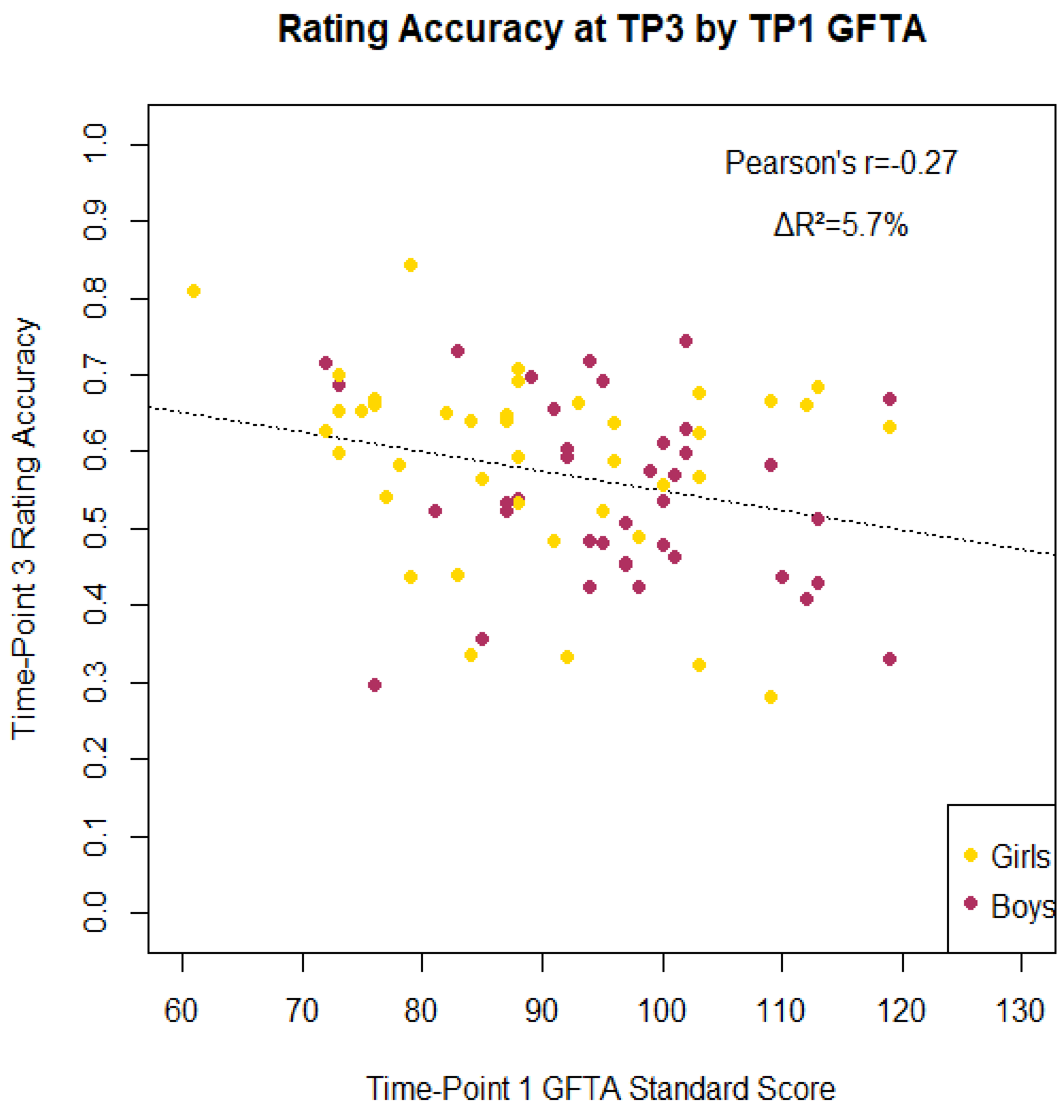

Figure 4. Rating Accuracy at TP3 by TP1 GFTA score.

To examine whether this relationship was robust, a second regression was conducted, in which TP1 accuracy and TP3 GFTA scores were also entered into the regression. In this regression, the effect of TP1 GFTA was significant even when these other variables were included in the equation. Comparing the variance accounted for in regressions with and without the TP1 GFTA score included showed that GFTA accounted for 5.7% of variance in TP3 accuracy scores when TP1 accuracy and TP3 GFTA were accounted for.

4. Discussion and Conclusions

To summarize, our study reaffirmed what has previously been discovered that boys and girls as young as 4 ½ years old, and even as young as 2 ½ years old, are rated differently on a scale of gender typicality as judged by untrained, unfamiliar listeners. We also found greater variation between boys and girls at TP3 than at TP1, which confirms our second research question regarding growth in the difference between boys and girls with time. However, the growth in the difference was due to the boys being rated as sounding more boy-like. Our third research questions investigated predictions of change in ratings of gender typicality from TP1 to TP3 based on standardized test scores of vocabulary and/or articulation from TP1. Our findings suggest that low GFTA scores at TP1 predicted more defined ratings of gender-typical speech in boys at TP3. This does not align with the hypothesis that better scores of articulation would correlate with more skill in emulating preferred gender identities.

