

**CREATION, PERCEPTION, AND USE OF GENDER EXPANSIVE  
SYNTHETIC VOICES**

by

Maxwell Hope

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Linguistics

Summer 2024

© 2024 Maxwell Hope  
All Rights Reserved

**CREATION, PERCEPTION, AND USE OF GENDER EXPANSIVE  
SYNTHETIC VOICES**

by

Maxwell Hope

Approved: \_\_\_\_\_  
Robin Andreasen, Ph.D.  
Chair of the Department of Linguistics and Cognitive Science

Approved: \_\_\_\_\_  
Debra Hess Norris  
Interim Dean of the College of Arts and Sciences

Approved: \_\_\_\_\_  
Louis F. Rossi, Ph.D.  
Vice Provost for Graduate and Professional Education and  
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Katherine Verdolini Abbott, Ph.D.  
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Kathryn Franich, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Sayako Earle, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Benjamin Munson, Ph.D.  
Member of dissertation committee

## ACKNOWLEDGMENTS

A dissertation cannot be done without a supportive community and network of people who believe in you even when you stop believing in yourself. The people who kept believing no matter what are undoubtedly my partners, GL and Xan (who stayed close through my self-doubt and pain), my good friends – too numerous to name (which is a wonderful problem!), and my committee members: Drs. Verdolini Abbott, Franich, Earle and Mx. Munson. Without willing and enthusiastic participants, there would be nothing to write about, so I thank, from the bottom of my heart, the myriad members of the gender expansive community who took part in this work. Jason Lilley, my academic “partner-in-crime”, and all of those at the Nemours’ Center for Pediatric Auditory and Speech Sciences, lent me time, resources, and camaraderie necessary to complete my experiments. I was extremely fortunate to have a committed, disciplined, and curious undergraduate research assistant, Vince Yonek. While I battled with a chronic shoulder injury, Marc Kealhofer, Charlotte Ward, and Thea Flurry assisted in the transcription and formatting of this dissertation. For the financial support I received from the University of Delaware during this time, I send my heartfelt appreciation while recognizing every single one of my colleagues deserved the same (and more). A twelve-month contract on a livable wage is key to both surviving and thriving in a PhD program. Finally, I thank my family, especially my parents and my ancestors; I cannot repay the debt for being alive. Ah! Life! Suffering. Joy. To all beings in this world, thank you!

## TABLE OF CONTENTS

LIST OF TABLES .....	ix
LIST OF FIGURES .....	xi
ABSTRACT .....	xiv

### Chapter

1	INTRODUCTION .....	1
1.1	Overview: Creation, Perception, and Use of Gender Expansive Synthetic Voices .....	1
1.2	Chapter 1: Gender Encoding in Speech .....	2
1.3	Chapters 2 & 3: Gender Perception in Speech .....	5
1.4	Chapter 4: Synthetic Voices and Gender Representation .....	6
1.5	Conclusion .....	8
2	GENDER EXPANSIVE SPEECH CHARACTERISTICS .....	10
2.1	Gender in Speech .....	10
2.1.1	Fundamental Frequency ( $f_0$ ) and Voice Quality .....	11
2.1.2	COG and Peak Frequency of [s] in English .....	12
2.1.3	Vocal Tract and Formant Frequencies .....	13
2.2	Questions and Hypotheses .....	14
2.3	Methods and Materials .....	15
2.3.1	Participants .....	15
2.3.2	Data Collection .....	16
2.3.3	Acoustic Analyses .....	18
2.3.4	Statistical Analyses .....	18
2.4	Results .....	19
2.4.1	Whole Group Summary .....	19
2.4.2	Categorical Gender Groups .....	20
2.4.3	Gradient Gender Results .....	23

2.4.3.1	Correlations .....	25
2.4.3.2	Effects of Gender on <i>f1</i> .....	26
2.4.3.3	Effects of Gender on <i>f2</i> .....	27
2.4.3.4	Effects of Gender on <i>f3</i> .....	29
2.5	Discussion.....	30
2.6	Conclusion.....	34
3	GENDER PERCEPTION IN GENDER EXPANSIVE SYNTHETIC VOICES.....	35
3.1	Introduction .....	37
3.1.1	Fundamental frequency .....	37
3.1.2	Formant Frequencies .....	39
3.1.3	Sibilants /s/ and /ʃ/ in English.....	42
3.1.4	Listener characteristics .....	43
3.1.5	Synthetic speech and gender perception.....	45
3.1.6	Questions and hypotheses.....	49
3.2	Methods and Materials .....	50
3.2.1	Synthetic Voice Creation.....	50
3.2.2	Sentence/Paragraph generation .....	51
3.2.3	Participants .....	53
3.2.4	Procedures .....	53
3.2.5	Statistical analyses.....	54
3.3	Results .....	55
3.3.1	Differences between voices .....	55
3.3.2	Gradient gender perception of the voices .....	59
3.4	Discussion and Conclusion.....	61
4	CATEGORIZATION OF THE GENDER EXPANSIVE SYNTHETIC SIBILANTS: DECOUPLING OF [S] PERCEPTION AND GENDER .....	63
4.1	Introduction .....	63
4.1.1	Gender and sibilant production in English .....	67
4.1.2	Gender and sibilant perception in English.....	70
4.1.3	Listener characteristics affect sibilant perception .....	72
4.1.4	Questions and Hypotheses.....	76

4.2	Methods and Materials .....	78
4.2.1	Stimuli Creation.....	78
4.2.1.1	Synthetic Voice Creation.....	78
4.2.1.2	“Sack” to “Shack” Continuum Creation.....	78
4.2.2	Participants .....	80
4.2.3	Procedures .....	81
4.2.4	Statistical analyses.....	82
4.3	Results .....	83
4.4	Discussion & Conclusion .....	86
5	NONBINARY VOICE GENDER ENCODING IN SYNTHETIC VOICE ....	88
5.1	Introduction .....	88
5.1.1	Synthetic voice and gender.....	91
5.1.2	Nonbinary identity and synthetic voice.....	93
5.1.3	Questions and hypotheses.....	95
5.2	Methods and Materials .....	96
5.2.1	Recruitment .....	96
5.2.2	Synthetic voice construction.....	96
5.2.3	Voice audition .....	97
5.2.4	Acoustic profiles of the top three voices .....	99
5.2.5	Daily journal entries .....	100
5.2.6	Post-trial Questionnaire .....	101
5.2.7	Analyses .....	102
5.3	Results .....	102
5.3.1	Overview of journal entries .....	102
5.3.2	Survey results .....	106
5.3.3	Qualitative Data Results .....	107
5.3.3.1	Feelings and sentiments.....	111
5.3.3.2	Acoustic cues.....	112
5.3.3.3	Social Interactions .....	112
5.3.3.4	Technology .....	113
5.4	Discussion.....	114
5.5	Conclusions .....	115

6	CONCLUSION .....	117
	REFERENCES .....	120
Appendix		
A	PERMISSIONS .....	130
B	IRB/HUMAN SUBJECTS APPROVAL .....	131
C	SUPPLEMENTAL INFORMATION .....	132

## LIST OF TABLES

Table 2.1 Descriptive statistics for acoustic measurements for the whole group of gender expansive talkers .....	19
Table 2.2 Descriptive acoustic statistics for each gender group.....	21
Table 2.3 Correlations per formant per vowel. Values represent Pearson’s $r$ with $r > 0.50$ bolded. ....	25
Table 3.1 Fundamental frequency ( $f_0$ ) averages and standard deviations for each of the sentences and passages for each synthetic voice.....	51
Table 3.2 Average vowel space dispersion (VSD) for the four voices calculated by averaging the Euclidian distance of vowels [i],[u],[a],[o],[e] from all the stimuli used for that voice to center of the vowel space.....	52
Table 4.1 Fundamental frequency ( $f_0$ ) averages and standard deviations for each of the sentences and passages for each synthetic voice.....	79
Table 4.2 Sibilant spectrum measures of intensity in decibels (dB) and spectral mean in Hertz (Hz) for the ALL sibilant spectrum from step 1, which contained only the [s] signal to step 9 which contained none of the [s] e.g. contained only [ʃ].....	80
Table 4.3 Results of the logistic regression, including coefficients, standard errors (std. error), $z$ -values and $p$ -values (* indicates that the $p$ -value was < 0.1; *** indicates that the $p$ -value was < 0.001). ....	86
Table 5.1 Results of the demographic questionnaire.....	97
Table 5.2 Results from the auditioning process showing the voice, its prosody model, its acoustic model, and the average affirmation score.....	99
Table 5.3 Acoustic characteristics of Voice1, Voice2 and Voice3. ....	100
Table 5.4 TT’s <i>Feminine</i> , <i>Masculine</i> , and <i>Other gender perception</i> of Voice1, Voice2 and Voice3. ....	101

Table 5.5 Overall themes and subthemes of the codes used and the total instances of each subtheme (n) used in all journal entries.....	108
Table 5.6 Percentage of instances of each code attributed to Voice1, Voice2, and Voice3. ....	109
Table 5.7 The subtheme “gender affirmation” broken down into its positive and negative instances with percent of those instances attributed to each voice. ....	110
Table S1. Correlation (Pearson’s $r$ ) results which differed in terms of statistical significance with and without the first author. (* represents that the $p$ -value of that correlation was statistically significant at the level of $p < 0.05$ ) .....	133
Table S2. Correlations per normalized formant per vowel, using the Johnson (2020) method. Bolded values represent Pearson’s $r$ correlations where $r > 0.50$ .....	134

## LIST OF FIGURES

<p>Figure 2.1 Strip chart of the six gender gradient gender variables. Horizontal black lines represent median values, with boxes indicating quartiles. MaleIdent represents gradient <i>Male gender identity</i>, FemaleIdent represents gradient <i>Female gender identity</i>, OtherIdent represents gradient <i>Other gender identity</i>, MascExpress represents gradient <i>Masculine gender expression</i>, FemExpress represents gradient <i>Feminine gender expression</i>, and OtherExpress represents gradient <i>Other gender expression</i>. .....</p>	17
<p>Figure 2.2 Unnormalized vowel spaces for each gender group. Nonbinary (dark green) represents the Nonbinary participants, Trans_Fem (dark orange) represents the Transfeminine participants, and Trans_Masc (dark blue) represents the Transmasculine participants. Each vowel shown represents an individual's average vowel production. ....</p>	22
<p>Figure 2.3 Boxplots showing the differences in A) <math>f_2</math> of [i], B) <math>f_2</math> of [e], and C) <math>f_2</math> of [o] for the three groups of Nonbinary, Transfeminine (Trans_Fem) and Transmasculine (Trans_Masc). ....</p>	24
<p>Figure 2.4 Linear regression of the relationship between <math>f_1</math> of [a] in hertz and gradient Other gender identity .....</p>	27
<p>Figure 2.5 Linear regressions of the relationships between <math>f_2</math> of [i] in hertz and gradient <i>Female gender identity</i> (A), <math>f_2</math> of [i] in hertz and gradient <i>Other gender identity</i> (B), <math>f_2</math> of [e] in hertz and gradient <i>Other gender identity</i> (C), <math>f_2</math> of [e] and gradient <i>Other gender expression</i> (D), <math>f_2</math> of [o] gradient <i>Male gender identity</i> (E), and <math>f_2</math> of [o] and gradient <i>Masculine gender expression</i> (F). ....</p>	28
<p>Figure 2.6 Linear regressions of the relationships between <math>f_3</math> of [i] in hertz and gradient <i>Other gender identity</i> (A), <math>f_3</math> of [u] in hertz and gradient <i>Female gender identity</i> (B), and <math>f_3</math> of [u] in hertz and gradient <i>Other gender identity</i> (C). ....</p>	30

Figure 3.1. Perceptual ratings collapsed across gender expansive and cisgender groups. The x-axis shows the four different synthetic voices (ALL representing the voice created from all gender expansive talkers, NonBin representing the voice created from the nonbinary talkers, TF representing the voice created from the transfeminine talkers, and TM representing the voice created from the transmasculine talkers). The y-axis shows the perceptual gender rating on a scale of 0-100. Colors indicate gender of perception with *Feminine gender* in red, *Masculine gender* in the green and *Other gender* in the blue. .... 56

Figure 3.2. Boxplots illustrating differences in perception for the ALL voice made from all 16 GE talkers (A), the NonBin voice made from the eight nonbinary talkers (B), the TF voice made from the four transfeminine talkers (C), and TM voice made from the four transmasculine talkers (D), between groups where red boxes indicate the cisgender listener group and blue boxes indicate the gender expansive listener group. .... 57

Figure 3.3 Linear regression demonstrating the relationship between *Other gender identity* of the listener (x-axis) and *Other gender perception* of the voices (y-axis). The red line indicates the voice made from all GE talkers, the green line represents the voice made from the eight nonbinary talkers, the blue line indicates the voice made from the four transfeminine talkers, and the purple line indicates the voice made from the four transmasculine talkers. .... 60

Figure 4.1 Sibilant categorization curves showing the proportion of [ʃ] responses by cisgender (Cis, left) and gender expansive (GE, right) listeners for three different synthetic voices (F: a female synthetic voice, M: a male synthetic voice, and N: a nonbinary synthetic voice). .... 66

