ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

The Acoustic and Perceptual Correlates of Gender in Children's Voices

Riccarda Funk[a] **and Adrian P. Simpson**[a]

[a] Institute for German Linguistics, Friedrich-Schiller University, Jena, Germany

ARTICLE INFO

Article History:
Received November 30, 2022
Revision received March 9, 2023
Accepted May 23, 2023

Editor-in-Chief: Cara E. Stepp
Editor: Leah B. Helou

https://doi.org/10.1044/2023_JSLHR-22-00682

ABSTRACT

Purpose: This study investigates the perceptual and acoustic correlates of gender in prepubertal voices. The study is part of a longitudinal project analyzing recordings of circa 60 German primary school children from the first to fourth grades (6- to 10-year-olds).
Method: Spontaneous and content-controlled audio recordings were made of 62 first-grade children (29 girls, 33 boys; age: 6- to 7-year-olds) from two German primary schools. Information on gender conformity was also recorded. A total of 167 listeners judged the gender of the voices on a 7-point scale. The results of the listening experiments and gender conformity ratings were related to a range of typical acoustic parameters.
Results: Measures of self-reported gender conformity differ significantly between the boys and the girls. Sixteen of the 62 children show unambiguous gender attributions in the listening experiment. A hierarchical cluster analysis including gender perception, gender conformity, and acoustic parameters shows four different types of speakers. Two multiple regression models revealed a significant main effect of fundamental frequency on the gender perception ratings of the listening experiment across and within gender. Significant correlations were found between the center of gravity and skewness of the sibilants and gender conformity, especially for the male speakers.
Conclusions: Fundamental frequency plays an important role in influencing perceptual judgments, whereas sibilant spectra are correlated with gender conformity. In further listening experiments, we will examine in more detail the role of individual acoustic parameters and analyze how the vocal expression of gender and gender conformity in individual children develops before reaching puberty.

The expression of gender in the prepubertal voice represents one of the most complex areas of phonetic research into the vocal expression of gender in the human voice (Simpson & Weirich, 2020). Certain phonetic correlates of gender in the adult voice are attributable to biological differences resulting from developmental changes in the larynx and the vocal tract that happen during puberty (e.g., Titze, 1989). Most notably, the average adult female fundamental frequency (f_o) is higher due to shorter and thinner vocal folds, and the average male resonance frequencies are lower due to a disproportionate lowering of the larynx and resultant lengthening of the vocal tract (Simpson, 2009). Although behavioral differences play a significant role in shaping the vocal expression of gender in the adult voice, these anatomical and physiological factors play a major role in determining the phonetic differences that often lead to correct gender identification of voiced stimuli at a rate close to 100% (e.g., Whiteside, 1998).

Acoustic Differences of Prepubertal Children

By contrast to adult findings, anatomical and physiological differences in the larynx and the vocal tract of prepubertal boys and girls are negligible (Fitch & Giedd, 1999; Kahane, 1978). Despite this, studies have repeatedly found gender-specific acoustic differences in prepubertal voices, from children as young as 2.5 years of age (McCormack & Knighton, 1996). They suggest that, from an early age, children are exhibiting learned gender-

Correspondence to Riccarda Funk: riccarda.funk@uni-jena.de. ***Disclosure:*** *The authors have declared that no competing financial or non-financial interests existed at the time of publication.*

specific patterns. However, the differences that have been observed do differ from study to study, and differences found in one study have not always been replicated.

f_o was found to be significantly lower in boys from the age of 7 years onwards in Hasek et al. (1980) and Ferrand and Bloom (1996) in American English. Glaze et al. (1988) found an overall gender difference (girls > boys) in a group of 5- to 11-year-old Americans, but no age-related differences. By contrast, Whiteside and Hodgson (1999) did find age-related, but no gender-specific, differences in a group of 6- to 10-year-old English-speaking children.

Differences in voice quality have also been found. Boys have been found to have a higher noise component, with correspondingly lower harmonics-to-noise-ratio (HNR) values and a stronger auditory impression of hoarseness in different languages (Ferrand, 2000; Kallvik et al., 2015; Sederholm, 1995; Simpson et al., 2017; Yairi et al., 1974).

One consistent difference that has repeatedly been found in different languages is acoustically more open female vowel qualities reflected in higher female F1 values (Auszmann & Neuberger, 2014; Bennett, 1981; Busby & Plant, 1995; Lee et al., 1999; Perry et al., 2001; Pettinato et al., 2016; Whiteside, 2001; Whiteside & Hodgson, 2000). Furthermore, Cartei et al. (2019) found larger acoustic vocal spaces in 6- to 10-year-old British girls.

A further aspect concerning differences in resonance is the spectral structure of /s/ in English-speaking children: Boys have repeatedly been found to produce this sibilant with energy concentrated at lower frequencies, reflected in a lower male center of gravity (CoG) and higher (and positive) skewness values (Flipsen et al., 1999; Li et al., 2016; Shadle et al., 2011).

Finally, but again less conclusive, are findings for tempo differences. Some studies have found English-speaking girls to have a lower speech tempo than boys (Walker et al., 1992; Whiteside & Hodgson, 2000); others have found the reverse to be the case, with German-speaking boys having the lower tempo (Funk et al., 2018; Simpson et al., 2017).

Gender Perception of Prepubertal Voices

It is possible that the acoustic differences described above are the reason why adult subjects generally perform better than chance in correctly identifying a child's gender in listening experiments (Günzburger et al., 1987; Karlsson, 1987; Kaya et al., 2017) or why gender perception ratings for boys and girls are significantly different (Munson et al., 2022). Overall, average correct gender identification rates are at approximately 70%, but this figure in itself is not very informative and belies more interesting patterns once we begin to examine identification rates for subgroups of speakers. Although identification rates for a subset of children remain at chance-level, even when sentence-length stimuli are used, the speech samples of other children lead listeners to make systematic decisions for one gender, often at levels approaching those found for adult voice stimuli (e.g., Simpson et al., 2017). This, in turn, indicates that this reliably identified subset of children is producing robust acoustic cues, which adults use for decoding gender.

The ability of 6- to 10-year-olds to consciously modify their voices to imitate girls and boys further suggests that children are actively producing at least some of these gender-attributable acoustic cues (Cartei et al., 2014, 2019). Further support for this assumption is provided by work on the vocal expression of gender identity and gender roles. Such identification has been shown in children as young as 2–3 years old (Farr et al., 2018; Martin et al., 1995; Rohrmann, 2017). Studies have shown that gender identity and gender role self-concept is reflected in speech of adults (Weirich & Simpson, 2018); therefore, it is probable that this also plays a role in children's voices. So, for instance, Munson et al. (2015) compared the speech of a group of 5- to 13-year-old boys with GD (gender dysphoria) with that of a group of age-matched boys without and found that the GD group producing both word- and sentence-length utterances were rated as less male. This could be related to a range of acoustic parameters: higher f_o, higher formant frequencies, longer durations, and more diffuse spectrum of /s/.

Aims and Research Questions

Continuing the work begun in a pilot study (Simpson et al., 2017), this study investigates the acoustic and auditory patterns that lead to robust or more ambivalent ratings of gender in children's voices. Although hoping to replicate the findings of previous studies with overall identification rates at about 70% (see Gender Perception of Prepubertal Voices section), we place our main focus on the acoustic correlates being produced in voices that are systematically perceived by listeners to be female or male.

Because gender role self-concept is reflected in the speech of adults (Weirich & Simpson, 2018), it is possible that it also plays a role in the prepubertal voice. We want to find out if and how acoustic and perceptual correlates of gender are related to a child's gender role self-concept. This, in turn, could be helpful for voice therapy of children with gender dysphoria, who want to sound more female or male.

This study reports the first results from a longitudinal study involving recordings of approximately 60 German primary school children, as they progress from the first to the fourth grade. Here, we examine acoustic and perceptual aspects of the first-grade recordings and how they relate both to adults' listening judgments and the children's gender role self-concept.

Specific research questions are as follows:

1. How do acoustic patterns of voices that lead to a robust rating of a child's gender (regardless of birth-assigned sex) differ from those that result in more variable gender responses?
2. Which acoustic parameters can be used to predict gender perception in children's voices?
3. How are acoustic parameters and gender perception connected with a child's gender role self-concept? Are children with strong self-described gender conformity (GC) perceived as either female or male? What are the correlations between GC and acoustic characteristics?

Even though we are interested in overall gender perception, we will also look at the differences between boys and girls. It is probable that birth-assigned sex has an influence on gender perception and gender role self-concept; therefore, it is important to distinguish the results within and between these groups.

In further studies of this longitudinal research, our aim is to analyze how the vocal expression of gender and gender role self-concept in individual children develops over the first years in primary education before reaching puberty.

Method

This study was carried out following approval by the ethics committee of the Friedrich Schiller University Jena. Permission was also granted from the responsible school authority, and children were only recorded following parental consent.

Speakers

Acoustic voice recordings were made of 62 children (29 girls and 33 boys) aged 6- to 7-year-olds (M = 6.16 years). The recordings took place in September 2020 at two primary schools in neighboring East German villages near Gotha, Thuringia. All the children were native speakers of German. There were 31 children from each school. Each child could participate, provided there was no developmental delay and the parents signed a declaration of consent.