One possible explanation for this is that the boys whose articulation scores were low, had low motivation to improve them, as their speech aligned more closely with that of adult men. It has been documented that women tend to speak more clearly than men, rather the "phonetic correlates of phonological contrasts produced by women tend to be more distinct in acoustic and temporal space than those produced by men" (Heffernan, 2010, p. 67). In this well named article *Mumbling is Macho: Phonetic Distinctiveness in the Speech of American Radio DJs,* Heffernan (2010) utilizes Likert scales to measure listeners rating of eight male American DJs of different genres of music to observe with-sex variation. All phonetic variables being measured were correlated with masculinity with at least marginal significance. This led to the conclusion that "phonetic distinctiveness, that is, clarity of speech, indexes socially constructed gender categories" (Heffernan, 2010, p. 83).

As with much of the prior discussion of gender typical speech patterns, it seems that the tendency for men to speak less clearly may be due both to sociophonetic and physiological/anatomical factors. In a study by Weirich, Fuchs, Simpson, Winkler, & Perrier (2016), they investigated anatomical explanations of male-female differences in clarity of speech. Greater jaw opening was observed in female speakers compared to male speakers. Models analyzed in this study showed that, anatomically, men were likely to demonstrate smaller jaw openings in an effort to avoid total closure of the vocal tract resulting from a difference in size of pharyngeal anatomy and a difference in distance between these parts. They suggested that "In order to avoid the tongue getting too close to the rear wall of the pharynx and thereby narrowing the pharyngeal cavity, a male speaker may not open his jaw as widely" (p. 1588).

To examine this interaction further, an experiment is currently being completed with a fourth set of stimuli for an additional 15 pairs of boys and girls. These children were selected to have as wide a range of GFTA scores as we could find while still meeting our matching criteria. Indeed, the inclusion of these additional 30 children exhausts the full set of potentially usable subjects from the larger set of 164 who participated in the learning to talk project. A second purpose of this fourth set is to re-examine and compare the sample of children in the fourth set to the sample of children in this thesis, as the strength of the GFTA scores for our male children as compared to our female children was unexpected and may have impacted the results.

Our findings also suggest that growth in gender typicality in girls does not change substantially within the age ranges explored in this thesis. It has been noted before that growth in gender typicality in speech is more dramatic for younger boys than it is for girls (Amir et al., 2012). This could be due to the fact that the smaller, less developed vocal tracts of young children more closely emulate the vocal tracts of adult women, and as such, girls do not have to compensate as much in their adolescents to approximate adult female vocalizations.

It could also be that since women tend to have more standard features in their speech and language than boys, more effort is being made on the part of the boys to learn vernacular variants. Indeed, not only do boys learn to sound gender-typical from the men in their lives, it has also been shown that mothers of young boys are more likely to use vernacular variants than mothers of young girls (Foulkes et al., 2006). In a study intended to confuse the speaker's gender in the speech signal of adults by altering formant frequencies and F0, identification rates were better for men than for women. They concluded that gender-specific cues were more salient for men then they were for women (Gelfer & Mikos, 2005). Should this be the case, it would be natural for boys to learn male gender-typical cues at a faster rate since phonetic variants that are more accessible and clearly related to a specific social category should be easier to learn than those that are more ambiguously linked, or less salient in the speech signal (Foulkes et al., 2006).

Finally, while some within-sex phonetic differences have been observed in women, less research has been conducted on within-sex variation in younger girls. Li et al. (2016) observed within-sex variation of /s/ in young boys that was associated to variation in gender identity, though no significant within-sex variation was observed for girls. It is yet to be seen if the lack of research detailing within-sex variation of young girls is because there is indeed more variation being observed in boys or if it is due to other factors. At this time, it seems unlikely that there is a lack of within-sex variation in females when compared to males, rather that it has yet to be documented.

Future research should be completed to identify phonetic factors cueing the perceptual ratings of adult listeners for children. Also, research should be conducted to detail more specifically the longitudinal growth of gender perception in children. Specifically investigating why boys show more growth, when compared to girls, in gender typicality over time. This information is clinically relevant in that it develops further our understanding of language and speech acquisition in children. Understanding how children develop sociophonetic variation could guide future decisions of how to appropriately match individual children seeking speech-language therapy with an appropriate clinician.

References

Amir, O., Engel, M., Shabtai, E., & Amir, N. (2012). Identification of Children's Gender and Age by Listeners. *Journal of Voice, 26(3),* 313-321.

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59 (4)*, 390–412.

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*, 1054-1063.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, 68, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-9, https://CRAN.R-project.org/package=lme4.

