INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Listeners' Ability to Identify the Gender of Preadolescent Children in Different Linguistic Contexts

Shawn Nissen[1], *Sharalee Blunck*[1], *Anita Dromey*[1], *Christopher Dromey*[1]

[1]Department of Communication Disorders, Brigham Young University, Provo, Utah, USA

Shawn_nissen@byu.edu

Abstract

This study evaluated listeners' ability to identify the gender of preadolescent children from speech samples of varying length and linguistic context. The listeners were presented with a total of 190 speech samples in four different categories of linguistic context: segments, words, sentences, and discourse. The listeners were instructed to evaluate each speech sample and decide whether the speaker was a male or female and rate their level of confidence in their decision. Results showed listeners identified the gender of the speakers with a high degree of accuracy, ranging from 86% to 95%. Significant differences in listener judgments were found across the four levels of linguistic context, with segments having the lowest accuracy (83%) and discourse the highest accuracy (99%). At the segmental level, the listeners' identification of each speaker's gender was greater for vowels than for fricatives, with both types of phoneme being identified at a rate well above chance. Significant differences in identification were found between the /s/ and /ʃ/ fricatives, but not between the four corner vowels. The perception of gender is likely multifactorial, with listeners possibly using phonetic, prosodic, or stylistic speech cues to determine a speaker's gender.

Index Terms: gender identification, genderlect, speech recognition, speech perception, fricative production, voice identification

1. Introduction

Listeners' ability to distinguish between male and female speakers with a relatively high degree of accuracy has been well established. However, the factors that enable listeners to perceptually identify a speaker's gender are not well understood, especially in younger speakers. Some researchers believe the mechanisms that facilitate speaker gender identification may be acoustic-phonetic differences present in relatively short segments of speech [1, 2, 3], while others suggest that male and female speakers differ in their use of prosodic features that extend over longer periods of speech [4, 5]. The aim of this study was to investigate listeners' ability to identify the gender of young speakers from speech samples of varying length.

Perceptual experiments have indicated that listeners can identify the gender of children with a relatively high degree of accuracy based on their speech. A study by Bennet and Weinberg [1] found that adult listeners were able to correctly identify the gender of 6- and 7-year-old children 68% of the time. Additional studies have reported that adult listeners can correctly identify the gender of speakers as young as 4 years of age with as high as 81% accuracy [2, 6].

Many of the perceptual differences between the speech of adult men and women can be explained by differences in vocal tract anatomy, such as differences in vocal tract and vocal fold length [7, 8, 9]. Although researchers have hypothesized a small degree of sex-related differences in the vocal anatomy of younger children [10, 11], most evidence indicates that substantial sexual dimorphism of vocal tract structures in speakers typically begins to appear around 12 years of age or at the beginning stages of puberty [9, 12]. Thus, the slight differences in preadolescent boys and girls do not appear to fully account for the perceptual distinctions found in their speech communication.

In view of these anatomical data, it could be reasoned that gender-related perceptual differences in the speech of younger children are in part the result of learned sociophonetic factors. Sachs et al. [6] suggested that perceptual differences in children's speech may be due to formant frequency and intonation patterns that adhere to culturally determined male and female archetypes. Fitch and Giedd [12] concluded that because they found no differences in vocal tract length among preadolescent males and females, acoustic differentiation in speech patterns is likely the result of behavioral rather than anatomical factors. In particular, young males may protrude their lips more when speaking, thus behaviorally creating a longer vocal tract, which results in lower formant frequencies. Sachs et al. [6] also suggested that perceptual distinctions may be caused by variation in lip configuration during articulation, which results in allophonic differences in phonemes. Bennet [13] indicated that male participants used greater lip rounding, smaller jaw openings, and/or a lower larynx position than female participants.

In terms of production, studies have found notable differences in formant frequencies among preadolescent males and females [7, 13, 14, 15, 16]. Research has also reported gender differences in voice onset time patterns in the speech of preadolescent males and females, associated with age and development, the extent of which varied depending upon the type of plosive examined [3]. Fox and Nissen [17] evaluated voiceless fricatives in children 6 to 14 years of age and found that preadolescent children exhibited gender-specific differences in fricative articulation, though the acoustic differences were reduced or absent in the 6- to 7-year-old children. In another study the researchers examined the acoustic patterns of voiceless stop consonants in children 3 to 5 years of age. Results showed significant gender differences for the spectral measures of slope, mean, and skewness in the 5-year-old children [18]. Despite findings that report gender-related speech production differences in children [3, 6, 17, 19], the perceptual speech cues used by listeners to identify the gender of younger speakers remain unclear, especially how listeners might use prosody in such decisions [6].