Each child was assigned an index encoding the grade, school, and gender as well as a unique integer. So, for instance, a first-grade boy and girl at the first primary school were assigned the indices 1a 01 m and 1a 02w, respectively.

Acoustic Recordings

The children were recorded individually in a quiet room at school (school library in the first and music room in the second school). Background noises could not be excluded all the time, but they rarely occur since recordings were stopped during the lesson breaks.

Both spontaneous and read (copied) speech was elicited. The words in the picture-naming task and sentence material allowed for a careful and controlled elicitation of important acoustic aspects including vowel quality and vowel space size or the spectral structure of sibilants. The spontaneous samples, as well as the sentence material, allow for an analysis of typical phonetic aspects, such as tempo, intonation patterns, mean f_o, and voice quality. The following material was used for elicitation:

1. Disyllabic target nouns from picture-naming task: Hase /haːzə/ ("hare"), Blume /bluːmə/ ("flower"), Biene /biːnə/ ("bee"), Tasche /taʃə/ ("bag"), Nase /naːzə/ ("nose"), Kuchen /kuːxən/ ("cake"), Igel / iːgəl/ ("hedgehog"), Tasse /tasə/ ("cup"), Vase / vaːzə/ ("vase"), Tiger /tiːgɐ/ ("tiger"), and Lupe / luːpə/ ("magnifying glass"; see Figure A1 in the Appendix).
2. Spontaneous, unscripted description of three, more complex, pictures of a farmyard, a living room decorated for Christmas, and a playground (see Figure A2 in the Appendix).
3. Repetition of 10 simple prose sentences (see Table 1).

In the picture-naming task, the children were just asked: "What do you see in the picture?" For the spontaneous descriptions, the children were asked to describe the pictures with the questions: "What happens here? What can you see in this picture?" The sentence list to be repeated by each child was prerecorded by an adult female speaker (first author). We recognize that children copying prerecorded utterances will not only be reproducing the verbal content but may also accommodate to aspects of the adult speaker's voice. However, given the level of (il) literacy in first-graders, we saw no other alternative to elicit tightly controlled sentence-length utterances from children of this age and literacy level.

All recordings were made using a USB microphone with integrated audio interface (Røde NT-USB) located in a portable sound wall (Marantz Pro Sound Shield) to improve sound quality in school surroundings. We did not use head-mounted microphones not only to ensure a minimally intrusive recording environment but also to reduce

Table 1. Sentences to be repeated.

Number	Sentence	Translation
1	Im Sommer blühen die Blumen.	Flowers blossom in summer.
2	Kannst du die Rose riechen?	Can you smell the rose?
3	Die Kinder haben schöne Namen.	The children have nice names.
4	Er hat die Tasse ins Wasser getaucht.	He dipped the cup into the water.
5	Sie haben den Kuchen gebacken.	They baked the cake.
6	Das Backen hat lange gedauert.	The baking took a long time.
7	Zwei Mädchen haben den Kuchen gegessen.	Two girls ate the cake.
8	Du hast die Tasche gepackt.	You packed the bag.
9	Das Packen hat lange gedauert.	Packing took a long time.
10	Die Hunde haben lange gekämpft.	The dogs fought for a long time.

unwanted noise caused by head movement. The microphone was positioned about 30 cm from the speaker's mouth. Each child was asked to sit straight in front of the microphone. The sitting position was corrected if necessary. Recordings were made using Audacity (Audacity Team, 2021) with 32-bit amplitude resolution and a sampling rate of 44, 1 kHz.

The acoustic recordings are to be carried out once a year from the first to the fourth grade. In this study, we focus on the recordings of the target nouns (see Figure A1 in the Appendix), the Christmas picture (see Figure A2 in the Appendix), and Sentences 1 and 2 (see Table 1) from the first-grade recordings. The repetition of the first two sentences and the description of the Christmas picture elicited the most consistent data, meaning that the children made few errors and the durations of the descriptions were quite similar. This, in turn, made them most suitable for acoustic analysis as well as for the listening experiment.

Acoustic Analysis

For the acoustic analysis, the recordings of the children were segmented and annotated using Web-MAUS (Kisler et al., 2017; Schiel, 1999). The Praat (Boersma & Weenink, 2018) text grids were then manually checked and corrected. All acoustic analyses were also performed in Praat using our own scripts as well as publicly available by Jörg Mayer (Mayer, 2017), Shigeto Kawahara (http://user.keio.ac.jp/kawahara/resource.html), and David R. Feinberg (https://osf.io/dbrpf/).

f_o was measured in Sentences 1 and 2 and in the description of the Christmas picture. The formant frequencies were estimated in the vowels /aː/, /iː/, and /uː/ of the target nouns; the vowels /ɔ/, /yː/, and /uː/ of Sentence 1; and the vowels /a/, /oː/, and /iː/ of Sentence 2. The measurement was carried out at the midpoint of each vowel (time step 0.01; window length 0.025 s; maximum formant 5500 Hz). A two-dimensional acoustic vowel space (AVS) was created using the F1 and F2 of these vowels. The area of this polygon was calculated and, for legibility and ease of comparison, will be stated in kHz^2. The same vowels were used for determining voice quality (HNR). Tempo was calculated in canonical syllables per second (without pauses) in the first two sentences and the description of the Christmas picture. CoG and skewness were calculated from the sibilant spectra of /s/ in "kannst" (Sentence 2) and /z/ in "Sommer" (Sentence 1) and "Rose" (Sentence 2). The results of the acoustic measurements were used to answer the three research questions of this study (see Aims and Research Questions section).

Capturing GC

The basis for recording GC is the German Extended Personal Attributes Questionnaire (GEPAQ; Runge et al., 1981) and Children's Personal Attributes Questionnaire (CPAQ; Hall & Halberstadt, 1980), which measure stereotyped gender traits. Since Weirich and Simpson (2018) found different significant correlations between acoustics and gender role self-concept using GEPAQ, it would seem reasonable to investigate the relevance of these stereotypes in German children. From the adult questionnaires, we used five different statements that seemed to be transferable and transformed them into easily understandable questions (see Table 2). In addition, each child was asked three questions asked about stereotypical color, relationship, and toy preferences. The questions were proofread by an experienced primary school teacher.

Since the children in the selected primary schools already learn to evaluate each other with the help of smileys in the first grade, three different boxes (green, yellow, and red) were placed in the room to answer the questions. Emojis with corresponding expressions were depicted on each box (green: smiling face; yellow: neutral face; and red: sad face). After each question, the interviewed child was given a numbered table tennis ball that was to be placed in one of the boxes: green box for a positive answer, red for a negative answer, and

Table 2. Questions for capturing gender conformity.

Number	Question
1	**Bist du gut darin, neue Freunde zu finden?**
	Are you good at making friends?
GEPAQ	able/unable to pay attention to others
CPAQ	When I meet someone, I am always the first to try and make friends.
Answer	female: ☺ neutral: 😐 male: ☹
2	**Spielst du lieber mit Mädchen als mit Jungen?**
	Would you rather play with girls than with boys?
Answer	female: ☺ neutral: 😐 male: ☹
3	**Bist du bei vielen Sachen besser als die anderen in deiner Klasse?**
	Are you better at many things than the others in your class?
GEPAQ	self-confident/not self-confident
CPAQ	In most ways, I am better than most of the other kids my age.
Answer	female: ☹ neutral: 😐 male: ☺
4	**Hilfst du anderen gerne?**
	Do you like helping other people?
GEPAQ	helpful/not helpful to others
CPAQ	I do (not) help other people very much
Answer	female: ☺ neutral: 😐 male: ☹
5	**Stört es dich, wenn du in einem Spiel verlierst?**
	Do you get annoyed if you lose in a game?
GEPAQ	competitive/not competitive
CPAQ	I hate to lose a game or have other kids do better than me.
Answer	female: ☹ neutral: 😐 male: ☺
6	**Findest du es in Ordnung, zu weinen, wenn du traurig bist?**
	Is it OK to cry when you're sad?
GEPAQ	emotional/not emotional
CPAQ	I cry when things upset me.
Answer	female: ☺ neutral: 😐 male: ☹
7	**Welche dieser Farben gefällt dir am besten?**
	Which color do you like best?
Answer	female: pink neutral: yellow male: blue
8	**Mit welchem dieser Spielzeuge würdest du am liebsten spielen?**
	Which toy would you most like to play with?
Answer	female: princess neutral: wizard male: motorcyclist

Note. GEPAQ = German Extended Personal Attributes Questionnaire; CPAQ = Children's Personal Attributes Questionnaire.

yellow if the child was unsure/ambivalent/indifferent. This simulated a clear and child-friendly 3-point scale. Numerical values from 1 (stereotype "male") to 3 (stereotype "female") were then assigned for the evaluation of the eight questions.

The expected typical female, neutral, and male answers are listed in Table 2 together with an explanation of which statements were transformed from GEPAQ and CPAQ. The second question was aimed at the observation that children usually prefer relationships within their own sex (Rohrmann, 2017; Zosuls et al., 2011). For Question 7, three cards of different colors (pink, yellow, and blue) were shown to the children for selection. Expected was the stereotypical preference of pink for girls and blue for boys (Yeung & Wong, 2018). In Question 8, the children could choose from three different toy sets of the brand Playmobil. One set contained a pink princess stereotyped "for girls"; another set contained a motorcyclist representing a stereotypical toy "for boys." The toys also matched the experimental findings for toy preferences (Farr et al., 2018; Martin et al., 1995; Serbin et al., 2001; Todd et al., 2017). A third toy included a wizard without any specific gender categorization on the packaging (e.g., color pink for the princess) was used as a neutral variant.