Guillaume Barbier, Louis-Jean Boë, Guillaume Captier, Rafael Laboissière. Human vocal tract growth: A longitudinal study of the development of various anatomical structures. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), International Speech Communication Association, Sep 2015, Dresden, Germany. hal-01200990

Beckman, M. E., Munson, B., & Edwards, J. (2004). Vocabulary Growth and the Developmental Expansion of Types of Phonological Knowledge. In *Laboratory Phonology 9.* Mouton de Gruyter.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics, 24*, 187-208.

Eckert, P. (2014). The Problem with Binaries: Coding for Gender and Sexuality. *Language and Linguistics Compass 8(11)*, 529-535.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics 34*, 409-438.

Foulkes, P., Docherty, G., & Watt, D. (2005). Phonological Variation in Child-Directed Speech. *Language, 81(1)*, 177-206.

Fox, R. A., & Nissen, S. L. (2005). Sex-Related Acoustic Changes in Voiceless English Fricatives. *Journal of Speech, Language, and Hearing Research*, *48,* 753-765.

Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In *Turbulent sounds: An interdisciplinary guide* (pp. 281-302). Mouton de Gruyter.

Gelfer, M. P., & Mikos, V. A. (2004). The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels. *Journal of Voice, 19(4)*, 544-554.

Heffernan, K. (2010). Mumbling is Macho: Phonetic Distinctiveness in the Speech of American Radio DJs. *American Speech, 85(1)*, 67-09.

Johnson, A., Cline, S., Beckman, M.E., Munson, B., & Edwards, J. (2015). Child-Level Factors & Acquisition of the /t/-/k/ Contrast: Production. Poster presentation at the American Speech-Language-Hearing Association, Denver, CO, November 12.

Kane, E. W. (2006). "No Way My Boys Are Going to be like That!" Parents' Responses to Children's Gender Nonconformity. *Gender and Society, 20(2)*, 149-176.

Ladegaard, H. J., & Bleses, D. (2003). Gender differences in children's speech: the acquisition of sociolinguistic competence. *International Journal of Applied Linguistics, 13(2)*, 222-233.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*, 1455-1468.

Leung, y., Oates, J., & Pang Chan, S. (2018). Voice, Articulation, and Prosody Contribute to Listener Perceptions of Speaker Gender: A Systematic Review and Meta-Analysis. *Journal of Speech Language and Hearing Research, 61*, 266-297.

Li, F., Rendall, D., Vasey, P. L., Kinsman, M., Ward-Sutherland, A., & Diano, G. (2016). The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. *Journal of Phonetics*, *57*, 59-70.

Logerquist, M., Martell, A., Kim, H., Munson, B., & Edwards, J. (2018). Growth in the accuracy of preschool children's /r/ production: Evidence from a Longitudinal Study. Poster presentation at the Acoustical Society of America, Minneapolis, MN.

Martin, C. L., & Ruble, D. (2004). Children's Search for Gender Cues: Cognitive Perspectives on Gender Development. *Current Directions in Psychological Science, 13(2)*, 67-70.

Munson, B., & Babel, M. (2019). The phonetics of sex and gender. In W.F. Katz and P.F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 499-525). New York: Routledge.

Munson, B., Crocker, L., Pierrehumbert, J. B., Owen-Anderson, A., & Zucker, K. J. (2015). Gender typicality in children's speech: A comparison of boys with and without gender identity disorder. *The Journal of the Acoustical Society of America, 137*, 1995-2003.

Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics & Phonetics, 31(1)*, 56-79.

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America, 109*, 2988-2998.

Rvachew, S., & Brosseau-Lapre, F. (2018). *Developmental Phonological Disorders: Foundations of Clinical Practice 2nd Edition.* San Diego: Plural Publishing, Inc.

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

Stoel-Gammon C. (2001). Transcribing the speech of young children. *Topics in Language Disorders, 21,* 12–21

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In *Laboratory Phonology 9* (pp. 65-86). New York: Mouton de Gruyter.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85,* 1699-2007.

Van Bezooijen, R. (1995). Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women. *Language and Speech, 38(3)*, 253-265.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America, 117,* 338-350.

Vorperian, H. K., Wang, S., Schimek, E. M., Durtschi, R. B., Kent, R. D., Gentry, L. R., & Chung, M. K. (2011). Developmental Sexual Dimorphism of the Oral and Pharyngeal Portions of the Vocal Tract: An Imagining Study. *Journal of Speech, Language, and Hearing Research, 54*, 995-1010.

Weinberg, B., & Bennett, S. (1971). Speaker Sex Recognition of 5- and 6-Year-Old Children's Voices. *The Journal of the Acoustical Society of America, 50*, 1210-1213.

Weirich, M., Fuchs, S., Simpson, A., Winkler, R., & Perrier, P. (2016). Mumbling: Macho or Morphology? *Journal of Speech, Language, and Hearing Research, 59,* 1587-1595.

Whiteside, S., & Hodgson, C. (2000). Speech patterns of children and adults elicited via a picture-naming task: An acoustic study. *Speech Communication, 32,* 267-285.