http://dx.doi.org/10.21437/Interspeech.2019-1865

One way to provide some insight as to the origin of the perceptual differences in the speech of children is to present various levels of linguistic context to listeners (i.e., sound segments, words, sentences, and discourse). It is hypothesized that if listeners accurately identify a child's gender from their speech much of the time at the segment level, it can be suggested that formant frequencies are the basis for such judgments. On the other hand, if listeners correctly identify the speakers' gender more often at the sentence or discourse level, it can be suggested that suprasegmental aspects of speech, such as intonation, pause, stress, rate, or rhythm, may play an important role in speaker gender identification.

Historically, only a limited number of studies have evaluated gender-related differences in children's speech across more than a single level of linguistic context [6, 20]. Therefore, in this study children's speech will be evaluated across four different contexts of varying length (i.e., sound segment, word, sentence, and discourse). Three research questions will be addressed in this study. First, can listeners accurately identify the gender of a speaker from a speech sample? Second, does their accuracy increase for longer samples? Third, does a listener's ability to identify the speaker's gender from a single sound segment change depending on the phoneme being spoken?

2. Method

2.1. Speech Recordings

Speech samples were collected from 10 children (5 male and 5 female) between 8:0 and 9:11 years of age (M = 9:2) all of whom were monolingual speakers of American English [21]. The children had no diagnosed history of speech, language, or hearing problems at the time the speech samples were collected. If a child exhibited perceptual signs of poor vocal health (e.g., cold, laryngitis, vocal hoarseness, etc.), the recording sessions were postponed until a later date.

2.2. Listeners

The child speech recordings were evaluated by a group of 20 adult listeners (12 female and 8 male), who were native speakers of American English and reported no previous history of speech or language disorder. The adult listeners exhibited pure-tone air-conduction hearing thresholds of 25 dB HL at octave frequencies from 500 to 8000 Hz.

2.3. Stimuli

Each listener was presented with speech stimuli drawn from the child recordings. The stimuli included 60 isolated phoneme segments, 60 single words, 60 sentences, and 10 short segments of discourse. The phoneme stimuli consisted of two sibilant voiceless fricatives (/s/ and /ʃ/) and four corner vowels (/i/, /æ/, /u/, /ɑ/) extracted from CVC citation words produced by each of the 10 child speakers, elicited through picture description. Vowel segments were selected as stimuli because previous studies have shown that preadolescent male children exhibit lower vowel formant frequencies than female children [13, 14, 16]. Fricative sounds were selected as stimuli because studies have found significant gender-related differences in young children's fricatives in terms of spectral slope, mean, and skewness [18, 19]. The word stimuli included three monosyllabic (*horse, pool, fence)* and three bisyllabic words (*jumping, pizza, swimming*) produced by each child. Similar to the sound segments, the word stimuli were elicited through picture identification. The sentence stimuli were created by instructing each child to read the following sentences:

The boy is swimming in the pool.
The horse is jumping the fence.
The girl is mowing the lawn.
The boy is eating the pizza.
The lady is picking the flower.
The boy is baking the cookies.

The segments of discourse were approximately 1-minute in length, extracted from conversations with each child about the academic subjects they like or dislike in school and the reasons why.

All the stimuli were segmented at a zero-crossing, ramped over the initial and ending 10 ms. of the waveform, high-pass filtered, and normalized for relative intensity using the average RMS of each token.

2.4. Procedures

The adults were asked to listen to each stimulus item once and decide if the speaker identified themselves as male or female gender. In addition, the listeners were instructed to rate their level of confidence on a scoring sheet for each perceptual decision using a scale from 1-10 (not confident - completely confident). In order to test intra-rater reliability, 20% of the samples were randomly selected and replayed to the listeners a second time. The overall intra-rater reliability was found to be 89.90%, with a correlation of $r = .80$, $p < .001$.

The stimulus items were randomly presented to the listeners according to the level of linguistic context (e.g., sound segments, words, etc.). The order of presentation for each stimulus group and individual items within each group was randomized. The stimuli were presented to the listeners via Sennheiser HD 650 headphones while seated in a single-walled sound booth meeting ANSI S3.1 standards with ears covered [22]. Each participant listened to a practice token before rating the experimental stimuli. The listeners self-selected the intensity level of the presented stimuli, with a starting level of approximately 60 dB HL, which could be readjusted during testing if needed. The testing took place during one forty-minute session.