Listening Experiment

The aim of the listening experiment was to find out how "male" or "female" the voices of the individual children are perceived to be. This, in turn, provides us with a good indication of those children producing robust phonetic correlates of gender in their voices, which is important to answer the three research questions (see Aims and Research Questions section).

The recordings of Sentences 1 and 2 together with about 5–10 s of the description of the Christmas picture were used for the listening experiment. Prior to performing the hearing experiment, signals were downsampled to 22.1 kHz and normalized to 70 dB using Praat.

Pretest. Since the listening experiment contains spontaneous speech samples, a pretest was conducted to rule out the possibility that the verbal content of the utterances might have an effect on the perception of a child's gender. For the pretest, the same shortened recordings of the Christmas picture used for the listening experiment were transcribed orthographically. The participants decided on a 7-point scale how much they perceive the written representation of the spontaneous description as "boy" (Value 1) or "girl" (Value 7). In addition to this assessment, the sex and age of the participants were recorded. The pretest was created using SoSci Survey (Leiner, 2019) and performed online. The participants were acquired via Facebook and WhatsApp. Sixty-eight people took part (60 women and eight men, 16- to 85-year-olds). If speakers were assigned values between 1 and 2 or 6 and 7 in the picture description on average, these recordings were excluded from the listening experiment and further calculations.

Listeners

A different group of 167 listeners participated in the listening experiment. The listeners were acquired via Facebook, WhatsApp, and via e-mail distribution lists of the Friedrich Schiller University Jena. The participants were 16- to 71-year-olds (M = 33.2 years, SD = 13.7 years), representing a broadly distributed population sample. The listeners were also asked to identify their gender, mother tongue, and if they have any hearing problems. To evaluate whether contact with children or expertise in voice assessment have an effect on gender perception, the listeners were asked if they have their own children, if they have regular contact with children (e.g., at work or apprenticeship), and if they have any expertise in voice evaluation. All participants were asked to carry out the listening experiment in a quiet place and to use headphones if possible.

Table 3 shows the participant demographics. Many listeners were acquired in humanities courses, which are mainly attended by female students. This may be the reason why the majority of the listeners are female. Less than half of the participants reported some expertise in the evaluation of voices. Less than half of the listeners have regular contact with children, with approximately a third of the sample having their own children. Listeners who reported any hearing problems would have been excluded, but that was not the case.

Main Experiment

The listening experiment was conducted online using the Percy software (Draxler, 2011, 2014). Stimuli were presented in random order (and randomized for each session). Each stimulus could be listened to twice. Listeners decided on a 7-point scale whether they had heard a "boy" (Value 1) or a "girl" (Value 7). Before the listeners started the experiment, they read the following text: "Wenn Sie die gehörte Stimme eindeutig als Jungen wahrnehmen, wählen Sie bitte den linken Kreis. Wenn Sie die gehörte Stimme eindeutig als Mädchen wahrnehmen, wählen Sie bitte den rechten Kreis. Die anderen Kreise liegen dazwischen." ("If you clearly perceive the voice as a boy, please select the left circle. If you clearly perceive the voice as a girl, please select the right circle. The other circles are in between"). During the experiment, there was the same question for each stimulus: "Haben Sie eher einen Jungen (links) oder ein Mädchen (rechts) gehört?" ("Did you rather hear a boy (left) or a girl (right)?"). We used the same anchor terms *boy* and *girl* as in previous studies (e.g., Günzburger et al., 1987; Simpson et al., 2017). We relied on the assumption that listeners would choose "boy" when they clearly perceive the child as masculine and "girl" if they clearly perceive the child as feminine.

There were a total of 186 stimuli (two sentences + spontaneous sample × 62 children). The duration of the experiment for each subject was shortened to approximately 15 min by dividing the stimuli into three groups

Table 3. Demographic properties of the participants.

Item	Demographic properties		
Mother tongue	98.0% German	1.0% Spanish	1.0% Indonesian
Sex	71.3% Female	27.5% Male	1.2% Diverse
Expertise	38.9% Yes	55.7% No	5.4% No information
Contact with children	40.1% Yes	58.7% No	1.2% No information
Own children	33.5% Yes	65.9% No	0.6% No information

each with 63 or 60 stimuli. The program was written in such a way that the three groups of stimuli were approximately evenly distributed across different participants. Parts 1 and 3 were listened by 56 participants and Part 2 by 55 participants.

Statistics

To evaluate how acoustic patterns of voices that lead to a robust gender impression differ from those with more variable gender responses (Research Question 1), we calculated a mean gender perception index for each child of the listening experiment first. Children with values less than 2 or greater than 6 are considered as “unambiguous.” Intraclass correlation coefficients (ICCs) were calculated to show how strongly the perception of the listeners resembles each other. To compare the gender perception index of boys and girls, Wilcoxon rank-sum tests were run.

To identify different groups of speakers, we performed a hierarchical cluster analysis in R (R Core Team, 2022, Version 4.2.1) with the psych package (Revelle, 2022). All items were *z* score normalized, and we used the Euclidean distance metric and the ward.D2 algorithm. The number of clusters was determined with the cubic clustering criterion using the NbClust package (Charrad et al., 2022) in R, with two as minimum and six as maximum. Evidence of the cluster validity was provided by one-way analysis of variance (ANOVA) tests. The mean values of gender perception, GC, and acoustic measurements (which are described in the Method: Acoustic Analysis section) were used for clustering. We are aware that cluster analysis is an exploratory technique. However, due to the large number of acoustic parameters, this method is well suited to gain an impression of the possible factors influencing gender perception. In addition, we compared the results of the acoustic analysis of the children who are perceived as most female (gender perception index > 6) and male (gender perception index < 2) in two-sample independent *t* tests.

To find out which acoustic parameters best predict the perception of gender in children’s voices (Research Question 2), we ran two multiple linear regression models, one for each sentence of the listening experiment. We excluded the spontaneous description of the Christmas picture because the acoustic content is not comparable. We used Sentences 1 and 2 instead and considered the differences between the two. Gender perception was the dependent variable, and the acoustic parameters were independent variables in the models. To investigate the influence of demographic background (e.g., contact to children) on gender perception, each of these parameters were used as independent variables across several simple regression models.

To analyze the connection between acoustic parameters and gender role self-concept (Research Question 3), we determined a mean GC index for each child, based on the answers of the questionnaire (see Method: Capturing Gender Conformity section). Furthermore, we used Cronbach’s alpha coefficient to measure the reliability of the eight items. Wilcoxon rank-sum tests were run to compare the GC index of boys and girls. The connection between self-described GC, gender perception, and acoustic parameters was evaluated using Spearman ρ correlations as well as linear regression models. In addition, to get an idea of how acoustics, gender role self-concept, and gender perception could be related, the GC index was also included in the cluster analysis described above.

Results

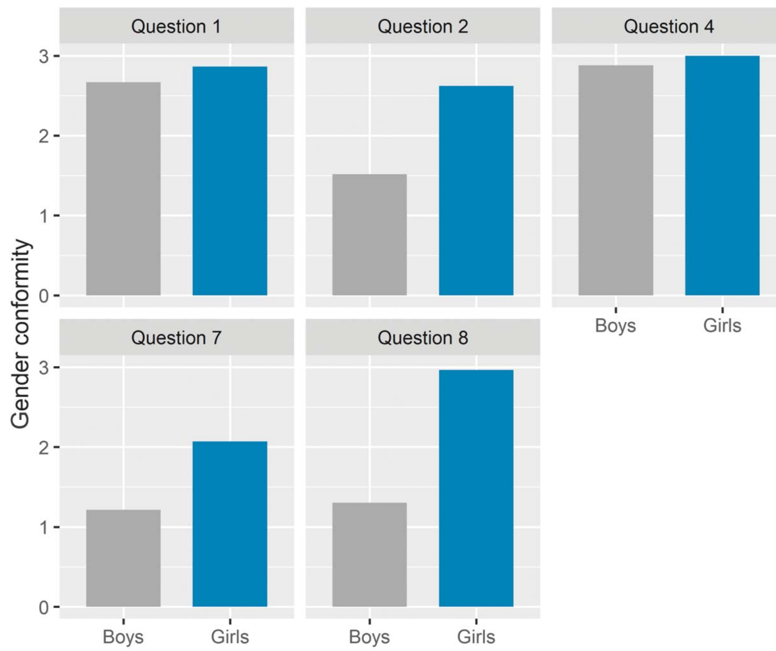
Listening Experiment

Pretest. In the pretest, listeners judged possible gender-bias in the verbal content of the children’s spontaneous utterances independently of the phonetic form on a 7-point scale (1 means *unambiguous male* and 7 means *unambiguous female*). Results show values of the individual children between 1.88 and 6.31. The mean for all subjects is 3.94 (SD = 1.11). If we split the group into birth-assigned sex, boys and girls do not differ significantly in a two-sample independent *t* test (boys M = 3.99, SD = 1.13, and girls M = 3.89, SD = 1.11; t = 0.32, and p = .75). These mean values of about 4 (average of the 7-point scale) suggest that the verbal content of the spontaneous image descriptions does not prejudice judgments of a child’s gender in general. However, the ICC is very low, ICC(2, 1) = .32, p < .01*), indicating poor interrater reliability according to Ko and Li (2016). From this, we can conclude that the impressions of the participants’ gender based on semantic content alone differ widely.