Figure 4.2 Production of the word “so” by the author with the tongue more forward in the mouth compared to that in figure 4.3. The A figure shows the spectrogram of the fricative and vowel and the B figure shows the spectrum of the fricative production. .... 68

Figure 4.3 A production of the word “so” by the author with the tongue in a more retracted position compared to that in 4.2. The A figure shows the spectrogram of the fricative and vowel and the B figure shows the spectrum of the fricative productions. .... 68

Figure 4.4 Spectrogram and spectrum from the “female” vocal tract “see” token from Hope and Lilley (2023). .... 74

Figure 4.5 Spectrogram and spectrum from the “male” vocal tract “see” token from Hope and Lilley (2023) .....	75
Figure 4.6 Spectrogram and spectrum from the “nonbinary” vocal tract “see” token from Hope and Lilley (2023).....	76
Figure 4.7 Overall sibilant categorization between gender expansive and cisgender listeners; the x-axis represents the “step” from 1 to 9 with 1 representing 100% [s] in the acoustic signal and 9 representing 0% [s] in the acoustic signal. ....	83
Figure 4.8 Sibilant categorization between gender expansive and cisgender listeners for the four different vocal tracts; the x-axis represents the “step” from 1 to 9 with 1 representing 100% [s] in the acoustic signal and 9 representing 0% [s] in the acoustic signal. ....	84
Figure 5.1 Responses from each of the days the voices were used for the 3 quantitative questions. The bottom voice label represents the voice that was primarily used that day.....	107
Figure S1: Boxplots showing the center of gravity (COG_s) and peak frequency (Peak_s) of [s] for the three gender groups. ....	132

## **ABSTRACT**

Gender expansive (transgender and nonbinary) listeners are typically left out of the study of sociophonetic perception, and even phonetic studies more broadly, although there is an increasing number of studies investigating gender expansive speech production. Various studies have been conducted on the speech of men and women, from which conclusions have been drawn about gender cues in speech. However, these prior investigations are restrictive because they largely consider only cisheterosexual men and women. Research of gender in speech has also impacted the available types of gendered synthetic voices for those relying on them, and little has been documented about the perception of gender in synthetic voices used by Speech Generating Device (SGD) users. This dissertation provides a series of four studies for broadening our conception of gender production and perception in speech and extending this domain beyond the “biological” voice, that is, the attributes of voices that can be traced solely and directly to anatomical or physiological factors. First, the series evaluates the speech of 16 gender expansive participants evaluating  $f_0$ , formant frequencies, and spectral qualities of [s]. Second, the series evaluates perception of synthetic voices constructed from the same talkers. Finally, the series investigates, in a detailed case study, how a nonbinary SGD user encodes their gender using the aforementioned gender expansive synthetic voices. The results from this dissertation series will enrich the respective fields of sociophonetics, psycholinguistics, and speech technology.

## Chapter 1

### INTRODUCTION

#### **1.1 Overview: Creation, Perception, and Use of Gender Expansive Synthetic Voices**

Language enables individuals to convey aspects of their identity alongside regular linguistic messages. Speech is a crucial component of human communication, serving as a powerful tool for sharing social and personal information. Gender significantly influences language, allowing people to express and interpret identity in a variety of modalities, including both gestural and spoken modalities. This dissertation focuses on the latter. Understanding how gender manifests in speech and how it is interpreted is essential, as this process impacts communication, social interactions, and their interconnections. Advancements in technology have expanded the possibilities for synthetic voices in devices that generate speech and conversational assistants, presenting new opportunities to explore gender representation and expression. This chapter provides an overview of background information relevant to the dissertation along three lines. First, it covers the different acoustic signals that convey gender in American English, such as fundamental frequency, formants, and fine phonetic details of sibilant sounds. Second, it examines how people perceive and assign gender based on these acoustic signals. Third, it focuses on gender expansive synthetic voices used in devices like communication aids, and how these voices convey gender. Each of these studies serves as a vital building block in constructing the bridge toward inclusive speech technology. As our understanding of gender representation in speech

evolves, we move closer to the creation and widespread adoption of gender expansive synthetic voices, ultimately fostering a more diverse and equitable synthetic speech landscape.

## 1.2 Chapter 1: Gender Encoding in Speech

Prior research on gender in speech has primarily focused on cisgender binary gender differences notably in fundamental frequency ( $f_0$ ), whereby men typically exhibit lower average speaking  $f_0$  values (107 – 132 Hz), while women have higher average  $f_0$  values (196 – 224 Hz) (Davies & Goldberg, 2006). However, this binary perspective fails to account for nonbinary and transgender speech. Recent studies have begun to address this gap. Our preliminary research suggests that nonbinary individuals'  $f_0$  values typically fall between the male and female ranges, with an average around 144 Hz (Schmid & Bradley, 2019). Leann et al. (2022) conducted a study comparing voice quality and pitch variation in nonbinary and binary individuals. They found that nonbinary talkers exhibit a brighter voice quality, defined as higher harmonics-to-noise ratios (HNR), indicative of clearer and more resonant voices compared to cisgender talkers. Nonbinary individuals also display more within-subject variation in  $f_0$  compared to binary talkers (Schmid & Bradley, 2019; Leann et al., 2022), suggesting a broader range of vocal inflections and intonations for self-expression. These findings indicate that nonbinary individuals may exhibit unique vocal characteristics linked to their gender identity. Spectral qualities of sibilants are also known to be correlated with gender. For example, the spectral mean, also called the center of gravity (COG), of [s] differs between genders. Women typically produce [s] at higher frequencies than men (Jongman et al., 2000; Fuchs & Toda, 2010). In a group of transmasculine talkers with diverse gender and sexual identities, the study

found that different transmasculine identities interacted with other queer identities to influence the COG of [s], demonstrating a complex interplay of sociophonetic cues (Zimman, 2017). Finally, vocal tract length (VTL) and formant frequencies are associated with gender. Men tend to produce lower formants, while women produce higher formants. Across a population, average formant frequencies correlate with the length of the vocal tract, as measured objectively (Lammert & Narayanan, 2015). However, people can raise or lower all of the formants by lengthening or shortening their vocal tracts, through raising or lowering the larynx and protruding or retracting the lips. VTL and the shape of the vocal tract, influenced by tongue positioning, contribute to these formant variations. Speech-language pathologists have utilized vocal tract manipulation techniques to train transfeminine individuals to achieve a more feminine voice, involving tongue placement and lip positioning adjustments (Carew et al., 2005).

It is important to note that these acoustic variables exist on a continuum, allowing individuals to encode their gender identity along a gradient. While men may generally have thicker vocal folds and longer vocal tracts than women, as noted prior, both groups can manipulate  $f_0$  and VTL, indicating that gender expression in speech is multifaceted and nuanced (Schmid & Bradley, 2019; Weirich & Simpson, 2018). However, the literature predominantly focuses on cisgender individuals, necessitating further research to explore these dynamics in gender expansive talkers. The speech encoding of gender encompasses the acoustic signals and language patterns that influence how gender is perceived. Gender expression in voice is a multifaceted phenomenon that researchers have begun to explore to grasp how individuals express their gender identity through spoken words.

More recent literature on gender in speech increasingly investigates the production and perception of nonbinary speech. Nonbinary individuals, who do not conform to the conventional male-female binary, may display speech patterns that differ from binary, cisgender speech norms. For example, we previously found that nonbinary talkers did not consistently use spectral mean of [s] to encode gradient gender while they did use vowel acoustics to encode gradient gender characteristics (Hope et al., 2023). This line of research holds immense importance in capturing the wide range of gender expressions and presentations found in speech. By studying nonbinary speech, researchers can discover additional or alternative acoustic indicators that play a role in encoding gender, thus not only deepening understanding of the various ways people can encode gender in speech but also allowing for improved commercial voice recognition and synthesis technologies.

The first study in this dissertation series demonstrates how gender expansive talkers from the mid-Atlantic region of the United States encode multidimensional gender (i.e. independent components of masculine, feminine, and nonbinary gender) in speech. Understanding how gender is represented in speech has practical implications in various fields. Such understanding can impact the development of voice recognition and synthesis technologies, as well as interventions in speech therapy and strategies for social communication. For example, knowing about the acoustic signals that convey gender can help create synthetic voices that are more accurate and inclusive compared to synthetic voices that implicitly encode cisgender, binary speech patterns. In speech therapy, knowledge about how gender is expressed can assist clinicians in supporting transgender and nonbinary individuals in achieving their desired voice and

speech goals and in giving their gender expansive clients both knowledge and tools they can use in their day-to-day communication.

### **1.3 Chapters 2 & 3: Gender Perception in Speech**

Experiences, cultural expectations, and inherent biases shape the perception of gender in speech. Certainly, the way people hear and perceive gender is influenced by characteristics such as pitch, vowel sounds, and sibilants. As previously noted regarding voice perception, higher pitches are perceived as more feminine, while lower pitches are associated with masculinity (Leung et al., 2018). Formant frequencies also contribute to how one perceives gender whereby lower formants are attributed to masculinity and higher formants attributed to femininity (Leung et al., 2019). The intensity and spectral properties of the high-frequency consonant sounds known as sibilants also contain information related to gender that can influence gender perception (Strand & Johnson, 1999; Munson, 2011).

However, how one perceives a person's gender through speech goes beyond just the sounds they make. Social and contextual factors play a significant role in assigning gender to a talker. The cultural and societal expectations surrounding gendered speech patterns heavily influence interpretation. Cognitive biases come into play as one interprets speech based on preconceived notions and stereotypes about gender. For example, in our previous study, we found that cisgender listeners were much more likely to use the  $f_0$  cue to ascribe a gender to a talker compared to gender expansive listeners (Hope & Lilley, 2020).

The second and third studies in this dissertation series investigate (1) how gender perception of gender expansive synthetic voices is influenced by group membership (e.g. gender expansive versus cisgender listeners) and individual identity

(e.g. gradient gender of the listener), and (2) how group membership influences sibilant perception of gender expansive synthetic voices. Recognizing and understanding how one perceives gender and speech is essential due to the implications for a variety of fields including sociolinguistics, psychology, and speech technology. Studying gender perception in sociolinguistics offers insights into the evolving nature of language and how social constructs shape gender identities. In psychology, better understanding gender perception aids the comprehension of how one perceives and classifies others based on gender; better understanding gender perception offers insight into the mechanisms of characterizing and labeling the self and others. In speech technology, understanding gender perception is crucial for developing voice recognition systems, conversational assistants, and artificial voices that are accurate and inclusive with respect to gender.

#### **1.4 Chapter 4: Synthetic Voices and Gender Representation**

Current speech-generating devices are restrictive in terms of identity expression. Although custom voices exist via voice-banking, not everyone has the ability or the desire to construct a voice that sounds exactly like them. We conducted a feasibility study on the intricate relationship between speech, in the form of synthetic voices in speech-generating devices, and the construction of gender identities, underlining the notion that gender is not an innate trait but rather a socially constructed and performed facet of one's identity. This study stresses the pivotal role of language and speech in the continual generation of gender identities through everyday actions and interactions.

Moreover, the study highlights the prevalence of traditional (i.e. in the context of white American culture) binary gender norms that are closely tied to specific vocal

features, such as pitch. This association between vocal characteristics and binary gender categories -- male and female -- can lead to profound experiences of dysphoria and discomfort for individuals whose gender identity does not align with these conventional norms; put in another way, these binary gender categories that are inherent in synthetic voices prevent the gender expansive user from experiencing gender euphoria. Misgendering, societal stigma, and their associated mental health consequences can result from voices that do not conform to one's experience of gender. This risk of mental and social harm emphasizes the critical need for speech technologies and voice representation that can accommodate the diverse spectrum of gender identities and expressions beyond the binary framework.

Acknowledging that previous research on synthetic speech and gender is highly restricted to binary, cisnormative representations, we call for more inclusive and diverse synthetic voices that accurately represent the rich tapestry of gender identities present in society. This feasibility study advocates for a shift from a predominantly binary perspective to one that can effectively mirror the multifarious gender identities and expressions characteristic of those who rely on such technology for communication.

In sum, this final study in this dissertation series demonstrates the urgent need for speech technologies that can accommodate the complexities of nonbinary identities. By creating more inclusive synthetic voices that represent the diverse gender identities of their users, these technologies can help protect against harm and promote inclusivity and recognition for nonbinary individuals. This intersection of nonbinary gender identity and synthetic voices is a promising field of research, with limited studies thus far, and the case study presented aims to shed light on the unique

experiences of nonbinary individuals who use synthetic voices while identifying areas for further research and development in this domain.