To examine the possible impact of voice pitch the mean speaking fundamental frequency (F0) of each child participant was calculated from the word, sentence, and discourse stimuli using Praat Analysis Software [23].

2.5. Statistical analyses

To meet the assumptions of independence, normality, and homoscedasticity, the percentage data were transformed using an arcsine transformation prior to using a repeated measures analyses of variance (ANOVA) to determine significant differences in the listeners' ability to accurately perceive the speakers' gender as a function of the amount of linguistic context and the segment type. It is important to note that although the arcsine transformation has historically been a standard statistical procedure in many fields, the efficacy of the transformation is continuing to be evaluated [24, 25]

3. Results

Across all levels of context, the overall accuracy with which listeners identified stimuli was found to be 90.66%. Individual listener percentages ranged from 86.32% to 94.74% accuracy, with an average confidence level of 7.26.

A repeated measures ANOVA, $F(3, 57) = 95.321$, $p < .001$, $\eta^2 = .83$, demonstrated statistically significant differences between the four levels of context. Pairwise comparisons showed statistically significant differences between all four levels of linguistic context. The comparison between the sentence and discourse levels of context was significant at $p < .02$, while pairwise comparisons between all other levels of context were significant at $p < .001$. As shown in Figure 1, as the amount of linguistic context increased, the accuracy of the listeners' identifications also increased. Percentages ranged from 83.42% at the segmental level to 99.00% at the discourse level. The accuracy with which listeners identified the gender of individual speakers ranged from 82.37% for speaker 7 to 98.68% for speaker 6, with a *SD* of 5.49%.

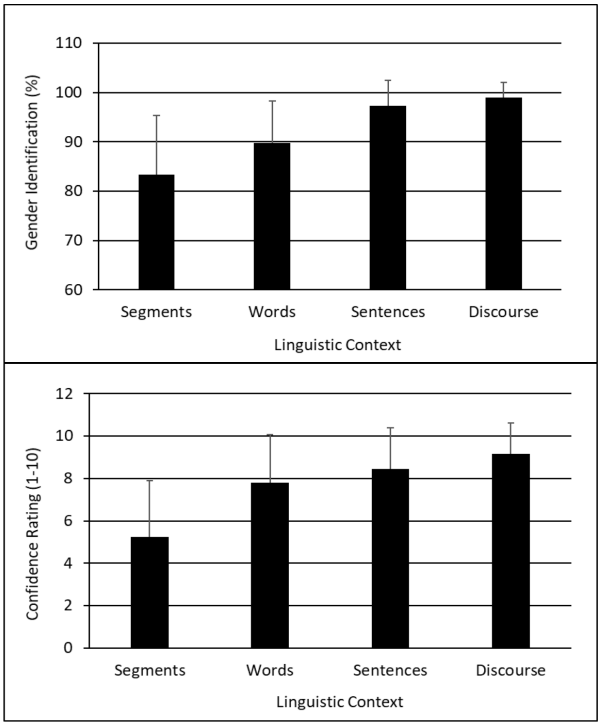

Figure 1: *Gender Identification Accuracy and Confidence Ratings across Linguistic Context.*

The response data were analyzed to determine whether listeners identified speakers of their own gender with a greater or lesser degree of accuracy. Female listeners identified female speakers with a significantly lower degree of accuracy than male speakers at a rate of 88.60% and 92.54% respectively, $F(1, 118) = 10.899$, $p < .01$, $\eta^2 = .09$. Male listeners identified both genders with the same degree of accuracy (90.79%).

At the segmental level of linguistic context, there was a higher degree of accuracy for vowels (87.13%) than for fricatives (76.00%), as shown in Figure 2. Although listeners' accuracy in identification of fricatives was well above chance, the listeners exhibited lower confidence ratings for fricatives (3.74) than for vowels (5.98), or stimuli at any other level of linguistic context. A repeated measures ANOVA was conducted to evaluate differences in speaker identification across individual vowel and fricative sound segments. No significant differences between the four vowels were found. However, the ANOVA did indicate statistically significant differences between the listeners' identification of speaker