Despite there not being any overall differences in the means, three speakers received unambiguous gender assignments with mean values between 1 and 2 or 6 and 7 in the spontaneous picture description. We excluded these recordings from the listening experiment.

Main experiment. Figure 1 shows the mean gender perception values for each child in the listening experiment separated by gender. The higher the value, the more female the voices of the children are perceived to be; the lower the value, the more male. In other words, the boys with the lowest values at the bottom of the left-hand plot in Figure 2 are perceived to be most male (e.g., 1b 37 m: 1.25); the girls with the highest values at the top of the right-hand plot, the most female (e.g., 1b 40w: 6.64).

Figure 1. Averages of the five reliable questions of the gender conformity questionnaire separated by gender.



The result of a Shapiro–Wilk test fail to indicate normal distribution (W = 0.94, p = .01*), so a Wilcoxon rank-sum test was used to compare gender perception of boys and girls. Results show that gender perception differs significantly between these groups (boys M = 3.06, SD = 1.37 and girls M = 4.97, SD = 1.03; W = 145, p < .01*). In other words, overall, boys are perceived more as boys and girls more as girls. This difference applies both to the stimuli of Sentences 1 (W = 211, p < .01*) and 2 (W = 157, p < .01*) as well as to the spontaneous picture description, even if we exclude the three children with values between 1 and 2 or 6 and 7 in the pretest (W = 121, p < .01*).

However, the plots in Figure 2 also suggest that any group means belie a more complex and interesting picture. On the one hand, if we take values less than 2 or greater than 6 as a sign of systematic gender perception, 16 children are producing phonetic patterns that receive a relatively unambiguous gender attribution by the listeners. The remaining 46 children, by contrast, are producing acoustic patterns that lead to more ambivalent, variable gender responses from the listeners. In other words, to varying degrees, a majority of the children's utterances lack the cues required by listeners to make a uniform gender attribution. Therefore, it is not surprising that the ICC including the ratings of all children shows a poor interrater reliability, ICC(1, 1) = .47, p < .01*, whereas the interrater reliability of the 16 children with an unambiguous gender attribution is good, ICC(1, 1) = .77, p < .01* (Ko & Li, 2016).

GC

Mean values of GC calculated for each child range between 1.38 and 2.88 on the 3-point scale of the questionnaire (one stereotype "male," three "stereotype," and "female"; see Method: Capturing Gender Conformity section). When we compare GC of boys and girls, the male mean conformity index is 2.0 (SD = 0.25); the female is 2.5 (SD = 0.17). To test for differences between the groups, Wilcoxon rank-sum tests were used for all eight questions because a Shapiro–Wilk test indicates the lack of normal distribution in the data (W = 0.96, p = .04*). To avoid alpha error accumulation, the results were Bonferroni–Holm corrected.

Although the GC between boys and girls is significantly different overall (W = 30, p < .01*), the groups only differ significantly in their responses to Questions 2 and 8. Specifically, both groups usually prefer to play with children of their own sex (boys M = 1.52, SD = 0.67 and girls M = 2.62, SD = 0.56; W = 129, p = .01*). Furthermore, in the choice of the preferred toy, the responses of boys and girls differ significantly (boys M = 1.30, SD = 0.53 and girls M = 2.97, SD = 0.19; W = 19, p < .01*). We used Cronbach's alpha coefficient to measure the reliability of the eight items. If we include all questions, the result is only α = .58. Questions 3, 5, and 6 have to be removed to achieve an acceptable value of α = .71. For this reason, only the average values from the five other questions were included in further calculations (see Figure 2).

Figure 2. Average of the perceived gender rating in the listening experiment. Boys left (gray), girls right (blue).

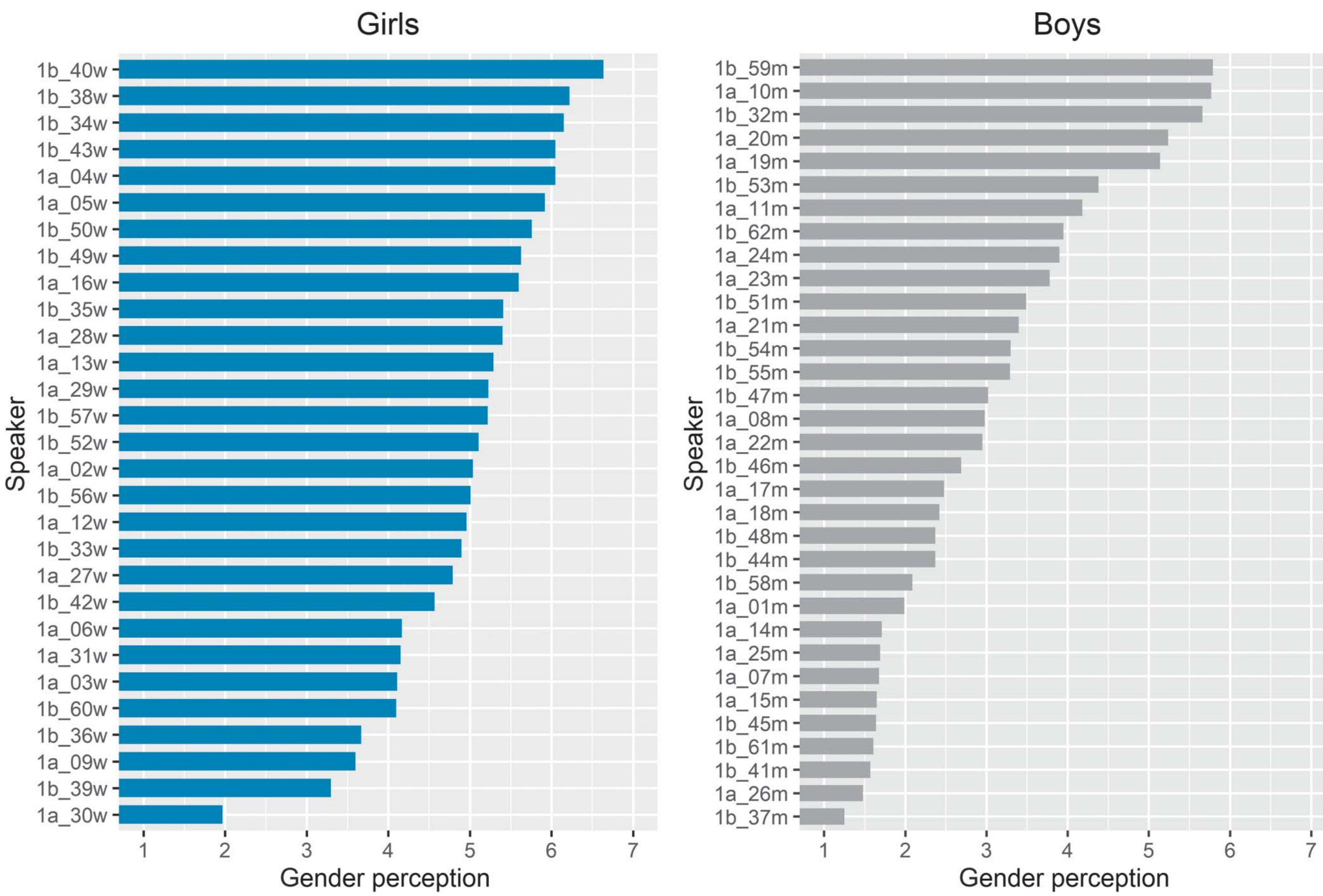


Figure 3. Hierarchical cluster of the children.

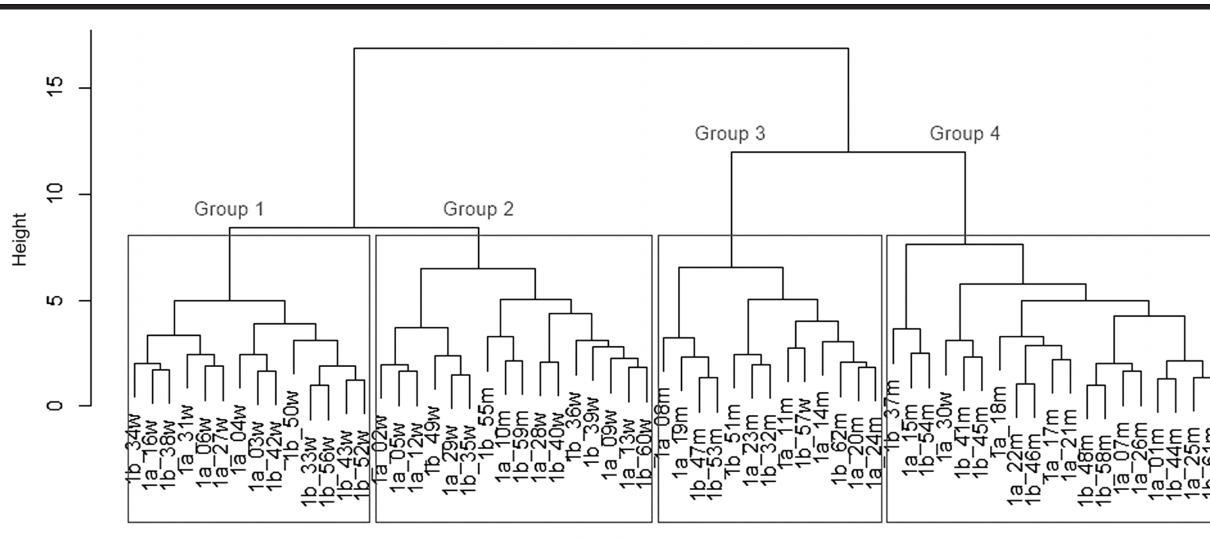


Table 4. Arithmetic means, standard deviations, one-way ANOVAs, and *p* values (Bonferroni–Holm corrected) of the parameters used for the four different clusters (raw values).