## **1.5 Conclusion**

This dissertation series addresses a gap in existing investigations by delving into gender cues in synthetic speech creation and perception. In particular, the project ultimately centers around the frequently ignored group of individuals who identify as nonbinary and use speech-generating devices. By breaking down the acoustic signals that convey gender, examining how gender is seen in speech, and creating synthetic voices that embrace a more extensive comprehension of gender, this series will offer significant insights into the complex aspects of gender performance in verbal correspondence. The descriptions in this section of the document are merely a preview of what will be expanded upon in the coming chapters, each of which serves as a standalone study to be published. Specifically, the following chapters address gaps in the literature by examining the encoding of gender in speech using a community-informed approach and examining how synthetic voices created by using gender-expansive speech are perceived by people of a variety of genders as well as how the use of these voices encodes nonbinary gender identity for a nonbinary speech-generating device (SGD) user. This research is crucial for further understanding how gender-related traits manifest in speech and aims to explore how people with non-conforming gender identities express themselves through speech and how these expressions relate to their gender in a gradient and multidimensional way. By exploring the variables that impact gender encoding in speech, such as individual identity and group membership, this series will illuminate multifaceted elements at play and lends to a diversity science framework of gender perception (Tripp &

Munson, 2021). This work will improve our understanding of how gender signals are conveyed in speech and will illuminate implications for inclusive augmentative and alternative communication.

## **Chapter 2**

### **GENDER EXPANSIVE SPEECH CHARACTERISTICS**

Part of this paper was published in the INTERSPEECH 2023 proceedings and is housed in the International Speech Communication Association (ISCA) archives. The citation for the published paper is noted in Appendix A.

#### **2.1 Gender in Speech**

The vast majority of previous research has attributed differences in speech between people of different genders (e.g. between men and women) as being due to anatomical differences between cisgender binary genders or as a result of physiological differences alone, e.g. differences in vocal fold size and tension (Dabbs & Mallinger, 1999; Evans et al., 2008; Glaser et al., 2016). However, such differences are now better understood as the products of a complex phenomenon that involves social and/or intentional articulatory factors rather than just anatomical factors. Gender cues in speech are influenced by language and culture (Van Bezooijen, 1995; Yuasa, 2008), socialization (Ferrand & Bloom, 1996), and individual identity (Weirich & Simpson, 2018), and can change over time (Yuasa, 2008). In terms of individual identity, investigations are lacking in how those who are gender expansive (e.g. transgender and/or nonbinary) may encode their gender into voice in ways that move beyond a cisheteronormative framework.

In the discussion that follows, the focus is adult voice. Comparatively little research has been done on children's voices along dimensions of interest for this dissertation project.

### **2.1.1 Fundamental Frequency ( $f_0$ ) and Voice Quality**

Much of previous research on gender in speech has focused on gender as a binary and has contrasted findings for the two binary genders: men and women. This research clearly documented some predictable differences between acoustic properties of adult male and female speech, such as in fundamental frequency ( $f_0$ ). Men tend to produce a lower  $f_0$  with averages between 107 – 132 Hz, and women tend to produce a higher  $f_0$  with averages between 196 – 224 Hz (Davies & Goldberg, 2006).

While previous research on binary genders has brought attention to gender in speech, this research did not account for nonbinary or transgender speech. Our preliminary research addressing this gap has shown that nonbinary individuals'  $f_0$  tends to fall in the middle of transgender and cisgender men's and women's ranges, with an average around 144 Hz (Schmid & Bradley, 2019) and large variability in  $f_0$  production across individuals. These properties do not fit into the patterns of typical cismale or cisfemale voices and inter-talker variability defies binary categorization (Zimman, 2017). A recent study by Brown et al. (2022) described voice quality and fundamental frequency of nonbinary talkers. The main objective of the study was to investigate voice quality and pitch variation in nonbinary individuals and compare findings to those of binary talkers. Recordings were analyzed from 10 nonbinary and 10 binary men and women talkers reading a set of standardized sentences. Acoustic analyses extracted fundamental frequency ( $f_0$ ), an approximate correlate of pitch, and harmonics-to-noise ratio (HNR), an approximate correlate of voice quality. Main

findings were: (1) nonbinary talkers had a “brighter” voice quality compared to binary talkers, as indicated by larger HNR values - that is, nonbinary voices had less noise and more harmonics numerically, giving their speech a clearer, more "ringing" quality – and (2) nonbinary talkers showed more variation in  $f_0$ , specifically a wider range of pitch values and more frequent pitch changes in speech. These studies indicate that nonbinary individuals may use a wider range of vocal inflections and intonations to express themselves. If replicated, these findings suggest that nonbinary individuals may have unique vocal characteristics that distinguish them from binary individuals, and that these characteristics may be related to their gender identity.

### **2.1.2 COG and Peak Frequency of [s] in English**

While  $f_0$  is indeed a gender-conveying variable for voice production and perception, it is not the only one. More recent research has found that other variables are also important for conveying gender. In fact, when it comes to gender perception,  $f_0$  only accounted for 41.6% of the perceptual ratings of voice gender in one study (Leung et al., 2018). Other factors such as acoustic center of gravity (COG) of [s] (Jongman et al., 2000; Fuchs & Toda, 2010; Zimman, 2017), the average of the frequencies of a segment weighted by their amplitudes, also play a role. Similarly to  $f_0$ , women tend to produce [s] at a higher spectral frequency than men (Flipsen et al., 1999; Jongman et al., 2000; Fuchs & Toda, 2010; Heffernan, 2004; Schwartz, 1968; Stuart-Smith, 2007). However, it should be noted that these values are often obtained from a subset of women and men that may reflect the dominant culture. The typical COG averages for [s] in women range between 6,400 – 8,500 Hz, and for men between 4,000 – 7,000 Hz (Avery & Liss, 1996; Flipsen et al., 1999; Fuchs & Toda, 2010; Nittrouer, 1995; Nittrouer et al., 1989; Stuart-Smith, 2007; Tjaden & Turner,

1997). In addition to these findings for binary talkers, Zimman (2017) found that in a diverse group of transmasculine talkers, different identities within transmasculinity (e.g. trans men vs nonbinary transmasculine individuals) were encoded in [s] and that these different identities also intersected with queer sexualities. The group of queer transmasculine individuals showed what Zimman calls a “stylistic bricolage,” i.e., the mixing and matching of sociophonetic cues: those who identified as not only very masculine but also as queer used a low  $f_0$  combined with high COG of [s] to signal queer masculinity.

### **2.1.3 Vocal Tract and Formant Frequencies**

Vocal tract length (VTL) and formant frequencies are also known to be correlated with gender. Men tend to produce lower formants in general, while women produce higher formants (Davies & Goldberg, 2006; Weirich & Simpson, 2018). One factor that influences formants is indeed VTL. Longer vocal tracts result generally in lower formants and smaller vowel spaces overall, while shorter vocal tracts result in higher formants and larger vowel spaces (Carew et al., 2007; Weirich & Simpson, 2018). However, other factors such as the shape of the vocal tract, influenced greatly by tongue positioning, also have an impact on formant frequencies and these factors are also tied to gender (see Raphael et al., 2012, p. 95-109; Lammert and Narayanan, 2015; Johnson, 2020 for an overview).

Vocal tract manipulation has been used as a technique in speech-language pathology to train transfeminine people to achieve a more feminine voice. One such technique is to encourage a forward tongue carriage (which raises  $f_2$ ) and lip spreading (which shortens the vocal tract, raising all formants, and raises  $f_3$  in particular) (Carew et al., 2007).

These acoustic variables are not “all or nothing” – they exist on a continuum on which individuals can encode identity along a gradient. This means that although men may on average have longer vocal tracts than women, both genders can still manipulate VTL. For example, one study found that more feminine men had larger vowel spaces were, reflecting a shortening of the acoustic VTL (Weirich & Simpson, 2018). However, that study did not explicitly include gender expansive talkers and only looked at (presumably) cisgender men and women.

## **2.2 Questions and Hypotheses**

The use of a binary categorization of gender or even a continuum of masculinity to femininity is not conducive to the study of nonbinary voices. Instead, a paradigm that includes continuous (masculine, feminine, and other scales) as well as categorical (male, female, nonbinary) variables of gender can be informative. We define “other” as simply any identity or expression that is neither masculine nor feminine. Such an approach accommodates the variability of gender expansive voices, provides more opportunities for grouping in analysis, and allows for a more holistic view of a given person and their gender. These types of scales were used in our previous studies (Hope & Lilley, 2020; Hope & Lilley 2022), which gave participants both identity and expression variables to describe themselves and the stimulus voices. This framework of gender is also based on the Gender Unicorn which was formulated by and for transgender individuals (TSER, 2015). While this approach does not fully capture all aspects of gender, it provides a basis for exploring gender in ways which are multidimensional.

Our questions were broad because of the descriptive nature of this study: (1) how do different groups of gender expansive people (e.g. transfeminine,

transmasculine, and nonbinary individuals) encode gender into speech and (2) how do gender expansive people encode gender into speech along the continua of *Masculine*, *Feminine*, and *Other gender*? We hypothesized that different groups of gender expansive people would encode gender into speech differently. Given that our previous research has found that nonbinary talkers produce average fundamental frequency in between the averages for men and women (Schmid and Bradley, 2019), we hypothesized that transfeminine individuals would have higher COG of [s] compared to transmasculine individuals, and that nonbinary people would pattern in the middle of transfeminine and transmasculine talkers for different acoustic variables such as COG of [s] and formant frequencies. Our second hypothesis was that there would be (1) a positive correlation between *Feminine gender* (i.e. identity and expression) and  $f_0$ , COG and peak frequency of [s], and vowel formants, and (2) a negative correlation between *Masculine gender* and those acoustic variables. For *Other gender*, we anticipated that there would be a more nuanced and complex relationship between *Other gender* and the acoustic variables investigated.

## **2.3 Methods and Materials**

### **2.3.1 Participants**

Sixteen gender expansive (e.g. transgender or nonbinary) participants over the age of 18 years were recruited online and via word-of-mouth to participate in a speech study that was approved by the IRB of the University of Delaware (see Appendix B). Most of these participants came from the mid-Atlantic region of the United States. All were lifelong talkers of American English who have always used it as their primary language at home. Participants ranged in age from 20 to 42 years (*mean* = 28.3,

$SD = 5.4$ ), and had a variety of genders including but not limited to “transmasculine”, “transfeminine”, “agender”, “genderqueer”, and overlapping identities, e.g. “nonbinary and transmasculine” or “transmasculine and agender.” I participated as a subject in this study as a member of the gender expansive community as the work was preliminary and descriptive.

### **2.3.2 Data Collection**

Before recording, the participants read and electronically signed an informed consent statement. They then completed a survey with demographic questions relating to gender identity, age, race, and an open-ended question about gender and speech (see Appendix C for the survey text). Participants provided their *Male*, *Female* and *Other gender identity* on three independent scales of 0 to 100 and were told that these did not have to add up to 100. They repeated this scaling for *Masculine*, *Feminine*, and *Other gender expression*, resulting in six total gender variables.

Participants’ gradient gender along the six continua are shown in Figure 2.1. The variables *Other gender identity* and *Other gender expression* had the highest average, median, and max scores across the continua, reflecting the nonbinary identities of this sample.

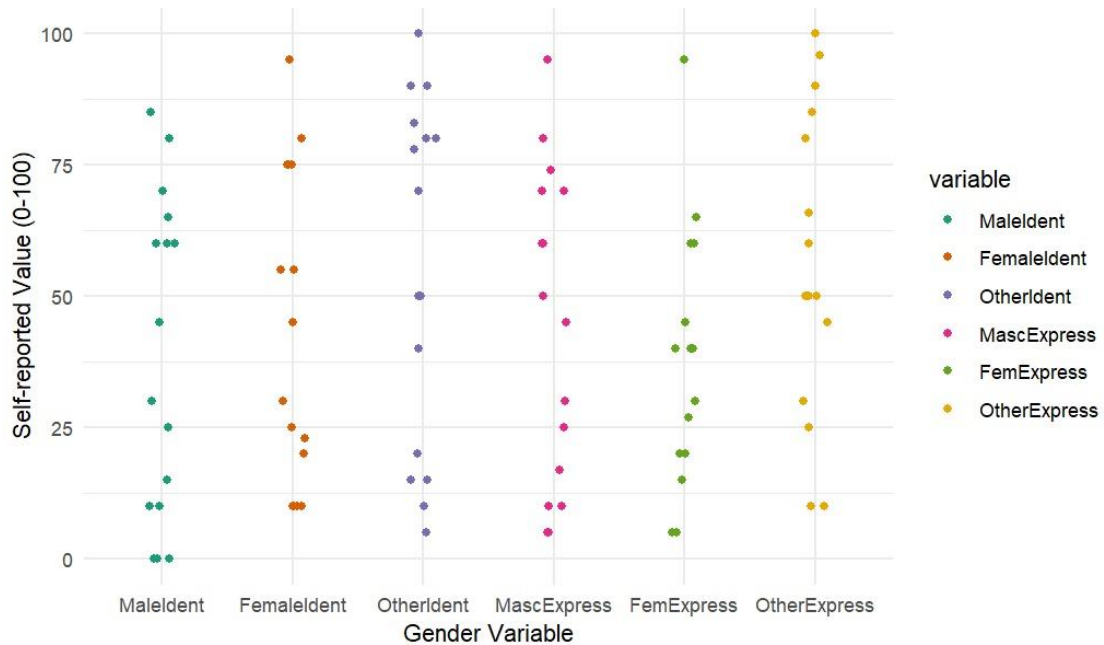


Figure 2.1 Strip chart of the six gender gradient gender variables. MaleIdent represents gradient *Male gender identity*, FemaleIdent represents gradient *Female gender identity*, OtherIdent represents gradient *Other gender identity*, MascExpress represents gradient *Masculine gender expression*, FemExpress represents gradient *Feminine gender expression*, and OtherExpress represents gradient *Other gender expression*.