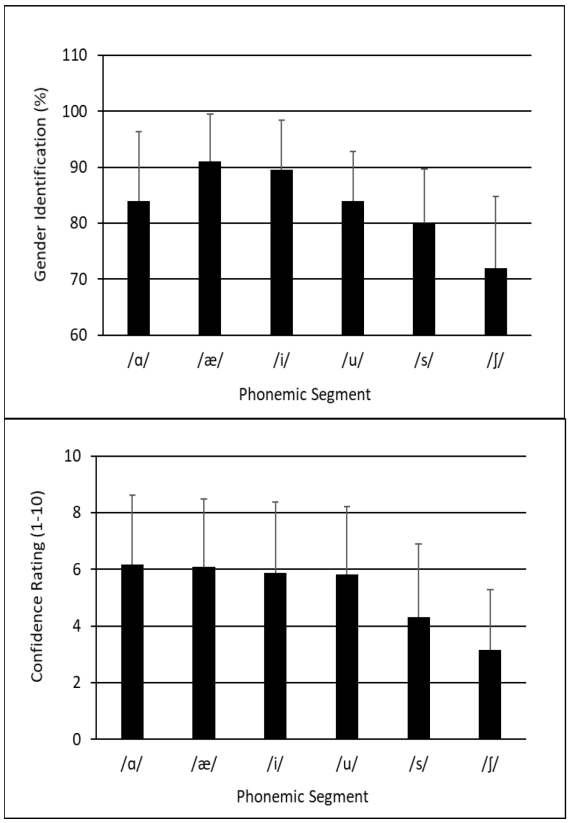

Figure 2: *Gender Identification Accuracy and Confidence Ratings across Phoneme Segments.*

gender when presented with /s/ and /ʃ/ segments, $F(1, 19) = 6.057$, $p < .02$, $\eta^2 = .24$.

The mean speaking F0 was found to be 241.89 Hz (*SD* = 20.87) for females and 231.30 Hz (*SD* = 10.14) for males. A correlation between speaker F0 and listener accuracy in identifying the speaker gender was found to be non-significant, $r = -.25$. Refer to Table 1 for a detailed listing of the mean F0 values and rate of identification percentages.

Table 1: *Speaker F0 and Listener Gender Identification Accuracy*

Speaker	Gender	Mean F0	Listener Accuracy %	Listener Accuracy *SD*
1	Female	224.48	85.26	8.97
2	Female	251.62	95.79	5.30
3	Female	253.32	90.53	5.29
4	Female	215.45	88.42	7.75
5	Female	264.58	87.37	8.61
6	Male	226.04	98.68	2.34
7	Male	247.08	82.37	9.24
8	Male	229.08	91.58	5.51
9	Male	220.24	98.42	2.47
10	Male	234.05	88.16	10.51

4. Discussion

The first aim of this study was to determine whether listeners could identify the gender of the child participants from a speech sample with relatively high accuracy. Similar to previous findings [1, 2, 6], the adult listeners were able to correctly identify the gender of children a majority of the time when listening to their speech.

The second aim was to examine whether listeners were more accurate in identifying the gender of the speaker when listening to a sample with more linguistic context. The data showed that as the amount of linguistic context increased, the accuracy of the listeners' identifications also increased, with percentages ranging from 83.42% when listening to sound segments to 99.00% when presented with a 1-minute speech sample from each speaker. Although speaker gender was identified from segments with the lowest percentage of accuracy among the four levels of linguistic context, the identification rate was still well above chance.

Listeners' higher accuracy for words compared with segments may be attributable to words containing multiple instances of phoneme segments. Differences may also be the result of additional suprasegmental information that can be heard when listening to words as opposed to sound segments in isolation, such as stress patterns and coarticulation. Some authors have hypothesized that children learn culturally determined suprasegmental patterns that are appropriate or archetypal for each gender [6, 12].

The suprasegmental aspects of speech that are present in sentences, such as speaking rate, rhythm, and sentential stress patterns, may underlie the higher identification scores for sentences. These results are consistent with data reported by Ferrand and Bloom [26], who found that young male speakers between the ages of 7 and 8 years began to restrict their intonational ranges while females did not. Although not documented in depth with child speakers, gender-related differences in speaking rate have also been found [5, 27].

While the difference was less pronounced, listeners identified the gender of the speakers with greater accuracy when presented with 1-minute sections of discourse than when hearing sentence length stimuli. Although the topics discussed in the conversational samples in the present study were meant to be gender-neutral, it is possible that when the segmental or suprasegmental cues were ambiguous regarding speaker gender, listeners used differences in lexicon [6].