	1		2		3		4			
Group	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***F***	***p***
GC	2.73	0.27	2.56	0.23	1.89	0.41	1.99	0.32	16.0	< .01*
Perception	5.19	0.78	4.94	1.03	4.05	1.11	2.14	0.62	42.3	< .01*
f_o (Hz)	268	16	252	20	275	21	242	17	10.1	.01*
AVS (kHz^2)	1.07	0.33	1.28	0.54	1.10	0.26	1.09	0.58	0.7	1.0
HNR (dB)	19.8	2.4	20.1	2.1	19.4	2.0	20.3	2.5	0.4	.72
Tempo (syll/s)	4.24	0.38	4.16	0.33	4.68	0.25	4.29	0.42	5.6	.01*
CoG (Hz)	3889	430	5251	641	2965	688	3969	820	28.5	< .01*
Skewness	0.87	0.31	–0.10	0.36	1.45	0.84	0.75	0.64	18.5	< .01*

Note. ANOVA = analysis of variance; GC = gender conformity; f_o = fundamental frequency; AVS = acoustic vowel space; HNR = harmonics-to-noise-ratio; CoG = center of gravity.

* = significant difference ($p < .05$)

Cluster Analysis

The cubic clustering criterion determined four different clusters in the hierarchical cluster analysis with a value index of 149.8. The resulting dendrogram is shown in Figure 3. The differences between the four groups are described in Table 4 (raw values). The analysis includes gender perception, GC, f_o, AVS, HNR, tempo, CoG, and skewness (see Method: Statistics section).

Group 1 contains only girls with feminine GC and high f_o values. These children also have the highest gender perception rate, which means they were judged to be most female. Therefore, it is possible that GC and f_o have a large influence on gender perception in this group. The values for AVS, HNR, tempo, CoG, and skewness are similar to the values in Group 4.

Group 2 contains 19% boys and 81% girls with feminine GC and gender perception similar to Group 1. By contrast to Group 1, however, these children have larger AVS, which previous studies have also found for girls (Cartei et al., 2019). In addition, the sibilants of this group show the highest CoG and the lowest (negative) skewness values, which has also repeatedly been found in female sibilants (Flipsen et al., 1999; Li et al., 2016; Shadle et al., 2011). However, the f_o of this group is lower than in Group 1. Group 2 indicates that high gender perception ratings could be also related to a large AVS, high CoG, and low (negative) skewness.

Group 3 contains 92% boys with the highest f_o and tempo values of all groups. Simpson et al. (2017) showed that children with high f_o and high speech tempo are often perceived as girls. On the other hand, the children in Group 3 produce typically male sibilants with the highest (positive) skewness and lowest CoG values. That could be the reason why these children receive the most neutral or ambivalent gender ratings with a mean perception value of 4.05. This is supported by the fact that the 16 children with the most unambiguous gender perception ratings carried out in the listening experiment cannot be found in Group 3.

Group 4 is dominated by boys (95%) with the most male ratings. The values for GC are also more masculine than in Groups 1 and 2, and these children have the lowest f_o. In this group, gender perception seems to be related only to f_o and maybe GC (like in Group 1). The one-way ANOVA tests reveal significant differences between the

Table 5. Arithmetic means, standard deviations, and results of two-sample independent *t* tests together with effect sizes of the most female- (> 6) and male- (< 2) sounding children of the listening experiment.

	Male (n = 11)		Female (n = 5)				
Variable	***M***	***SD***	***M***	***SD***	***t***	***p***	***d***
f_o (Hz)	242	20	269	15	–2.7	.02*	1.5
AVS (kHz^2)	1.04	0.37	1.30	0.23	–1.4	.17	0.79
HNR (dB)	20.20	2.91	21.23	3.54	–0.6	.54	0.34
Tempo (syll/s)	4.29	0.53	4.36	0.30	–0.3	.78	0.15
CoG (Hz)	4194	789	4435	691	–0.6	.57	0.32
Skewness	0.76	0.76	0.53	0.47	0.6	.55	0.33

Note. f_o = fundamental frequency; AVS = acoustic vowel space; HNR = harmonics-to-noise-ratio; CoG = center of gravity.

clusters of gender perception, GC, f_o, AVS, CoG, and skewness but not of AVS and HNR (see Table 4).

Acoustic Analysis

Table 5 describes the results of the acoustic analysis (means of all acoustic measurements; see Method: Acoustic Analysis section) of the children who are perceived as most female (> 6) and male (< 2) in the listening experiment. Significant differences together with large effect size can only be found for f_o, with male-sounding children having lower values than female-sounding children. The female-sounding children show higher values for AVS than the male-sounding children with a medium effect. Small effect sizes can be found for HNR and CoG with higher values for female-sounding children. There is also a small effect in the comparison of skewness; in this case, the mean values of male-sounding children are higher than the female ones. However, a large standard deviation can be found in female- as well as in male-sounding children, suggesting a large variation within both groups.

There is no significant difference between the average tempo of female- and male-sounding children. As tempo was calculated with mean values from Sentences 1 and 2 together with the picture description, we also compared the number of syllables of the spontaneous recording. No significant differences were found (boys M = 21.3, SD = 6.6, and girls M = 19.9, SD = 6.3; W = 537, p = .42).

Gender Perception and Acoustic Analysis

To find out which acoustic parameters best predict the gender perception ratings, we ran two multiple linear regression models, one for each sentence of the listening experiment. We used Sentences 1 and 2 to consider the differences between the two. Gender perception was the dependent variable, and f_o, AVS, HNR, tempo, CoG, and skewness were independent variables in the models (means of acoustic easurements; see Method: Acoustic Analysis section).

Table 6 shows the estimated values, p values and effect sizes (Cohen's f^2) for each different parameter and the F values, p values, and the adjusted R^2 of the whole model. As we can see, only f_o has a significant influence on gender perception in both sentences with medium effect size (see Figure 4). With f_o, we can predict higher female gender ratings. By contrast, AVS, HNR, tempo, CoG, and skewness do not play a role in the gender perception of the children in the present sample. If we split the group into birth-assigned sex, the main effect of f_o on gender perception is still significant in both sentences (Sentence 1 boys: *Coeff.* = 2.98, p = .01* and girls *Coeff.* = 0.03, p = .01*; Sentence 2 boys: *Coeff.* = 0.24, p = .01* and girls *Coeff.* = 0.28, p = .02*).

No significant effect of listener's gender, $F(2, 10348) = 1.18$, p = .31; self-reported expertise, $F(2, 10348) = 1.35$, p = .26; regular contact to children, $F(2, 10348) = 0.36$, p = .69; or having own children, $F(2, 10348) = 1.89$, p = .15, on gender perception can be found.

GC and Acoustic Analysis/ Gender Perception

Table 7 shows Spearman ρ correlations between the reported GC of the children and the means of acoustic parameters, which we described in the Method: Acoustic

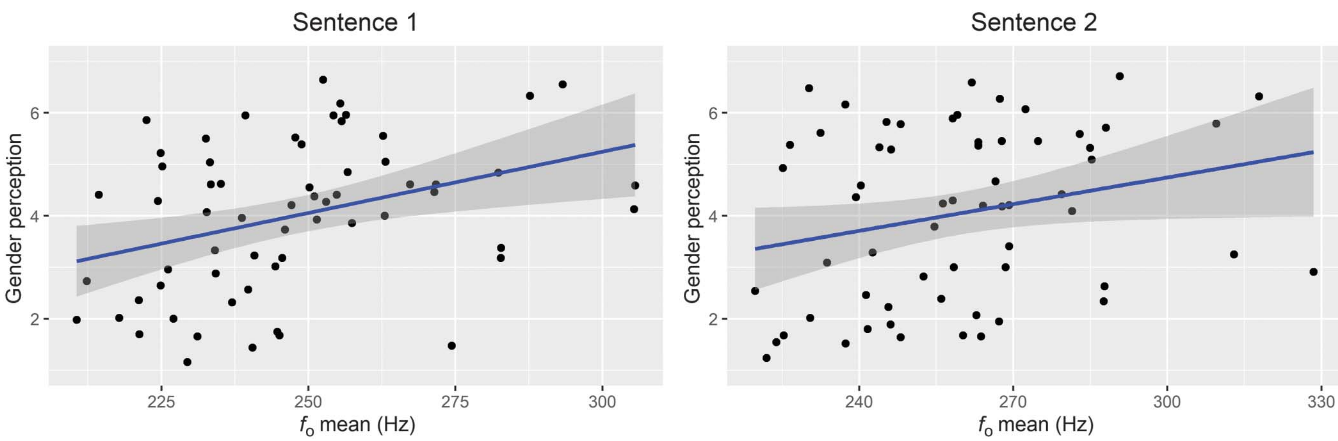
Table 6. Coefficients, standard errors, t values, p values, and effect sizes of the multiple linear regression models (Sentences 1 and 2) testing the main effects of acoustic parameters on gender perception ratings of the listening experiment.