Next, participants proceeded to record 10 test sentences used in the ModelTalker database (Bunnell et al., 2017). These recordings were used as a screening set to ensure participants' microphone and physical environment were adequate for recording. The screening set was evaluated for background noise or noise generated by their microphone that could cause an issue in speech analysis; it was also evaluated to make sure the participants were reading the sentences at an appropriate pace and in a natural manner, e.g. not focusing on overly enunciating or being too theatrical. Following completion of these tasks, participants recorded the first 400 sentences in the ModelTalker database. Sentences in this database were chosen to

cover the widest possible range of the most commonly occurring diphones and triphones in English.

### 2.3.3 Acoustic Analyses

$f_0$  was extracted using a Praat (Boersma & Weenink, 2021) script that assessed recorded segments with a minimum of 75 Hz and a maximum of 400 Hz for each sentence; then  $f_0$ s were averaged for the 400 sentences to obtain the average  $f_0$  per talker. COG and peak frequency of [s] were extracted using a separate Praat script; we limited the context to word-initial prevocalic singletons, yielding 39 [s] tokens per talker. The COG and peak frequencies were averaged across all tokens to obtain each talker's average COG and peak frequency for [s]. Vowel formants were extracted using a modified Praat script that calculates formant frequencies at the midpoint of the vowel (Kent & Vorperian, 2018; Chung et al., 2012). In total, each talker had 371 [i] tokens, 176 [a] tokens, 203 [u] tokens, 201 [e] tokens, and 175 [o] tokens; formant frequencies were averaged per vowel per talker. Then, using the average formant frequency spacing ( $\Delta f$ ) method described by Johnson (2020), we computed the  $\Delta f$  for each talker and used this value to compute the acoustic vocal tract length ( $aVTL = 34,000 (cm/s)/[\Delta f \times 2]$ ).

### 2.3.4 Statistical Analyses

Statistical analyses, including pairwise two-tailed t tests, Pearson's correlations, and linear regressions, were conducted in R. We used raw formant values for correlations and regressions. We chose to highlight the raw values in this analysis because of the likelihood of over-normalization that can occur especially for a group that may be actively manipulating vocal tract length.

Correlations using the Pearson’s method were computed between each of the acoustic variables and each of the gradient gender variables. All acoustic variables are continuous greater than 0 in Hz. Gradient gender variables are encoded as whole numbers between 0 and 100. Simple linear regressions were then performed for the 10 correlations with  $r > 0.50$  to establish the trendlines and  $R^2$  values that indicate the amount of variance in the dependent variable (the acoustic measure in this case) explained by the independent variable (the six gradient gender variables). The acoustic variables consisted of overall  $f_1$ ,  $f_2$  and  $f_3$  values, average  $f_1$ ,  $f_2$  and  $f_3$  for each of five vowels ([i], [a], [u], [o], and [e]), COG for [s], peak frequency for [s], and average fundamental frequency.

Because this is exploratory, preliminary work, our priority is to not inflate our Type II error rate. As such, we have decided to not make multiple-testing corrections, as that will lower Type I errors at the expense of Type II (Berry, 1990; Roy et al., 2004; Rubin, 2017).

## 2.4 Results

### 2.4.1 Whole Group Summary

Average and standard deviation for  $f_0$ , COG of [s], peak frequency of [s],  $f_1$ ,  $f_2$ , and  $f_3$ , and acoustic vocal tract lengths (aVTL) are shown in Table 1 for the participant pool as a whole.

Table 2.1 Descriptive statistics for acoustic measurements for the whole group of gender expansive talkers

<u>MEASURE</u>	<u>MEAN</u>	<u>STANDARD DEVIATION (SD)</u>
----------------	-------------	--------------------------------

$f_0$ (HZ)	163.1	37.5
CENTER OF GRAVITY (HZ)	5612.5	999.8
PEAK FREQUENCY (HZ)	5727.3	1022.7
AVG $f_1$ (HZ)	544.2	63.6
AVG $f_2$ (HZ)	1770.0	103.6
AVG $f_3$ (HZ)	2730.7	119.8
aVTL (CM)	14.9	1.3

This Table illustrates the characteristics for the pooled group of gender expansive participants (n=16). The mean  $f_0$  for the group was within the “androgynous” or “neutral” range (Davies & Goldberg, 2006) and had a large standard deviation. The center of gravity and peak frequency for /s/ were both roughly the same and had considerable within-group variability.  $f_1$ ,  $f_2$  and  $f_3$  were all within expected ranges and the aVTL varied greatly between group members.

#### 2.4.2 Categorical Gender Groups

To answer our first question about how different groups of gender expansive individuals encode gender into speech, we computed the mean and standard deviations for those groups for  $f_0$ , COG of [s], peak frequency of [s],  $f_1$ ,  $f_2$ ,  $f_3$ , and aVTL. These results are presented in Table 2.2. In order to visualize group differences in vowel production, we created vowel plots of each of the groups, shown in Figure 2.2. Finally, Figure 2.3 shows differences between groups for specific vowels.

Table 2.2 Descriptive acoustic statistics for each gender group

MEASURE	NONBIN (N=8)		TRANSFEM (N=4)		TRANSMASC (N=4)	
	MEAN	SD	MEAN	SD	MEAN	SD
$f_0$ (HZ)	168.6	33.6	145.0	37.5	155.8	48.2
COG (HZ)	6094.3	340.3	5603.5	942.3	4658.0	1416.2
PEAK (HZ)	6257.7	349.4	5553.1	1237.3	4840.7	1254.9
$f_1$ (HZ)	539.8	36.0	552.0	77.9	521.5	28.8
$f_2$ (HZ)	1822.2	82.6	1688.6	71.2	1709.4	42.8
$f_3$ (HZ)	2768.0	108.0	2674.2	162.1	2694.5	12.0
aVTL (CM)	14.9	0.7	14.7	2.7	15.3	0.4

Table 2.2 shows that the Nonbinary group had the highest average  $f_0$ , while the Transfeminine group had the lowest. The Transmasculine group had the most variability (SD) in  $f_0$  across participants. The Nonbinary group also had the highest mean COG and peak frequency of [s], while the Transfeminine group fell in the middle of the Nonbinary and Transmasculine groups for these values. Additionally, the Nonbinary group had smaller variability between talkers in COG and peak frequency of [s] when compared to the larger variabilities of the Transfeminine and Transmasculine participants.  $f_1$  had similar values between groups, and the Transmasculine group had both the lowest  $f_1$  and the smallest variability in  $f_1$  production between talkers. For  $f_2$  and  $f_3$ , the Nonbinary group had the highest average values, while the Transfeminine and Transmasculine groups showed lower values. However, the Transfeminine group had a very large variability across talkers

while the Transmasculine group had very small variability. Finally, for the aVTL the Transmasculine group had the longest average aVTL, while the Transfeminine group had the shortest average aVTL and the greatest aVTL variability across talkers. These aVTL differences were not statistically significant between groups.

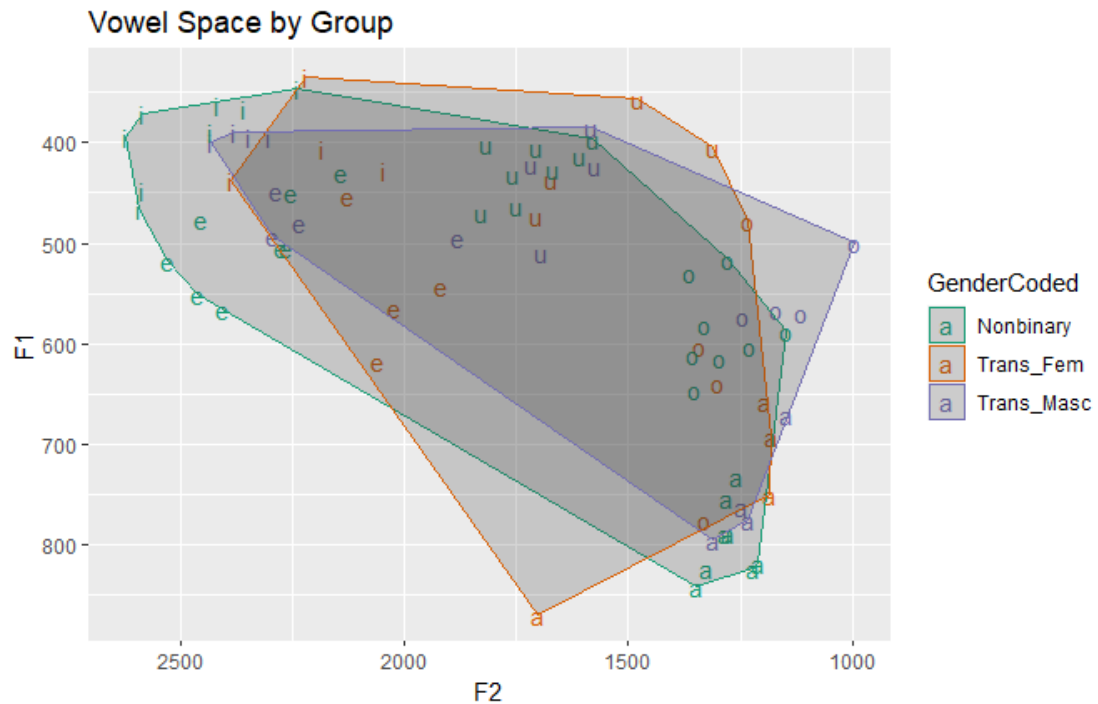


Figure 2.2 Unnormalized vowel spaces for each gender group. Nonbinary (dark green) represents the Nonbinary participants, Trans\_Fem (dark orange) represents the Transfeminine participants, and Trans\_Masc (dark blue) represents the Transmasculine participants. Each vowel shown represents an individual’s average vowel production.

Figure 2.2 shows the major group differences in the vocal tract, seen by the higher  $f_2$  of [i] and [o] in the Nonbinary group and the lower  $f_2$  of [o] in the Transmasculine group. The Transfeminine group had a more varied vowel space with

vowels spread further apart compared to the clustering observed in the Nonbinary and Transmasculine groups.

Pairwise two-tailed t tests were conducted to examine differences between groups for the various acoustic variables. Several of the t tests showed significant differences between categorical gender groups and these significant findings pertained only to the vowel acoustics; t tests for spectral mean, peak frequency, and  $f_0$  were all statistically insignificant. The significant findings are shown in the boxplots in Figure 2.3. Removing the author from the t tests showed no change in statistical significance for any of the variables.

The t test for  $f_2$  of [i] between groups was significant between the nonbinary and transfeminine groups ( $p = 0.011$ ). The t test for  $f_2$  of [e] between groups was also significant between the nonbinary and transfeminine groups ( $p = 0.009$ ). The t test for  $f_2$  of [o] between both the nonbinary and transmasculine groups as well as between the transmasculine and transfeminine groups were significant ( $p = 0.013$  and  $p = 0.025$  respectively) (see Figure 2.3).

### **2.4.3 Gradient Gender Results**

To answer the question “how do gender expansive people encode gender into speech along the continua of masculine, feminine and other gender?”, we analyzed correlations between the various acoustic features of interest and the six gradient gender variables.