The third aim of this study was to evaluate the extent to which listeners' abilities to identify a speaker's gender from a single sound segment might be related to the specific phoneme stimulus. Listeners' accuracy varied depending on the phoneme being spoken, with higher accuracy for vowels than for fricatives. These results are similar to those reported by Lee et al. [20], who also found that speaker gender was identified more accurately for vowels than fricatives. The accuracy with which vowels were identified provides support for the perceptual salience of formant frequency and F0 patterns in determining the gender of a speaker. As noted in the introduction, authors have attributed differences in formant frequencies between genders to anatomical factors, behavioral factors, or a combination of both.

While some researchers have found that F0 does not differ as a function of gender until children reach puberty [14], others have found differences in F0 beginning around age 8 [15]. It is notable that the female speaker with the highest mean F0 was not the speaker most often identified as female, although the two male speakers with the lowest mean F0s were the most frequently identified males. Such results suggest that while F0 may play a role in gender identification, it is not the only factor listeners rely on when determining gender. As mentioned in the results section, the correlation between speaker F0 and listener accuracy in identifying the speaker gender was non-significant, $r = -.25$. This lack of a significant correlation indicates that listeners were likely using cues other than pitch to determine the speaker gender.

It is noteworthy that the fricatives /s/ and /ʃ/ allowed gender identification at a rate much greater than chance, even though the fricatives had extremely short durations and contained no voicing. A study by Schwartz (1968) found that women tended to have higher frequency spectra when producing /s/ and /ʃ/ than men [28]. Another study found gender-specific differences in the acoustic properties of fricatives when analyzing spectral slices and spectral moments of adults [29]. Evidence that supports differences in fricative production in male and female children between 6 and 14 years of age was presented by Fox and Nissen [17]. The authors concluded that, because dimorphism of the vocal tract primarily occurs at peripubertal and postpubertal stages of development, a portion of the gender-linked differences found in children may be attributed to learned or behavioral factors, emerging as early as 6-7 years of age. A follow-up study [19] found evidence that gender differences in fricative production started at around 5 years of age. As noted in the results, there were no statistically significant differences in accuracy of gender identification between the four vowels.

Future research in this area might involve larger numbers of both speakers and listeners across a more diverse age range. In addition, it would be of interest to examine the impact that individual acoustic parameters (e.g., overall duration, spectral mean, intensity) of a phoneme might have on the perception of speaker gender, possibly through the manipulation of synthetic speech stimuli. This study correlated the perceptual judgments of the listeners with the acoustic measure of F0; however, it would also be of interest to more fully examine other perception-production links by conducting a detailed spectral and acoustic analysis of the speakers' productions.

5. Conclusions

The high accuracy with which listeners identified the gender of the young speakers, who according to previous anatomical research likely have limited dimorphism in their vocal tract structures, may indicate the presence of learned, gender-specific characteristics in the speech of the child participants. A combination of acoustic distinctions, possibly resulting from both anatomical and behavioral differences, as well as suprasegmental or prosodic aspects of speech, likely provide perceptual cues which enable listeners to identify gender in children. Despite the limitations of the current study, these data may provide additional insight into the ability of listeners to identify the gender of a young speaker and the associated phonetic, prosodic, or stylistic speech factors.

6. Acknowledgements

We would like to thank the McKay School of Education at Brigham Young University for their financial support.

7. References

[1] S. Bennet and B. Weinberg, "Sexual characteristics of preadolescent children's voices," *J. Acoust. Soc. Am.*, vol. 65, pp. 179–189, 1979.

[2] T. L. Perry, R. N. Ohde, and D. H. Ashmead, "The acoustic bases for gender identification from children's voices," *J. Acoust. Soc. Am.*, vol. 109, pp. 2988–2998, 2001.

[3] S. P. Whiteside and J. Marshall, "Development trends in voice onset time: Some evidence for sex differences," *Phonetica,* vol. 58, pp. 196-210, 2001.

[4] M. L. Andrews and C. P. Schmidt, "Gender presentation: perceptual and acoustical analyses of voice," *Journal of Voice*, vol. 11, pp. 307-313, 1997.

[5] M. Fitzsimons, N. Sheahan, and H. Staunton, "Gender and the integration of acoustic dimensions of prosody: Implications for clinical studies," *Brain and Language*, vol. 78, pp. 94-108, 2001.