Variable	Sentence 1					Sentence 2				
	Coeff.	*SE*	*T*	f^2	*p*	Coeff.	*SE*	*t*	f^2	*p*
Intercept	−5.07	3.43	−1.48	.04	.14	−3.29	3.48	−0.95	.02	.35
f_o (Hz)	0.03	0.01	3.5	.22	.001*	0.03	0.01	2.8	.15	.007*
AVS (kHz^2)	0.21	0.22	1.0	.02	.34	0.09	0.22	0.4	.00	.67
HNR (dB)	−0.006	0.08	−0.1	.00	.93	−0.005	0.06	−0.9	.02	.36
Tempo (syll/s)	−0.08	0.36	−0.2	.00	.82	0.23	0.46	0.5	.00	.63
CoG (Hz)	0.0004	0.01	1.8	.06	.07	0.0003	0.01	1.3	.03	.20
Skewness	0.2	0.25	1.2	.02	.25	−0.16	0.24	−0.7	.01	.49
	$F(6, 55) = 2.33$, p = .045*					$F(6, 55) = 2.50$, p = .033*				
	R^2 (adj.) = .12					R^2 (adj.) = .13				

Note. Coeff. = coefficient; *SE* = standard error; f^2 = effect size; f_o = fundamental frequency; AVS = acoustic vowel space; HNR = harmonics-to-noise-ratio; syll = syllable; CoG = center of gravity; R^2 (adj.) = adjusted R^2.

* = significant difference (p < .05)

Figure 4. Relationship between perception and f_o of Sentences 1 (left) and 2 (right). f_o = fundamental frequency.



Analysis section. All p values were Bonferroni–Holm corrected. f_o, AVS, HNR, and tempo do not correlate significantly with GC, but we do find appreciable correlations between GC and acoustic characteristics of the sibilants of Sentences 01 and 02 (/s/ in *kannst* "can" and /z/ *Sommer* in "summer" and *Rose* "rose"). The higher the value of GC, that is, the more "female" the children have responded to the interview questions, the higher the CoG values and the lower (negative) the skewness values are. The negative correlation between GC and skewness achieves significance. If we split the group into birth-assigned sex, the negative correlation between GC and skewness is still significant in boys (see Figure 5), but not in girls. There is also a positive (but not significant) correlation between GC and CoG in the boy group, but not in the girl group. A linear regression model confirms the effect of GC on skewness, $F(1, 60) = 15.64$, $p < .01^*$, and in interaction with gender, $F(3, 58) = 8.40$, $p < .01^*$. The effect remains in the boy group, $F(1, 31) = 14.06$, $p < .01^*$, but not in the girl group, $F(1, 27) = 0.15$, $p = .70$. A significant effect of GC on CoG, $F(1, 60) = 10.94$, $p < .01^*$, and in interaction with gender, $F(3, 58) = 5.10$, $p < .01^*$, can be found too. As with skewness, the effect is significant in boys, $F(1, 31) = 6.08$, $p = .02^*$, but not in girls, $F(1, 27) = 0.34$, $p = .57$.

Furthermore, as shown in Table 7, there is a significant correlation between GC and gender perception across boys and girls, but not within gender. A linear regression model verifies a significant effect of GC on gender perception, $F(1, 60) = 21.19$, $p = .02^*$, and in interaction with gender, $F(3, 58) = 12.52$, $p < .01^*$. If we spilt the children by birth-assigned sex, there is neither an effect of GC on gender perception in boys, $F(1, 31) = 0.18$, $p = .68$, nor in girls, $F(1, 27) = 0.54$, $p = .47$.

Discussion

Acoustic Patterns of Female- and Male-Sounding Children

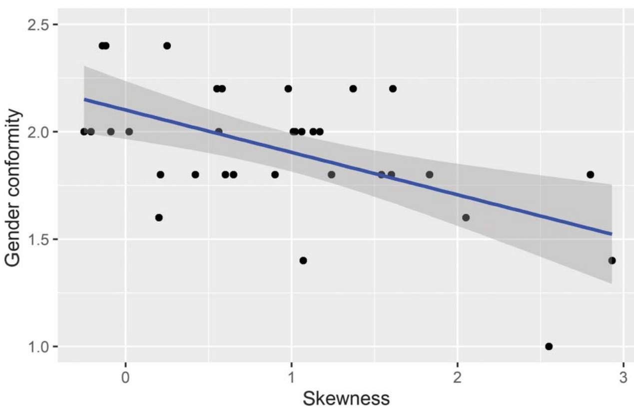
Our first research question was how acoustic patterns of voices, which lead to a robust rating of a child's gender, differ from those that result in more variable

Table 7. Spearman ρ correlations between gender conformity (GC) and acoustic parameters (p values Bonferroni–Holm corrected).

	Rho			*p*		
GC	**All**	**Boys**	**Girls**	**All**	**Boys**	**Girls**
Perception	.53	.16	−.14	.04*	1.0	1.0
f_o (Hz)	−.03	.63	.10	1.0	.40	.62
AVS (kHz^2)	−.12	.15	−.40	1.0	.82	.22
HNR (dB)	−.02	−.06	−.39	.89	.76	.23
Tempo (syll/s)	−.29	.22	−.30	.08	.88	1.0
CoG (Hz)	.39	.32	.06	.01*	.42	.76
Skewness	−.35	−.46	.12	.03*	.05*	1.0

Note. f_o = fundamental frequency; AVS = acoustic vowel space; HNR = harmonics-to-noise-ratio; syll = syllable; CoG = center of gravity.
* = significant difference ($p < .05$)

Figure 5. Relationship between gender conformity and skewness for boys.



gender responses. To answer this, we conducted a listening experiment on gender perception and an acoustic analysis of the stimuli used in the listening experiment. A hierarchical cluster analysis was used to identify different types of speakers. Furthermore, we compared the results of the acoustic analysis of the children who are perceived as most female and most male.

Results show that gender perception of boys and girls differs significantly (see Results: Listening Experiment section). This substantiates a common finding in previous studies that, given stimuli with enough linguistic content, listeners are on average able to identify the birth-assigned sex of a prepubertal child at a level better than chance (Günzburger et al., 1987; Karlsson, 1987; Kaya et al., 2017; Simpson et al., 2017). Similar to Simpson et al. (2017) and Funk et al. (2018), it was apparent that 16 of the 62 children exhibit very strong markers of gender in their voices and achieve very high (> 6) or very low (< 2) values on the 7-point scale of the listening experiment. The other children received more gender-ambivalent rating, suggesting that they are not producing such unequivocal acoustic patterns.

If we have a look at the acoustic analyses (see Results: Acoustic Analysis section), we found that significant differences only exist for f_o with lower frequencies for male- than female-sounding children. In addition, the female-sounding children have shown higher values for AVS than the male-sounding children.

In a cluster analysis (see Results: Cluster Analysis section), four different clusters were emerged. Two clusters include children with high perceptual female ratings, who also have high self-reported female GC. The children in these clusters have high f_o, have a large AVS, or produce sibilants with high CoG and low (negative) skewness values, which have repeatedly been found in female speakers. One cluster contains children who are perceived as boys with low f_o. The cluster with the most ambiguous children carried out in the listening experiment shows on average high f_o and high speech tempo, which is associated with girls. On the other hand, these children have typically male sibilants with low CoG and high (positive) skewness values. This mixture of female and male characteristics could be the reason why listeners find it difficult to judge the gender of these children.

From these results, we can conclude that there is a range of acoustic parameters that may influence gender perception and lead to a robust rating of a child's gender. At the same time, it seems that a particular female or male gender rating does not rely on a single acoustic parameter or cluster of parameters. In other words, one child may receive a strong female rating based on high average f_o, while another strong male rating based on sibilant spectra.

Relationship Between Gender Perception and Acoustics

With Research Question 2, we examined which acoustic parameters predict gender perception. Two multiple linear regression models were run, one for each sentence of the listening experiment.

We found that only f_o has a significant main effect on gender perception across and within gender (see Results: Gender Perception and Acoustic Analysis). In other words, the higher a speaker's f_o is, the higher their female rating will be. This result remains even if the group of children is divided by birth-assigned sex. With a lack of any significant gender-specific anatomical differences in the prepubertal larynx or vocal tract (Fitch & Giedd, 1999; Kahane, 1978), these differences must have been learned and are being used by the children to encode and the listeners to decode a speaker's gender. In other words, it is probable that prepubescent girls are already adapting to the higher pitched voices of women and boys to the lower pitched voices of men before reaching puberty, especially if they identify themselves with this gender role. There is no significant effect of listener's demographic characteristics (gender, self-reported expertise, or contact to children) on gender perception.