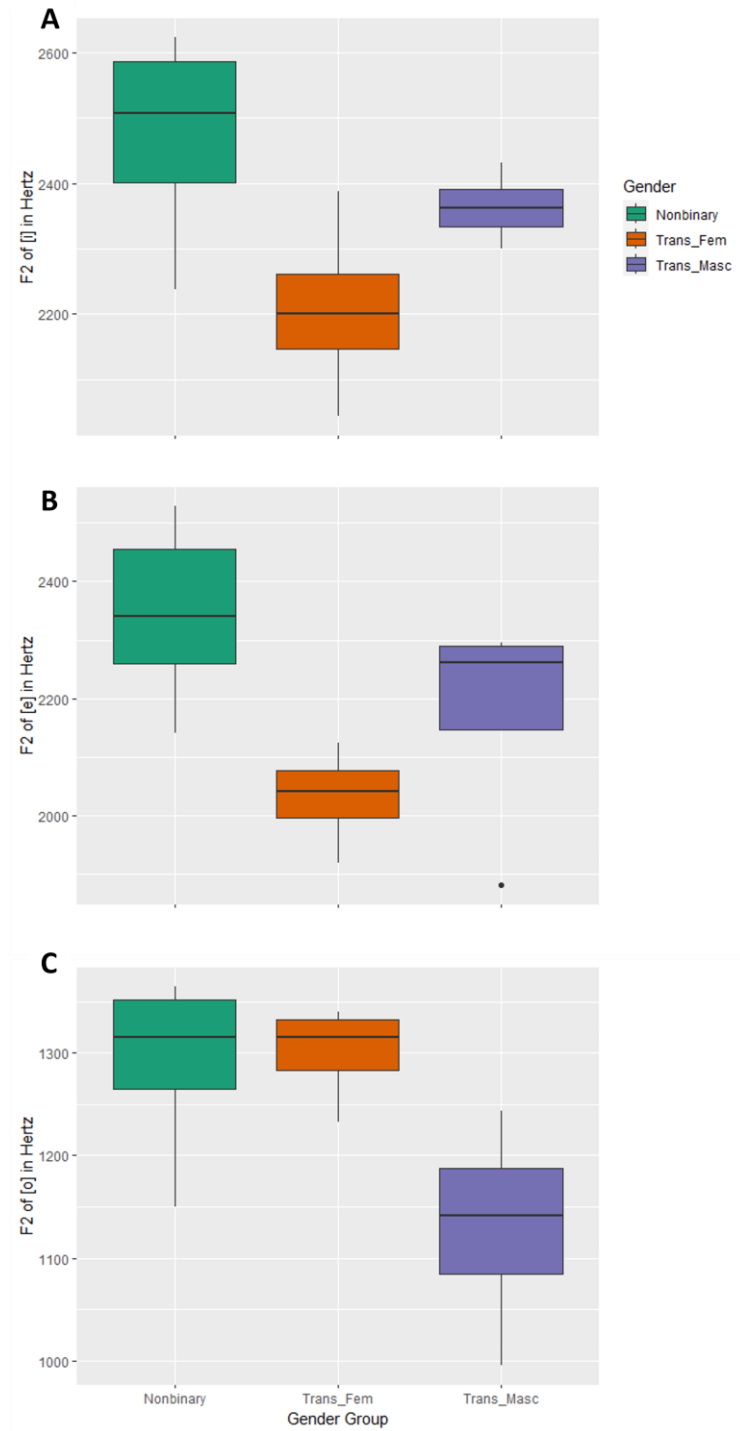


Figure 2.3 Boxplots showing the differences in A)  $f_2$  of [i], B)  $f_2$  of [e], and C)  $f_2$  of [o] for the three groups of Nonbinary, Transfeminine (Trans\_Fem) and Transmasculine (Trans\_Masc).

### 2.4.3.1 Correlations

Fundamental frequency, COG for [s] and peak frequency for [s] had no statistically significant correlations with any of the gender variables. Several of the formant frequency acoustic measures did have statistically significant relationships and are shown in Table 2.3. For further transparency, Table S1 is provided in Appendix C showing the *p*-values and Pearson's *r* values for the relationships between the gender and acoustic variables with and without the first author when removing the author caused a change in statistical significance. Overall, excluding the first author changed the significance in only a few of the results (four out of 90, or ~4% of all results). Results including the author are presented below.

Table 2.3 Correlations per formant per vowel. Values represent Pearson's *r* with *r* > 0.50 bolded.

		FEM	MALE	OTHER	FEM	MASC	OTHER
		GENIDENT	GENIDENT	GENIDENT	GENEXP	GENEXP	GENEXP
[i]	$f_1$	-0.12	-0.24	0.11	0.34	-0.27	-0.08
	$f_2$	<b>-0.58</b>	0.21	<b>0.69</b>	-0.32	0.32	0.48
	$f_3$	-0.40	-0.01	<b>0.56</b>	-0.16	0.11	0.32
[a]	$f_1$	-0.31	0.10	<b>0.54</b>	0.01	0.13	0.33
	$f_2$	0.18	-0.22	0.15	0.17	-0.17	0.11
	$f_3$	0.11	-0.29	0.20	0.37	-0.32	0.28
[u]	$f_1$	-0.23	0.16	-0.09	0.16	0.05	-0.32
	$f_2$	-0.38	0.12	0.35	-0.08	0.12	0.10
	$f_3$	<b>-0.61</b>	0.14	<b>0.61</b>	-0.18	0.22	0.27
[o]	$f_1$	0.13	-0.30	0.15	0.47	-0.33	0.04

	$f_2$	0.47	<b>-0.68</b>	0.30	0.50	<b>-0.69</b>	0.29
	$f_3$	-0.30	-0.06	0.50	0.09	-0.02	0.43
[e]	$f_1$	0.26	-0.42	0.01	0.48	-0.45	0.02
	$f_2$	-0.44	0.12	<b>0.64</b>	-0.16	0.22	<b>0.55</b>
	$f_3$	-0.43	-0.06	<b>0.56</b>	-0.09	0.07	0.37

The significant correlations were modeled using linear regressions as described in the following subsections. We chose to show the unnormalized formants in the regression analysis because unnormalized formant values will always reflect both articulation and overall vocal-tract size. The perception of gender involves the perception of both of these factors (articulation and overall size). Additionally, normalization is a method that often focuses on listener perceptions and how listeners perceive distinct vowels, while this study focused on speech production. Analyzing the unnormalized data led to the greatest preservation of trends in the data. Comparing Table 2.3 with Table S2 (see Appendix C), the normalized method shows no significant correlations with  $f_3$ ; this is likely due to the fact that the Johnson (2020) normalization method tries to account for differences in vocal tract length. By reducing differences in vocal tract length in this population, we could be missing crucial information about how talkers are purposefully manipulating vocal tract length to signal gender.

#### 2.4.3.2 Effects of Gender on $f_1$

There was only one statistically significant relationship between any of the gender variables and the first formant frequency. This was the relationship between *Other gender identity* and  $f_1$  of [a].

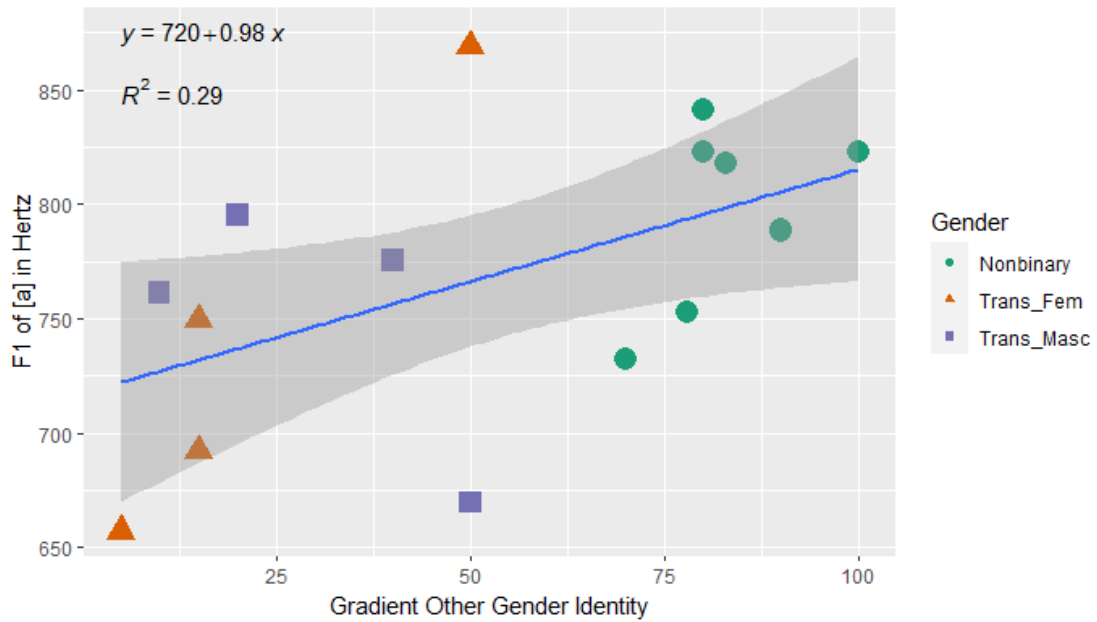


Figure 2.4 Linear regression of the relationship between  $f_1$  of [a] in hertz and gradient Other gender identity.

As illustrated in Figure 2.4,  $f_1$  of [a] increased as gradient *Other gender identity* increased. Additionally, nonbinary talkers tended to have higher *Other gender identity* scores and higher  $f_1$  of [a] compared to the transfeminine and transmasculine talkers.

### 2.4.3.3 Effects of Gender on $f_2$

$f_2$  was shown to be significantly correlated with *Female gender identity* and *Other gender identity*. Figure 2.5 shows linear regressions of the relationships between  $f_2$  of [i] in hertz and gradient *Female gender identity* (A),  $f_2$  of [i] in hertz and gradient *Other gender identity* (B),  $f_2$  of [e] in hertz and gradient *Other gender*

identity (C),  $f_2$  of [e] and gradient *Other gender expression* (D),  $f_2$  of [o] gradient *Male gender identity* (E), and  $f_2$  of [o] and gradient *Masculine gender expression* (F).

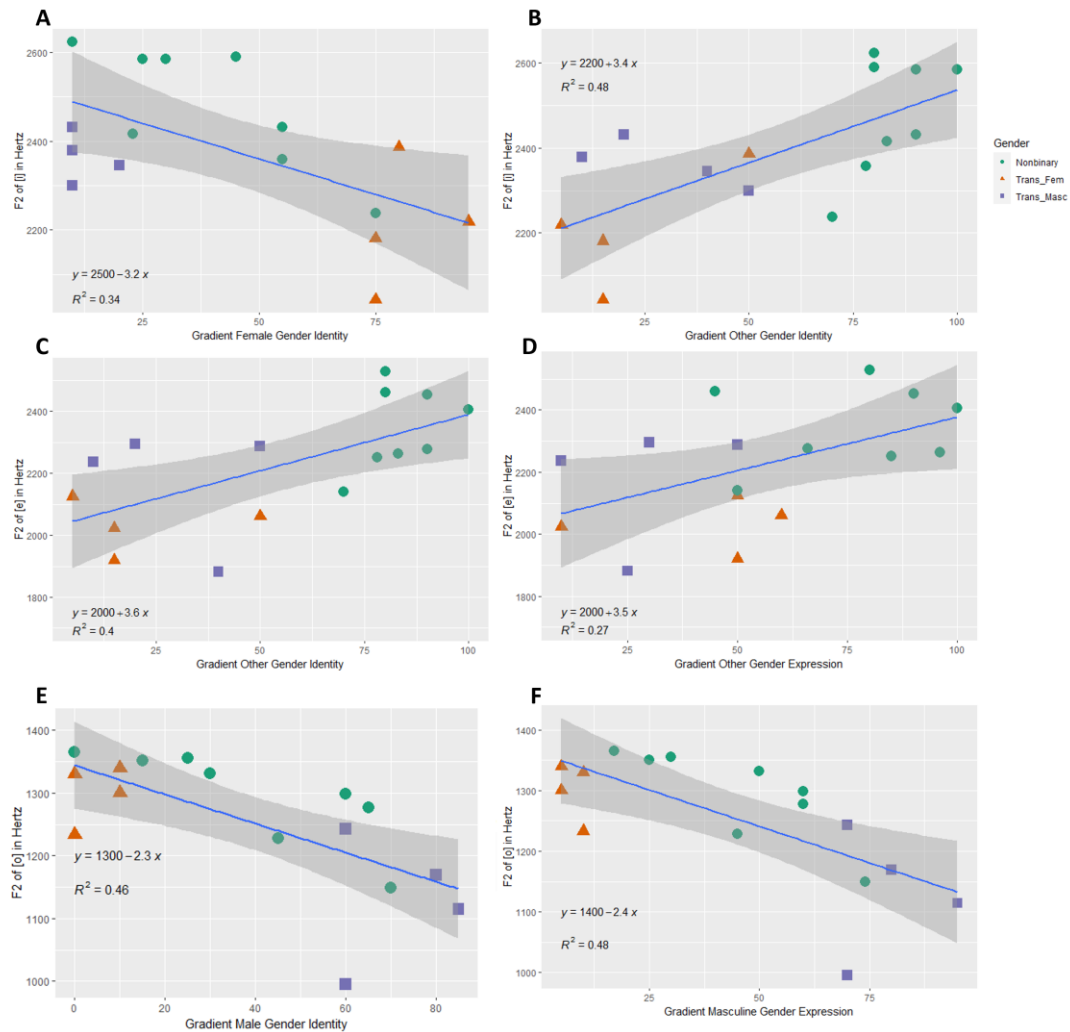


Figure 2.5 Linear regressions of the relationships between  $f_2$  of [i] in hertz and *gradient Female gender identity* (A),  $f_2$  of [i] in hertz and *gradient Other gender identity* (B),  $f_2$  of [e] in hertz and *gradient Other gender identity* (C),  $f_2$  of [e] and *gradient Other gender expression* (D),  $f_2$  of [o] gradient *Male gender identity* (E), and  $f_2$  of [o] and *gradient Masculine gender expression* (F).

The relationships between  $f_2$  of [i] and gradient *Female gender identity* and  $f_2$  of [i] and gradient *Other gender identity* show reverse trends with the relationship with  $f_2$  of [i] getting lower as *Female gender identity* increases and  $f_2$  of [i] getting higher as *Other gender identity* increases (Figure 2.5 A and B). Both *Other gender identity* and *Other gender expression* were significantly correlated with  $f_2$  of [e] with  $f_2$  of [e] increasing as those two variables increased (Figure 2.5 C and D). Finally, *Male gender identity* and *Masculine gender expression* are both negatively correlated with  $f_2$  of [o] showing that as those gender variables increase,  $f_2$  of [o] decreases (Figure 2.5 E and F).