[6] J. Sachs, P. Lieberman, and D. Erickson, "Anatomical and cultural determinants of male and female speech," in *Language Attitudes*, R. W. Shuy and R. W. Fasold, Eds. Washington DC: Georgetown University Press, 1973, pp 74-83.

[7] A. J. Seikel, D. W. King, and D. G. Drumright, *Anatomy and Physiology for Speech, Language, and Hearing.* New York: Thomson Delmar Learning, 2005.

[8] I. R. Titze, "Physiologic and acoustic differences between male and female voices," *J. Acoust. Soc. Am.*, vol. 85, pp. 1699-1707, 1989.

[9] H. K. Vorperian, S. B. Wang, M. K. Chung, M. E. Schimek, R. B. Durtschi, R. D. Kent, A. J. Ziegert, and L. R. Gentry, "Anatomic development of the oral and pharyngeal portions of the vocal tract: an imaging study," *J. Acoust. Soc. Am.*, vol. 125, pp. 1666-1678, 2009.

[10] C. Hasek, S. Singh, and T. Murry, "Acoustic attributes of preadolescent voices," *J. Acoust. Soc. Am.*, vol. 68, pp. 1262-1265, 1980.

[11] H. K. Vorperian, R. D. Kent, M. J. Lindstrom, C. M. Kalina, L. R. Gentry, and B. S. Yandell, "Development of vocal tract length during early childhood: A magnetic resonance imaging study," *J. Acoust. Soc. Am.*, vol. 117, pp. 338-350, 2005.

[12] T. W. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 106, pp. 1511-1522, 1999.

[13] S. Bennet, "Vowel formant frequency characteristics of preadolescent males and females," *J. Acoust. Soc. Am.*, vol. 69, pp. 231-238, 1981.

[14] P. A. Busby and G. L. Plant, "Formant frequency values of vowels produced by preadolescent boys and girls," *J. Acoust. Soc. Am.*, vol. 97, pp. 2603-2606, 1995.

[15] S. P. Whiteside and C. Hodgson, "Some acoustic characteristics in the voices of 6- to 10-year-old children and adults: A comparative sex and developmental perspective," *Logop. Phoniatr. Voco.*, vol. 25, pp. 122-132, 2000a.

[16] S. P. Whiteside and C. Hodgson, "Speech patterns of children and adults elicited via a picture-naming task: An acoustic study," *Speech Communication*, vol. 32, pp. 267-285, 2000b.

[17] R. A. Fox and S. L. Nissen, "Sex-related acoustic changes in voiceless English fricatives," *J. Speech Lang. Hear. Res.*, vol. 48, pp. 753-765, 2005.

[18] S. L. Nissen and R. A. Fox, "Acoustic and spectral patterns in young children's stop consonant productions," *J. Acoust. Soc. Am.*, vol. 126, pp. 1369-1378, 2009.

[19] S. L. Nissen and R. A. Fox, "Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective," *J. Acoust. Soc. Am.*, vol. 118, pp. 2570-2578, 2005.

[20] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Am.*, vol. 105, pp. 1455-1468, 1999.

[21] A. Dromey, "An acoustic and perceptual investigation of contrastive stress in children," M.S. Thesis, Dept. Communication Disorders, Brigham Young University, Provo, UT, 2010.

[22] *Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms*, American National Standards Institute S3.1, 1999.

[23] P. Boersma and D. Weenink, *Praat Signal Processing* [Computer Software]. Amsterdam, NL: University of Amsterdam Institute of Phonetic Sciences, 2009.

[24] S. Lo and S. Andrews, "To transform or not to transform: using generalized linear mixed models to analyse reaction time data," *Frontiers in Psychology*, vol. 6, pp. 1-16, 2015.

[25] D. I. Wharton and F. K. C. Hui, "The arcsine is asinine: the analysis of proportions in ecology," *Ecology*, vol. 92, pp. 3-10, 2011.

[26] C. T. Ferrand and R. L. Bloom, "Gender differences in children's intonational patterns," *Journal of Voice*, vol. 10, pp. 284-291, 1996.

[27] S. P. Whiteside, "Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences," *J. Int. Phon. Assoc.*, vol. 26, pp. 23-40, 1996.

[28] M. F. Schwartz, "Identification of speaker sex from isolated, voiceless fricatives," *J. Acoust. Soc. Am.*, vol. 43, pp. 1178–1179, 1968.

[29] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *J. Acoust. Soc. Am.*, vol. 108, pp. 1252–1263, 2000.