Relationship Between GC and Acoustics/Gender Perception

To find out how acoustic parameters and gender perception are related to a child's gender role self-concept (Research Question 3), we used a questionnaire to calculate a GC index for each child. Our results verify that GC is significantly different when the children are grouped according to birth-assigned sex (see Results: Gender Conformity section). Boys and girls prefer to play with partners of their own sex and have different toy preferences, which is consistent with the findings of previous studies (Farr et al., 2018; Martin et al., 1995; Rohrmann, 2017; Serbin et al., 2001; Todd et al., 2017; Zosuls et al., 2011).

GC seems to be related to the spectral structure of sibilants. A significant positive correlation between CoG and GC was found. Furthermore, significant negative correlations between skewness and GC were found in the whole group, but above all in the male speakers. This finding is in line with the results of Li et al. (2016), who found significant differences in skewness values between boys and girls and also correlations between gender role self-concept and skewness (significant for boys).

The GC index correlates with listeners' perception of gender across boys and girls, but not within the groups. A linear regression model verifies a significant effect of GC on gender perception in general: The more female the children judge themselves to be, the more female their voices are perceived to be by adult listeners. This effect does not persist if only boys or girls are included in the model. Therefore, it cannot be ruled out that birth-assigned sex has an influence on this connection. Results of the cluster analysis (see Results: Cluster Analysis section) also suggest an influence of GC on gender perception, especially for two of the four groups (Groups 1 and 4).

Limitations

In our study, 62 children were recorded (see Method: Speakers section). We are aware that more children would be needed in order to achieve high statistical power (> .80) in some of the statistical tests used. A post hoc power analysis with G*Power Software (Faul et al., 2009) was used to calculate the power of significant results. With 62 children, the power is .98 in Sentence 1 and .81 in Sentence 2 if we consider the factor f_o. The differences of f_o between male- and female-sounding children result in a power of .84. In the listening experiment, we can calculate a power of .74 for the different gender perception of boys and girls. The differences of GC between boys and girls achieve a statistical power of .92. Nevertheless, it was not possible for us to record more children. Since we wanted to keep the context (dialect and living environment) as similar as possible, only two primary schools were considered. From these, all first-grade children whose parents had consented were included.

In this study, we can only present results of German children who are perceived by mainly German listeners. We do not assume that our findings are representative of or transferable to other languages. It is possible that German children may be perceived differently by other native speakers or that children with a different mother tongue may have a different effect on German listeners (see Karlsson & Rothenberg, 1992).

In the recordings (see Method: Acoustic Recordings section), it is possible that the children had accommodated some aspects of the adult speaker's voice. Since the first grade children were not yet able to read, repeating prerecorded utterances was seen as the only way to elicit tightly controlled sentence-length utterances.

We tried to eliminate interference in the recordings by using a sound shield in quiet rooms at the schools and stopping recordings during the lesson breaks. However, background noise from other pupils could not be ruled out. For this reason, we manually checked sensitive measurements such as f_o.

The questionnaire used to capture GC (see Method: Capturing Gender Conformity section) was based on the PAQ and GEPAQ, which are primarily designed to measure personality traits. However, given the few alternative options, we used these questions to examine stereotyped gender role self-concepts. In addition, the questionnaire was age appropriate and child friendly, since a 3-point scale and only eight questions were used to ensure that the children were not overwhelmed. It is clear that a more finely grained scale is likely to be more informative. Moreover, we had to exclude three questions to get a reliable questionnaire. For these reasons, a more differentiated scale together with some open-ended questions will be used for the second, third, and fourth graders. Instead of creating a questionnaire with smiley boxes, it might have been more convenient to ask the children's parents about the estimated gender conformity. However, we decided to exclude the children's parents from the study, as we specifically wanted to acquire self-reported data on gender role.

In the listening experiment (see Method: Listening Experiment section), we used a 7-point scale with the anchor terms *boy* and *girl* to evaluate the gender perception of the children. It is possible that the terms *masculine* and *feminine* would have been also suitable to capture gender perception, but we have chosen these anchor terms to receive comparable results to previous studies (e.g., Günzburger et al., 1987; Simpson et al., 2017). It is also possible that using terms such as masculine/feminine could have been variably interpreted by listeners, with some continuing to make gender judgments, whereas others might indeed be judging masculine/feminine aspects of the voices beyond any initial judgment of gender.

We used a cluster analysis (see Results: Cluster Analysis section) to identify different types of speakers and to gain an impression of the possible factors influencing gender perception. However, this kind of analysis is an exploratory technique. Furthermore, the one-way ANOVA tests do not reveal significant differences between the clusters in the parameters AVS and HNR. In addition, the cubic clustering criterion could be influenced by correlating parameters such as GC and gender perception. Therefore, it is possible that another number of clusters could be more suitable to find different types of speakers, even when the cubic clustering criterion fits best for four clusters. However, the cluster analysis was only one of the methods used to explore the data, and more rigorous tests were also employed, such as the comparison of the acoustic analysis of the children who are perceived as most female/male using two-sample independent t tests, or the calculation of linear regression models to examine the effect of different acoustic parameters on gender perception. Nevertheless, one important outcome of the cluster analysis was to identify a group of children exhibiting variable (ambiguous) gender assignment, as well as the acoustic parameters that are likely to have played a role in causing this ambiguity for the listeners.

Further Steps

It has become clear that f_o plays an important role in gender perception (see Results: Gender Perception and Acoustic Analysis section) and that sibilant structure correlates with GC in the first graders, especially in the boy group (see Results: Gender Conformity and Acoustic Analysis/Gender Perception section). The results of our cluster analysis (see Results: Cluster Analysis section) suggest that a range of acoustic parameters may influence gender perception.

The next steps in our investigation include

- Refining the GC questionnaire (see Method: Capturing Gender Conformity section) with a larger scale for older children and more questions including typical female and male preferences for colors, toys, or sports.
- Examining in more detail the role played by individual acoustic parameters by using sophisticated voice morphing techniques provided by TANDEM-STRAIGHT software (Kawahara et al., 2008).
- Examining possible relationships between perceptual attribute ratings and the corresponding acoustic parameters by conducting listening experiments where listeners judge perceptual attribute pairs (e.g., high–low) of female- and male-sounding children on a 7-point scale (see Simpson et al., 2017).
- Analyzing how the vocal expression of gender and GC in individual children develops over the first years in primary education before reaching puberty.

Author Contributions

Riccarda Funk: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Adrian P. Simpson:** Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Data Availability Statement

The data set is not available yet. We will publish the data set in a public repository as soon as it is complete.

Acknowledgments

The project is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Grant 440882807, received by Adrian P. Simpson, http://www.dfg.de/). We would like to thank Tu Anh Nguyen Thi for drawing the beautiful pictures and Christoph Draxler (MLU Munich) for his help in constructing and creating the listening experiment. Most of all, we are very grateful to the children and their parents agreeing to take part in this project.

References

Audacity Team. (2021). *Audacity(R): Free Audio Editor and Recorder*. https://audacityteam.org/

Auszmann, A. & Neuberger, T. (2014). Age-and gender-related differences in formant structure during the stabilization process of vowels. *Olomouc linguistics colloquium 2014: Language use and linguistic structure* (pp. 663–676).

Bennett, S. (1981). Vowel formant frequency characteristics of preadolescent males and females. *The Journal of the Acoustical Society of America, 69*(1), 231–238. https://doi.org/10.1121/1.385343

Boersma, P. & Weenink, D. (2018). *Praat, doing phonetics by computer* [Computer software]. http://www.praat.org

Busby, P. A., & Plant, G. L. (1995). Formant frequency values of vowels produced by preadolescent boys and girls. *The Journal of the Acoustical Society of America, 97*(4), 2603–2606. https://doi.org/10.1121/1.412975

Cartei, V., Cowles, R., Banerjee, R., & Reby, D. (2014). Control of voice gender in pre-pubertal children. *British Journal of Developmental Psychology, 32*(1), 100–106. https://doi.org/10.1111/bjdp.12027

Cartei, V., Garnham, W., Oakhill, J., Banerjee, R., & Reby, D. (2019). Children can control the expression of masculinity and femininity through the voice. *Royal Society Open Science, 6*(7), 190656. https://doi.org/10.1098/rsos.190656

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2022). *NbClust: Determining the best number of clusters in a data set*. https://CRAN.R-project.org/package=NbClust

Draxler, C. (2011). Percy—An HTML5 framework for media rich web experiments on mobile devices. *Twelfth Annual Conference of the International Speech Communication Association*. (pp. 3339–3340).

Draxler, C. (2014). *Online experiments with the Percy software framework—Experiences and some early results* (pp. 235–240).