#### **2.4.3.4 Effects of Gender on $f_3$**

There were several statistically significant relationships between the third formant frequency and the various gender variables. Figure 2.6 shows linear regressions of the following relationships:  $f_3$  of [i] in hertz and gradient *Other gender identity* (A),  $f_3$  of [u] in hertz and gradient *Female gender identity* (B), and  $f_3$  of [u] in hertz and gradient *Other gender identity* (C).

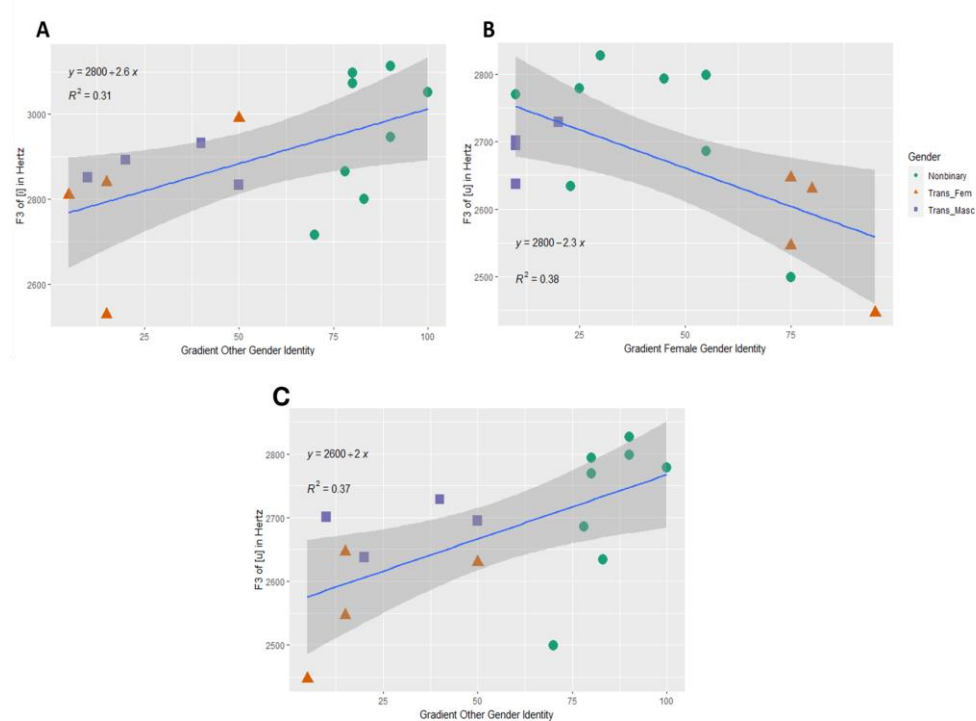


Figure 2.6 Linear regressions of the relationships between  $f_3$  of [i] in hertz and gradient *Other gender identity* (A),  $f_3$  of [u] in hertz and gradient *Female gender identity* (B), and  $f_3$  of [u] in hertz and gradient *Other gender identity* (C).

Figure 2.6 demonstrates several relationships with different directions between  $f_3$  values and gradient gender. As gradient *Other gender identity* increased,  $f_3$  of [i] and [u] both increased. Meanwhile, as gradient *Female gender identity* increased,  $f_3$  of [u] decreased.

## 2.5 Discussion

Based on the findings, gender expansive individuals may be encoding gender in speech both in terms of categories and along a gradient, predominantly through the

use of vowel acoustics in this study. Fundamental frequency and acoustics of [s] that we explored in this study were not found to be statistically significantly related to any of the six gradient gender variables (*Male*, *Female* and *Other gender identity* and *Masculine*, *Feminine* and *Other gender expression*) and did not categorically distinguish any of the three subsets of talkers from each other. In Figure 2.2, we presented a vowel space of the different gender groups. We noted the presence of /u/ fronting, which is consistent with Mid-Atlantic dialects in North America (Havenhill, 2024). Another note is that the presence of group differences exhibited in the vowel space may be attributed to differences in the vocal tract sizes of participants. The Nonbinary participants had aVTL of 14.9, the Transfeminine participants had aVTL of 14.7, and the Transmasculine participants had aVTL of 15.3. While these do reflect differences across groups that may have an anatomical component, we did not want to exclude the effects of participants deliberately manipulating vocal tract length. With this consideration in mind, findings from the vowel space show that transmasculine GE talkers tend to produce vowels that are more back (i.e.  $f_2$  is lower) compared to nonbinary GE talkers. Additionally, transmasculine talkers tended to have a smaller and lower vowel space compared to the nonbinary and transfeminine talkers. This is consistent with results previously reported for male and female talkers who were made to “feminize” or “masculinize” their voices. Both male and female talkers showed that when they were asked to “masculinize” their voice, their  $f_1$ - $f_4$  values became lower and their vowel space smaller. When they were asked to “feminize” their voice, their  $f_1$ - $f_4$  values increased as did vowel space (Cartei et al., 2012). The vowel space in the present study also showed that nonbinary GE talkers have lower [e] and [a] productions in terms of formant frequencies compared to transmasculine and most

transfeminine talkers. Transmasculine people categorically produced [o] with lower  $f_2$  than nonbinary or transfeminine individuals, and there was a significant difference between the transfeminine and transmasculine groups for overall  $f_2$ .

However, in contrast to previous research, we did not find significant relationships between *Feminine gender expression/Female gender identity* and  $f_0$  or *Masculine gender expression/Male gender identity* and  $f_0$ ; indeed, none of the six gender variables explored were significantly correlated with  $f_0$ . In this study we did not find evidence that gender expansive people did use  $f_0$  to encode gender. These results correspond with our recent perceptual findings, which showed gender expansive people do not utilize  $f_0$  in the perception of voice gender (Hope & Lilley, 2020; Hope & Lilley, 2022). While we did not explore perception in the present study, these together are suggestive of a production-perception interface in voice gender processing. Contrary to prior research, our findings reveal that pitch is uncorrelated with gender encoding and perception, potentially indicating an expanded range of cues and internal processing mechanisms influenced by exposure to diverse voices, particularly within the gender expansive community.

We found that  $f_2$  of [o] became significantly lower as gradient *Male identity* increased and *Masculine expression* increased (Figure 2.5), whereby *Masculine expression* had the greater impact of the two. Thus, we have a potential marker of masculinity in this group. However, voice gender expression in this group may be multidimensional with talkers mixing and matching different acoustic properties. For example, those who were nonbinary may have been more or less masculine and used other acoustic markers, such as raising  $f_2$  of [i] to signal *Other gender identity* (Figure 2.5) and/or raising other formants to signal simultaneous *Female gender identity* (as

these gender variables are not mutually exclusive). Thus, a gender expansive person who has both strong other gender and male identity could raise  $f_2$  for [i], signaling other gender identity, while lowering  $f_2$  of [o], to signal male gender identity. This interpretation is speculative and awaits confirmation in future data sets – however, based on the current data set, this is an interpretation that might be entertained.

Another way these talkers may encode simultaneous aspects of gender is by using sibilant production. Although the finding was not statistically significant, nonbinary individuals produced [s] with the highest COG and peak frequency (see Table 2.2) and thus, nonbinary talkers who embody both masculine gender and feminine gender may choose to mix and match a lower  $f_2$  of [o] with a higher [s] COG. In these ways, nonbinary talkers are utilizing “stylistic bricolage” – the mixing and matching of sociophonetic cues – to signal a multidimensional and nonbinary identity.

One limitation of this study was that we were not completely able to control for environmental factors in the recordings. Participants recorded remotely, and, although they passed a screening set of test sentences, they all had variable recording environments. Additionally, because we looked only at a small group of gender expansive talkers, results may not be generalizable to a larger group. Even for a larger group, gender expansive talkers are diverse both in identities and in speech. Finally, while we modeled the *Other gender* variables’ relationships with acoustics using linear regression, this group largely had high “other” gender identity and expression; thus, the low variance in these scores resulted in linear regression being less appropriate. However, we were not powered for more complex modeling. This analysis was preliminary, given our small number of talkers, and further investigation is warranted to examine more complex relationships between gender identity and

formant production and acoustics, in accordance with our hypotheses that *Other gender* has a more complex relationship with acoustic variables than *Female* and *Male gender* do.

## 2.6 Conclusion

This study was an exploratory, descriptive, community-informed and -directed investigation into the production of speech of gender expansive individuals and the gradient encoding of multidimensional, nonbinary gender in speech. Sixteen gender expansive talkers recorded 400 English utterances, which were analyzed for multiple acoustic variables known to be correlated with the encoding of gender. We found significant correlations between formants of different vowels and multidimensional gender variables (gradient masculine, feminine, and “other” gender identity and expression). In particular, we uncovered a strong correlation between  $f_2$  of [o] and gradient *Male identity*, as well as gradient *Masculine gender expression*, indicating that this group may be using tongue backing of [o] (which drops  $f_2$ ) as a way to signal masculinity. Additionally, we found cues for encoding gradient *Other gender identity*, namely raising  $f_2$  of [i]. This work has implications for sociophonetics and speech language pathology. These findings could prompt further research into vocal techniques and therapies for gender expansive people of a variety of genders, as well as the creation of more inclusive speech technology.

### **Chapter 3**

#### **GENDER PERCEPTION IN GENDER EXPANSIVE SYNTHETIC VOICES**

Gender perception in speech is a complex sociophonetic phenomenon that involves both the talker's own encoding of gender using different acoustic cues as well as the listener's biases and experiences in interpreting those cues and mapping them onto different voice genders. Various sociophonetic cues in English, including fundamental frequency, formant frequencies, and spectral properties of sibilants, influence how we perceive gender in speech. These acoustic cues to speech must also be extended to the realm of synthetic voices.

In recent years, the development and use of synthetic voices, particularly in speech-generating devices (SGDs) and virtual assistants, have expanded significantly. These synthetic voices, commonly referred to as text-to-speech (TTS), are now ubiquitous in technology like Alexa and Siri, but are also of vital importance for meeting the communication needs of those who use SGDs. However, the design and implementation of these voices have historically adhered to binary gender norms, resulting in voices that are explicitly or implicitly coded as either male or female. This binary framework has significant limitations, especially for individuals who identify outside the traditional cisheteronormative spectrum. Transgender, nonbinary, and other gender expansive individuals often find themselves marginalized by these synthetic voice systems, which fail to represent their identities adequately (Hope & Lilley, 2023).

The perception of gender expansive synthetic voices is a crucial area of study, as it directly impacts the usability and effectiveness of these technologies for gender diverse populations. Historically, research on speech production and perception has focused predominantly on cisgender individuals, leading to a substantial gap in our understanding of how gender expansive individuals encode and perceive gender in speech. There is a clear need to address this gap by examining how gender expansive synthetic voices are perceived by both gender expansive and cisgender listeners. This understanding is vital for creating more inclusive and representative synthetic voices that can better serve the needs of all users.

The findings from Chapter 2 underscore several key differences in how gender expansive individuals produce speech compared to cisgender talkers. For instance, while traditional metrics like  $f_0$  and the center of gravity (COG) of /s/ have been used to differentiate male and female voices, these parameters do not necessarily apply to gender expansive talkers. Instead, formant frequencies, particularly the second formant ( $f_2$ ), play a significant role in encoding multidimensional, nonbinary gender for gender expansive individuals. These findings suggest that gender expansive synthetic voices might require different acoustic properties to accurately reflect the identities of their users, and further that they should be developed from gender expansive talkers themselves.

Moreover, the perception of these voices may be influenced by the listener's own gender identity and experiences. It is a possibility that gender expansive listeners are more attuned to recognizing and affirming nonbinary cues in synthetic voices than cisgender listeners. This implies that the effectiveness of a synthetic voice in conveying gender identity can vary widely depending on the listener's background and

familiarity with gender expansive speech patterns. Therefore, it is not only important to create synthetic voices that technically encode gender correctly but also to understand how these voices are perceived across different populations.

Investigating the perception of gender expansive synthetic voices is essential for advancing speech technology that is truly inclusive. By understanding how different listeners perceive these voices, we can develop better tools and technologies that affirm the identities of gender expansive individuals and enhance their communication experiences. This chapter aims to delve into the nuances of gender expansive speech perception, drawing on the latest research to highlight the importance of inclusive design in synthetic voice technology.

## **3.1 Introduction**

### **3.1.1 Fundamental frequency**

Fundamental frequency ( $f_0$ ), whose perceptual correlate is pitch, plays a critical role as a cue in gender perception. Extensive research has established that men tend to have lower  $f_0$  values than women. This difference in  $f_0$  between binary genders primarily stems from physiological factors such as vocal fold size and tension (Dabbs & Mallinger, 1999; Evans et al., 2008; Glaser et al., 2016). The majority of the relevant previous work has been done with cisgender individuals.

Additionally, it is important to note that reducing the differences in  $f_0$  to categorical differences between men and women does not fully account for all variations in  $f_0$  production and perception. Sociocultural and individual differences in voice pitch also contribute to  $f_0$  variation, influencing gender perception. For instance, the perception of women's voices varies along a gradient depending on their pitch.