Farr, R. H., Bruun, S. T., Doss, K. M., & Patterson, C. J. (2018). Children's gender-typed behavior from early to middle childhood in adoptive families with lesbian, gay, and heterosexual parents. *Sex Roles, 78*(7–8), 528–541. https://doi.org/10.1007/s11199-017-0812-5

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behaviour Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferrand, C. T. (2000). Harmonics-to-noise ratios in normally speaking prepubescent girls and boys. *Journal of Voice, 14*(1), 17–21. https://doi.org/10.1016/S0892-1997(00)80091-0

Ferrand, C. T., & Bloom, R. L. (1996). Gender differences in children's intonational patterns. *Journal of Voice, 10*(3), 284–291. https://doi.org/10.1016/S0892-1997(96)80009-9

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511–1522. https://doi.org/10.1121/1.427148

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeny, J. (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research, 42*(3), 663–677. https://doi.org/10.1044/jslhr.4203.663

Funk, R., Voigt-Zimmermann, S. & Simpson, A. P. (2020). Junge oder Mädchen? ZurGeschlechtsidentifikation präpubertärer Stimmen [Boy or girl? Gender identificantion in prepubertal children]. *Proc. Phonetik und Phonologie 14*, 58–61. https://doi.org/10.25365/phaidra.159

Glaze, L. E., Bless, D. M., Milenkovic, P., & Susser, R. D. (1988). Acoustic characteristics of children's voice. *Journal of Voice, 2*(4), 312–319. https://doi.org/10.1016/S0892-1997(88)80023-7

Günzburger, D., Bresser, A., & Keurs, M. (1987). Voice identification of prepubertal boys and girls by normally sighted and visually handicapped subjects. *Language and Speech, 30*(1), 47–58. https://doi.org/10.1177/002383098703000104

Hall, J. A., & Halberstadt, A. G. (1980). Masculinity and femininity in children: Development of the children's personal attributes questionnaire. *Developmental Psychology, 16*(4), 270–280. https://doi.org/10.1037/0012-1649.16.4.270

Hasek, C. S., Singh, S., & Murry, T. (1980). Acoustic attributes of preadolescent voices. *The Journal of the Acoustical Society of America, 68*(5), 1262–1265. https://doi.org/10.1121/1.385118

Kahane, J. C. (1978). A morphological study of the human prepubertal and pubertal larynx. *Journal of Anatomy, 151*(1), 11–19. https://doi.org/10.1002/aja.1001510103

Kallvik, E., Lindström, E., Holmqvist, S., Lindman, J., & Simberg, S. (2015). Prevalence of hoarseness in school-aged children. *Journal of Voice, 29*(2), 260.e1–260.e19. https://doi.org/10.1016/j.jvoice.2013.08.019

Karlsson, I., & Rothenberg, M. (1987). Sex differentiation cues in the voices of young children of different language background. *The Journal of the Acoustical Society of America, 81*(S1), S68–S69. https://doi.org/10.1121/1.2024359

Karlsson, I., & Rothenberg, M. (1992). Inter-cultural variations in gender-based language differences in young children. *Quarterly Progress and Status Report,* 1–17.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. & Banno, H. (2008). Tandem straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *IEEE international conference on acoustics, speech and signal processing* (pp. 3933–3936).

Kaya, H., Salahb, A. A., Karpovc, A., Frolovae, O., Grigoreve, A., & Laykso, E. (2017). Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech & Language, 46,* 268–283. https://doi.org/10.1016/j.csl.2017.06.002

Kisler, T., Reichel, U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45,* 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Ko, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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*(3), 1455–1468. https://doi.org/10.1121/1.426686

Leiner, D. J. (2019). *SoSci Survey* (Version 3.1.06) [Computer software]. https://www.soscisurvey.de

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. https://doi.org/10.1016/j.wocn.2016.05.004

Martin, C. L., Eisenbud, L., & Rose, H. (1995). Children's gender-based reasoning about toys. *Child Development, 66*(5), 1453–1471. https://doi.org/10.2307/1131657

Mayer, J. (2017). *Phonetische Analysen mit Praat. Ein Handbuch für Ein- und Umsteiger* [Praat manual]. https://praatpfanne.lingphon.net/das-praat-handbuch/

McCormack, P. F. & Knighton. T. (1996). Gender differences in the speech patterns of two and a half year old children. *Proc. Sixth Australian International Conference on Speech Science and Technology* (pp. 337–342).

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*(4), 1995–2003. https://doi.org/10.1121/1.4916202

Munson, B., Lackas, N., Koeppe, K. (2022). Individual differences in the development of gendered speech in preschool children: Evidence from a longitudinal study. *Journal of Speech, Language, and Hearing Research 65,* 1311–1330. https://doi.org/10.1044/2021_JSLHR-21-00465

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*(6), 2988–2998. https://doi.org/10.1121/1.1370525

Pettinato, M., Tuomainen, O., Granlund, S., & Hazan, V. (2016). Vowel space area in later childhood and adolescence. Effects of age, sex and ease of communication. *Journal of Phonetics, 54,* 1–14. https://doi.org/10.1016/j.wocn.2015.07.002

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing [Computer software]. https://www.R-project.org/

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. https://CRAN.R-project.org/package=psych

Rohrmann, T. (2017). Kindheit. Entwicklung und Sozialisation im Blick der Geschlechterforschung [Childhood. Development and socialization in the focus of gender research]. In B. Kortendiek, B. Riegraf, & K. Sabisch (Eds.), *Handbuch Interdisziplinäre Geschlechterforschung* (pp. 1–11). Springer.

Runge, T. E., Frey, D., Gollwitzer, P. M., Helmreich, R. L., & Spence, J. T. (1981). Masculine (instrumental) and feminine (expressive) traits. *Journal of Cross-Cultural Psychology, 12*(2), 142–162. https://doi.org/10.1177/0022022181122002

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Proc. of the ICPhS* (pp. 607–610).

Sederholm, E. (1995). Prevalence of hoarseness in ten-year-old children. *Scandinavian Journal of Logopedics and Phoniatrics, 20*(4), 165–173. https://doi.org/10.3109/14015439509098744

Serbin, L. A., Poulin-Dubois, D., Colburne, K. A., Sen, M. G., & Eichstedt, J. A. (2001). Gender stereotyping in infancy: Visual preferences for and knowledge of gender-stereotyped toys in the second year. *International Journal of Behavioral Development, 25*(1), 7–15. https://doi.org/10.1080/01650250042000078

Shadle, C. H., Koenig, L. L. & Preston, J. L. (2011). Acoustic characterization of /s/ spectra of adolescents: Moving beyond moments. In *Proceedings of Meetings on Acoustics 12*.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass, 3*(2), 621–640. https://doi.org/10.1111/j.1749-818X.2009.00125.x

Simpson, A. P., Funk, R. & Palmer, F. (2017). Perceptual and acoustic correlates of gender in the prepubertal voice. In *Interspeech 2017* (pp. 914–918).

Simpson, A. P. & Weirich, M. (2020). Phonetic correlates of sex, gender and sexual orientation. In *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.749

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Todd, B. K., Barry, J. A., & Thommessen, S. A. O. (2017). Preferences for 'gender-typed' toys in boys and girls aged 9 to 32 months. *Infant and Child Development, 26*(3), Article e1986. https://doi.org/10.1002/icd.1986

Walker, J. F., Archibald, L. M. D., Cherniak, S. R., & Fish, V. G. (1992). Articulation rate in 3- and 5-year-old children. *Journal of Speech, Language and Hearing Research, 35*(1), 4–13. https://doi.org/10.1044/jshr.3501.04

Weirich, M., & Simpson, A. P. (2018). Gender identity is indexed and perceived in speech. *PLOS ONE, 13*(12), Article e0209226. https://doi.org/10.1371/journal.pone.0209226

Whiteside, S. P. (1998). Identification of a speaker's sex: A study of vowels. *Perceptual and Motor Skills, 86*(2), 579–584. https://doi.org/10.2466/pms.1998.86.2.579

Whiteside, S. P. (2001). Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *The Journal of the Acoustical Society of America, 110*(1), 464–478. https://doi.org/10.1121/1.1379087

Whiteside, S. P., & Hodgson, C. (1999). Acoustic characteristics in 6-10-year-old children's voices: Some preliminary findings. *Logopedics Phoniatrics Vocology, 24*(1), 6–13. https://doi.org/10.1080/140154399434508

Whiteside, S. P., & Hodgson, C. (2000). Speech patterns of children and adults elicited via a picture-naming task: An acoustic study. *Speech Communication, 32*(4), 267–285. https://doi.org/10.1016/S0167-6393(00)00013-3

Yairi, E., Currin, L. H., Bulian, N., & Yairi, J. (1974). Incidence of hoarseness in school children over a 1 year period. *Journal of Communication Disorders, 7*(4), 321–328. https://doi.org/10.1016/0021-9924(74)90014-8

Yeung, S. P., & Wong, W. I. (2018). Gender labels on gender-neutral colors: Do they affect children's color preferences and play performance? *Sex Roles, 79*(5–6), 260–272. https://doi.org/10.1007/s11199-017-0875-3

Zosuls, K. M., Martin, C. L., Ruble, D. N., Miller, C. F., Gaertner, B. M., England, D. E., & Hill, A. P. (2011). 'It's not that we hate you': Understanding children's gender attitudes and expectancies about peer relationships. *British Journal of Developmental Psychology, 29*(2), 288–304. https://doi.org/10.1111/j.2044-835X.2010.02023.x

Appendix

Pictures for Audio Recordings

Figure A1. Picture for elicitation of disyllabic target nouns.

Figure A2. Three scenes for the elicitation of spontaneous speech: farmyard, living room decorated for Christmas, playground.