Women with higher-pitched voices are often perceived as more feminine and attractive compared to those with lower-pitched voices (Pisanski et al., 2018). Conversely, lower mean  $f_0$  in men's voices has been linked to higher perception of dominance (Hodges-Simeon et al., 2010). These factors interact with other sociocultural variables, including race and culture (Van Bezooijen, 1995; Levy, 2023), further complicating the understanding of gender perception.

There are several ways that  $f_0$  varies. First, it can vary immensely within individuals; each person may fluctuate  $f_0$  within and between sentences and in different contexts. This variation can also differ between groups. For example, women tend to use more of their pitch range compared to men, including variations of  $f_0$  within individual words (Hancock et al., 2014). Therefore,  $f_0$  is found to vary within and between individuals and groups. It is crucial to acknowledge that gender is not strictly binary, and the variation in  $f_0$  between individuals who are nonbinary is high, reflecting their unique gender identities (Schmid & Bradley, 2019). This poses a challenge to the cisheteronormative framework when attempting to categorize individuals solely based on pitch. Nonbinary individuals may exhibit  $f_0$  values that do not conform to traditional male or female patterns. Therefore, relying only on  $f_0$  to determine gender would overlook the diversity within nonbinary experiences. The implications of reliance of on  $f_0$  for gender determination extend to binary individuals who talk in ways that are outside the general trends for their gender identity.

Technological advancements have provided researchers with the ability to manipulate  $f_0$  in speech synthesis, enabling investigations into how listeners perceive gender when  $f_0$  is modified. In a previous study from our lab, synthetic voices, constructed using voice data from (presumably) cisgender men and women, with

average  $f_0$  values within the "neutral range" were categorized as "nonbinary" more frequently and rated higher on an "other" gender scale by gender expansive individuals (e.g., transgender and/or nonbinary) compared to ratings by cisgender individuals (Hope & Lilley, 2022). Additionally, gender expansive listeners relied less on the  $f_0$  cue and placed more emphasis on other factors described shortly in their gender categorization compared to cisgender listeners (Hope & Lilley, 2020). However, these voices, as noted, were created from (presumably) cisgender men and women's speech data, and there has been limited research examining the perception of actual nonbinary voices.

In conclusion, while  $f_0$  is a crucial sociophonetic cue in gender perception, it is essential to consider sociocultural and individual factors that influence  $f_0$  differences and their relationship with gender. Gender cannot be solely determined by pitch, especially in light of nonbinary identities and the diversity they encompass. It is important to explore the perception of nonbinary voices further, considering multiple cues beyond  $f_0$  to better understand the intricate nature of gender perception.

### **3.1.2 Formant Frequencies**

While  $f_0$  offers valuable information on gender perception, it should not by any means be regarded as the sole determinant of gender. Recent research indicates that  $f_0$  accounts for only 41.6% of the variation in gender perception ratings in speech (Leung et al., 2018). Another significant factor influencing gender perception in speech is formant frequencies. Formants are a product of vocal tract size and the location of constrictions in the vocal tract. Any given pattern of formants--particularly the lowest two formants ( $f_1$  and  $f_2$ )--reflects both of these factors. Formant frequencies can cue gender because they represent a particular gendered way of

producing a vowel (such as a fronted or non-fronted /u/, at least in dialects where /u/-fronting is related to gender), because they cue the listener that the vocal tract has a particular size, or both. Longer vocal tracts tend to produce lower average formants overall compared to shorter vocal tracts. However, the place of constriction in the vocal tract also influences formant frequencies;  $f_1$  is typically associated with the size of the pharyngeal cavity (which is related to tongue height), and  $f_2$  is associated with the length of the oral/front cavity. Studies have consistently demonstrated gender-related differences in formant frequencies contributing to the perception of masculinity or femininity in speech. Men typically exhibit lower  $f_1$  and  $f_2$  values than women, primarily due to physiological differences in vocal tract length and shape (Raphael et al., 2012, pp. 95-109). These differences contribute to the perception of a more "masculine" voice for men and a more "feminine" voice for women (Leung et al., 2018). Despite these differences between men and women with respect to size and shape of the articulatory cavities, it is important to bear in mind that these differences are based largely on studies of demographically unspecified groups that most likely involved cisgender people exclusively.

Formant frequencies exhibit a correlation with voice gender, for which female talkers typically display higher formant frequencies compared to male talkers (Davies & Goldberg, 2006; Cartei & Reby, 2013; Leung et al., 2018; Nagels et al., 2020). Research by Kawitzky and McAllister (2019) revealed that higher  $f_2$  values were associated with higher ratings of femininity of speech and determined that both  $f_0$  and  $f_2$  jointly influence a listener's perception of the talker's gender. A production study assessing talkers' gender identity with regard to both masculine and feminine gender scales revealed that men self-identifying as less masculine tended to exhibit higher

fundamental frequencies ( $f_0$ ) and larger vowel spaces compared to men who reported higher degree of masculine gender (Weirich & Simpson, 2018). Weirich & Simpson furthermore performed a voice gender perception experiment that demonstrated an interaction between formant frequencies and the *listener's* gender in listener identification of gender.  $f_0$  has been generally thought to have a more pronounced impact on voice gender perception than formant frequencies; however, research regarding the influence of formant frequencies continues to put that traditionally held belief into question. One study indicated that an "androgynous"  $f_0$  contour, combined with either masculine or feminine formant frequencies, significantly influenced voice gender perception (Skuk & Schweinberger, 2014). However, it is worth noting that that study employed a binary choice approach for voice gender perception, allowing listeners to categorize voices as either "male" or "female." These findings highlight the importance of individual variation in the encoding of gender through vocal tract size and shape.

Nonbinary individuals reject binary categorization. As such, it would not be surprising if their speech defied binary categorization as well; this means they may not encode gender in traditional male or female speech patterns. These individuals may utilize techniques such as manipulating the size and shape of the pharyngeal or oral cavity to raise or lower  $f_1$  or  $f_2$ . Such manipulation may allow nonbinary individuals to influence gender perception and create a stylistic bricolage (Zimman, 2017), which is a mixing and matching of sociophonetic cues. These individuals may mix and match techniques, altering  $f_0$  alongside formants in unique ways, and utilize a combination of strategies to convey their gender identity. Consequently, speech produced in this

stylistic bricolage may be perceived differently from speech which is not produced in this style in terms of gender, whether categorically or on a gradient scale.

In summary, while  $f_0$  is essential for gender perception in speech, it explains only a portion of the variation in gender perception ratings. Formant frequencies, which are influenced by vocal tract size and shape, play a significant role in shaping perceptions of masculinity and femininity in speech. The individual variation and unique approaches employed by nonbinary individuals in manipulating formants and  $f_0$  contribute to the diverse ways gender is perceived in their speech. Further exploration of these factors is important to deepen our understanding of gender perception and representation in diverse linguistic contexts.

### **3.1.3 Sibilants /s/ and /ʃ/ in English**

Sibilant perception in English has also been identified as a relevant factor in gender perception. Sibilants are sounds such as /s/ and /ʃ/ that exhibit a distinct hissing or hushing quality and are characterized by a high-frequency energy concentration. These sibilant sounds serve as important sociophonetic cues in gender perception, providing additional information about the talker's gender identity (Strand & Johnson, 1996; Munson, 2011; Hope & Lilley, 2023).

Studies have shown that there are differences in sibilant production between men and women. Specifically, women tend to produce [s] with higher spectral means and energy peaks compared to men (Jongman et al., 2000; Fuchs & Toda, 2010; Weirich & Simpson, 2015). One possible explanation is the sociophonetic encoding of gender by women, involving the advancement of the tongue further forward in the mouth during [s] production. This tongue positioning leads to higher spectral means and peak frequencies (Fuchs & Toda, 2010). Additionally, the duration of sibilant

sounds can also play a role in gender perception, with research suggesting that women tend to produce longer sibilants than men (Weirich & Simpson, 2015). It is crucial to remember the demographics of these studies were likely to have been limited. The gender and sexual identities of the women and men in each of the studies (cis or trans, queer or straight) is often left unspecified.

Thus, the previous studies are limited and do not represent the diverse range of women's or men's voices, including gender expansive women and men. The existing research predominantly falls within the framework of binary gender categorization, leaving a significant gap in understanding sibilant perception in nonbinary speech. Therefore, more research is needed to explore how sibilant perception contributes to gender categorization among nonbinary individuals.

Further investigation into the perception of sibilant sounds in nonbinary speech will help expand our knowledge of gender perception beyond binary frameworks. By examining how nonbinary individuals manipulate and produce sibilant sounds, we can gain a deeper understanding of the sociophonetic cues and strategies these individuals employ to convey their gender identities. Such research will contribute to a more comprehensive and inclusive understanding of gender perception than is currently available in relation to sibilant perception.

#### **3.1.4 Listener characteristics**

The perception of gender in speech is influenced by various listener characteristics, including the listener's own gender and other group identities. Individuals often possess pre-existing gender stereotypes that shape their interpretation of sociophonetic cues. Cisgender men and cisgender women have been found to exhibit increased activation in the prefrontal cortex when exposed to voice gender

stimuli that become increasingly ambiguous (Junger et al., 2013). Stimuli that fit into previously constructed categories are processed by the reflexive centers of the brain, while stimuli that do not fit neatly into preconceived categories must be processed by the reflective center, largely the prefrontal cortex (Lieberman, 2007). The increased activation in the prefrontal cortex of cisgender listeners when confronted with nonbinary voices suggests that their voice gender categories may not extend beyond binary conceptions. On the other hand, trans men have demonstrated higher accuracy in identifying "ambiguous" male voices as "male" and exhibited reduced processing load during voice categorization compared to cisgender men (Smith et al., 2018), suggesting an expanded category of "masculine" voice. Dolquist (2023) found that gender expansive listeners were less likely to identify a transmasculine voice as male than cisheterosexual men and women; however, the methods were different from those in the study conducted by Smith et al. (2018) which (1) did not use transmasculine voices in particular but used "male" speech that had acoustic parameters shifted to be "ambiguous"; (2) forced a choice in their participants as opposed to giving their participants the chance to say they could not reasonably guess the person's gender as Dolquist (2023) did; and (3) found this increased sensitivity in trans men specifically while Dolquist (2023) did not look at subcategories of listeners (e.g. did not specifically look at trans men's perception of transmasculine speech but looked at perception in the whole group of gender expansive listeners). These studies highlight the impact of the listener's gender identity and group identity, particularly for gender expansive individuals, on voice gender perception and processing. Furthermore, research has shown that cisgender, queer individuals perceive trans women's speech as significantly more feminine compared to how cisgender, straight individuals perceive

the same voices (Hancock & Pool, 2017). This could be attributed to a broadening of the concept of "feminine" speech and the exposure to a wide range of femininities within the queer community. These findings support the notion that voice gender categories can change or broaden, influenced by the diversity of gender expressions and experiences encountered within different communities (Hope & Lilley, 2022). However, limited investigation has been conducted to understand how gender expansive individuals, particularly those who are nonbinary, perceive gender expansive voices.

To gain a comprehensive understanding of gender perception, it is crucial to explore how gender expansive individuals perceive and interpret gender expansive voices. Examining the experiences, preferences, and biases of gender expansive individuals can provide valuable insights into the diverse ways in which gender is understood and categorized in speech. Further research in this area is necessary to shed light on the nuanced dynamics of gender perception and expand our knowledge beyond traditional binary frameworks. By considering the perspectives of gender expansive individuals, we can foster a more inclusive understanding of gender identity and representation in speech perception.

### **3.1.5 Synthetic speech and gender perception**

The perception of gender in synthetic voices has been explored in several studies. Mullennix et al. (2003) investigated the perception of male and female synthetic voices and examined the effects of listener gender. These authors used preset voices from DECTalk system, which was widely commercially used TTS system. They found that male synthetic voices were preferred over female synthetic voices, particularly by female listeners. The male voices were perceived as more persuasive,

favorable, powerful, softer, and slower compared to the female voices, even with the same rate of synthesized speech. These findings were attributed to the overall better voice quality in the male synthetic voices, with an interaction with higher-level power-dynamic effects of gender. Furthermore, when comparing natural female speech to synthetic female speech, the natural speech from female talkers was preferred. It was perceived as slower, livelier, and less "nasal" compared to the synthetic female voices. In contrast, the male natural and synthetic voices were perceived similarly in all measured respects, indicating less disparity in voice quality between natural and synthetic male voices. This suggests an effect on perceived speaking rate and voice quality in the resulting synthetic female voices.

The preference for male synthetic voices has a historical bases in disparity of research on male versus female voice production. Henton (1999) states:

“There has been little synthesis of the female voice for two reasons. The first reason is there is insufficient data on female speech production. A cross-language survey of phonetic studies, which ostensibly provided 'representative' adult acoustic data, showed that among 42 studies, 30.9% had equal numbers of males and females, 40.5% assembled solely male speakers, a meagre 4.8% had only female speakers, and 21.4% incorporated more males than females (Henton 1986). Just one study (2.4%) incorporated more females than males. A second reason for the comparative lack of female synthetic speech is that female voices historically have been marginalized acoustically (and hence, disregarded in phonetic theory) owing to inadequacies in analytic hardware.” (p. 51-52)

This quote is not only a marker of the past but a warning of the future: what could become the fate for nonbinary synthetic voices. Henton (1999) analyzes factors of the female voice that make the synthesizing process more difficult, which fall out from the reasons stated above; these factors ultimately amount to the more varied speech patterns that women used compared to men and the fact that men tend to have more

reduced forms in speech whereas women tend to use more distinct forms of both vowels and consonants and therefore, those make it more difficult for synthesis to capture.

In our previous study (Hope & Lilley, 2023), a "sibilant goodness" task was conducted to examine listener preferences for sibilant sounds ([s] and [ʃ]) produced by three different synthetic voices: a "female" voice constructed using data from 20 cis female talkers, a "male" voice constructed using data from 20 cis male talkers," and a "nonbinary" voice constructed using data from all 40 talkers and which was labeled as "nonbinary" by nonbinary listeners 100% of the time in an earlier study by the same authors (Hope & Lilley, 2022). The results showed that cisgender listeners more often chose the male synthetic voice as having the best sibilant overall, both in terms preference and exemplary acoustics. In contrast, gender expansive listeners chose the female and male voices about equally. Additionally, the male synthetic voice was rated as having the best [s] by both groups. However, the "nonbinary" synthetic voice was ranked the lowest in terms of the best [s] among the three for both the gender expansive and cisgender groups. It was hypothesized that the relatively low preference for the "nonbinary" voice may have been due to the fact that it was constructed using data from male and female talkers rather than actual nonbinary individuals, potentially lacking authentic sociophonetic cues of nonbinary speech. These findings indicate a general preference for the male synthetic voice, particularly among cisgender listeners.

Furthermore, Přebil et al. (2016) conducted a study evaluating the classification of synthetic speech by gender and age. These authors constructed a voice classification system that could classify a synthetic voice as female or male and as a "child," "young adult," "adult," or "senior". The results showed that the evaluation system classified

male synthetic voices with greater accuracy than female synthetic voices. In another study evaluating synthetic voice, Přibil et al. (2020) created an evaluation system for rating voice quality in synthetic voices. Voice quality in this study was determined as the “similarity [of the synthetic voice] with the original voice by evaluation of features derived from time durations of voiced and unvoiced speech parts” (Přibil et al., 2020, p. 78). In that study, the authors found that the male voices were evaluated as better by their evaluation system than female voices, e.g. the evaluated acoustic parameters were closer to the original raw acoustic measures for the male voices compared to the female voices. The authors noted that the difference in classification might be attributed to the higher variability observed in natural female speech, affecting supra-segmental features (sound intensity and  $f_0$  changes), spectral features, and changes in time duration relations. To address this disparity and improve the synthesis of nonbinary speech, proactive research is needed to understand the sociophonetic encoding of nonbinary identity and how it can be effectively captured and perceived in synthetic speech.

Overall, the studies discussed highlight the preference for male synthetic voices in various aspects, including voice quality, sibilant perception, and in speech synthesis evaluation. Given that this preference for male synthetic voices may reflect a historical and institutional research bias against women’s speech production, to ensure accurate and inclusive representation of nonbinary voices in speech synthesis, it is crucial to investigate and incorporate the sociophonetic cues of nonbinary speech in the development of synthetic voices. Specifically, research should focus on understanding the unique characteristics of nonbinary speech and explore ways to effectively capture and reproduce them in synthetic speech systems.

### 3.1.6 Questions and hypotheses

Gender perception in speech is a multifaceted process influenced by various sociophonetic cues, including fundamental frequency, formant frequencies, and acoustic properties of sibilants. More research needs to be conducted on nonbinary individuals to capture a fuller range of diversity in gender expression in voice. In a similar vein, a listener's characteristics also influence voice gender perception. Therefore, it is essential to investigate how different individuals' experiences of gender may interact with perception of gender in speech. In particular, perception of gender in gender expansive synthetic voices has not been explored. This next study looked at gender perception in four different gender expansive synthetic voices created from the speech of 16 gender expansive talkers.

Our questions and hypotheses were:

1. How do cisgender and gender expansive individuals perceive gender in synthetic speech produced from data for gender expansive individuals? In particular, do gender expansive individuals have different “other” gender perception of these voices compared to cisgender individuals? We anticipated that cisgender and gender expansive individuals would perceive the gender of the gender expansive synthetic voices differently.
2. Do cisgender and gender expansive people perceive the four voices differently? We hypothesized that the nonbinary and gender expansive synthetic voices would be perceived as significantly more “other” gender compared to the transmasculine and transfeminine voices which we anticipated would be rated more in a “binary” manner.
3. Do those with greater other gender perceive voices as more “other” gender? We hypothesized that the greater “other” gender a participant identifies with, the greater they would perceive the voices as “other” gender (e.g. not as male or female or as feminine or masculine).

## 3.2 Methods and Materials

### 3.2.1 Synthetic Voice Creation

Four voices (one created with data from all talkers in the study illustrated in Chapter 2 – ALL, one from 8 nonbinary participants – NB, one from 4 transfeminine participants – TF, one from 4 transmasculine participants – TM) were built using the standard Merlin DNN synthesis process (Wu et al., 2016) in which two DNN models were trained as follows: (1) a duration model that takes 229 linguistic features as input, and predicts the number of 5-msec frames per phone as output; and (2) an acoustic model with the same linguistic features as input, and a set of 187 acoustic features per frame as output, which include 180 mel-generalized coefficients (MGC), 3 band aperiodicity (BAP) features, 3  $\log f_0$  features, and voicing. The recordings of all 16 talkers were used to train the ALL model, the 8 nonbinary talkers for the NB model, the 4 transfeminine talkers for the TF model, and the 4 transmasculine talkers for the TM model. The variable sample size across groups is due to convenience sampling. All training used Theano (Al-Rfou et al. 2016). All DNN models used six fully connected layers of 1536 units apiece with tanh activation and no dropout. Training used stochastic gradient descent (initial learning rate 0.002 with exponential decay), with 10 warmup and 30 training epochs (25 for  $f_0$  models). Batch sizes were 64 for duration models and 256 for other models. For synthesis we used the WORLD vocoder (Morise et al., 2016). For the four voices, we used the trained duration models and matching acoustic models to generate the WORLD vocoder features.

The WORLD vocoder is known for its speech synthesis capabilities, especially in real-time applications. The system employs advanced signal processing techniques to analyze and synthesize speech signals, allowing for natural and intelligible output.

These synthetic voices are often used specifically for speech-generating devices. The WORLD vocoder software includes both the tools for parameterizing the speech signal and for reconstructing it from the parametric representation. The input (raw) speech signal is analyzed to extract its  $f_0$ , spectral envelope, and aperiodic components. This process develops a parametric description of the acoustic signal in terms of a sequence of time-varying measurements or parameters.  $f_0$  captures the rate of vocal fold vibration, while the spectral envelope captures the shape of the vocal tract and the vocal fold output, and the aperiodic components represent non-harmonic components, such as those occurring with breathiness or frication. The derived parameters can then be used for parametric synthesis using the WORLD vocoder. The synthesis process involves generating a waveform based on  $f_0$ , the spectral envelope, and aperiodic components obtained from the analysis stage. As a result of this process, synthesized words, phrases, and sentences can be generated.

### 3.2.2 Sentence/Paragraph generation

Each voice was made to generate two different Harvard sentences resulting in eight different Harvard sentences (IEEE, 1969). Each voice was also made to produce a slightly modified version of the Rainbow Passage (Fairbanks, 1960).

Fundamental frequency averages and standard deviations for each of the sentences and passages are in Table 3.1.

Table 3.1 Fundamental frequency ( $f_0$ ) averages and standard deviations for each of the sentences and passages for each synthetic voice.

<b>SENTENCE/PASSAGE</b>	<b>VOICE</b>	<b>AVG<math>f_0</math></b>	<b>SD<math>f_0</math></b>
The brown house was on fire to the attic.	NB	151.5	21.8

A fresh start will work such wonders.	NB	148.6	21.4
Cut the cord that binds the box tightly.	TM	129.9	20.1
New pants lack cuffs and pockets.	TM	147.3	18.1
Try to trace the fine lines of the painting.	TF	128.3	17.3
Throw out the used paper cup and plate.	TF	127.8	13.8
Carry the pail to the wall and spill it there.	ALL	140.7	19.1
The fur of cats goes by many names.	ALL	141.9	21.4
Rainbow Passage	NB	145.1	14.9
Rainbow Passage	TM	130.0	16.4
Rainbow Passage	TF	122.3	12.7
Rainbow Passage	ALL	137.4	15.7

The average Vowel Space Dispersion (VSD) for each of the four voices is shown in Table 3.2. VSD could not be accurately calculated for each sentence individually due to lack of sufficient vowel tokens.

Table 3.2 Average vowel space dispersion (VSD) for the four voices calculated by averaging the Euclidian distance of vowels [i],[u],[a],[o],[e] from all the stimuli used for that voice to center of the vowel space

<b>VOICE</b>	<b>VSD</b>
<b>ALL</b>	546.0 Hz
<b>NB</b>	606.6 Hz
<b>TF</b>	527.7 Hz
<b>TM</b>	612.8 Hz

### 3.2.3 Participants

Forty listener participants were recruited via the website Prolific ([www.prolific.com](http://www.prolific.com)) and participated in an online speech perception experiment that was approved by the IRB of the University of Delaware (see Appendix B). There were two rounds of recruitment; the first was set to allow 20 nonbinary individuals and the second was set to allow 20 men and women according to Prolific's identity settings, allowing both cisgender and transgender participants. Those who participated in the first round were excluded from participation in the second round. For both rounds, Prolific was set up to have a pre-screen that admitted only the participants who disclosed they were between the ages of 18 and 65 years, that they were fluent talkers of English, and that the place where they spent the most time before they turned 18 was in the United States. The participants' ages ranged from 18 to 61 years ( $mean = 29.7$ ,  $SD = 9.6$ ) and the racial identities that people used to describe themselves in response to an open-ended question were distributed as follows: 28 white, 2 Black, 2 Multicultural/mixed, 2 Latinx, 2 Hispanic, 2 Asian, 1 Native American, 1 Native American/mixed. Eighteen of the total participant pool were gender expansive (GE); that is, they were transgender and/or nonbinary (the other 22 were cisgender). Participants were compensated \$15.49/hour for completing the experiment, as set by Prolific. The experiment took a median completion time of 30 minutes.

### 3.2.4 Procedures

The experiment was built and run through Gorilla (Anwyl-Irvine et al., 2020). The first part of the experiment involved a comprehensive demographic survey that asked questions about the participants' age, race, and gender. The survey also included six scales of gender corresponding to three continuous scales (0 to 100) of gender

identity (female, male, and other gender) and three similar scales of gender expression (feminine, masculine, and other gender).

Participants listened to two different Harvard Sentences produced by each of the voices, for a total of eight sentences, and gave subjective gender ratings of masculine, feminine, and other gender perception. Finally, participants heard each voice produce the first five sentences of the Rainbow Passage, and gave an open-ended description of what the voice sounded like to them, which could include any features of the voice and was not limited to gender (e.g., age, race, class, education, emotional state, intelligence, etc.). Participants also completed a second experiment identifying sibilant fricatives, which is reported in Chapter 4.

### 3.2.5 Statistical analyses

Statistical analyses including 3 repeated-measures ANOVAs and a linear mixed model were conducted in R using functions `aov` from base R and `lmer` from `lme4`, respectively (Bates et al., 2015). Shapiro-Wilk normality tests and Bartlett tests of homogeneity of variances were also conducted in base R using `shapiro.test` and `bartlett.test`. To answer the question of if the four voices are perceived differently from each other, we conducted our repeated-measures ANOVAs between the gender perceptions of the voices for all participants:

$$Gender\ perception \sim Voice * Group + Error\left(\frac{Participant}{Voice}\right),$$

where *Gender perception* is encoded as a whole number between 0 and 100, *Voice* is a categorical variable with four categories (ALL representing the voice created from all gender expansive talkers, NonBin representing the voice created from the nonbinary talkers, TF representing the voice created from the transfeminine talkers,

and TM representing the voice created from the transmasculine talkers), and *GE* is a binary variable GE vs. Cis (encoded 1 vs. 0 respectively).

To answer the question of whether listeners with higher other gender identity perceive “other” gender to a greater degree than other listeners, a mixed linear regression was computed between *Other gender identity* of the listener and *Other gender perception* of the voices:

$$\textit{Other gender perception} \sim \textit{Other gender identity} + (1|\textit{Participant}),$$
where *Other gender perception* and *Other gender identity* were encoded as whole numbers between 0 and 100

In a similar vein to the statistical analyses in Chapter 2, although this is not an exploratory study, given the small sample size and preliminary nature of the work, we will adapt an alpha of 0.1 to again reduce the potential for Type II error (Roy et al., 2004; Schumm et al., 2013).

### **3.3 Results**

#### **3.3.1 Differences between voices**

Figure 3.1. shows the *Feminine*, *Masculine*, and *Other gender perception* of the four synthetic voices for the whole group of listener participants.

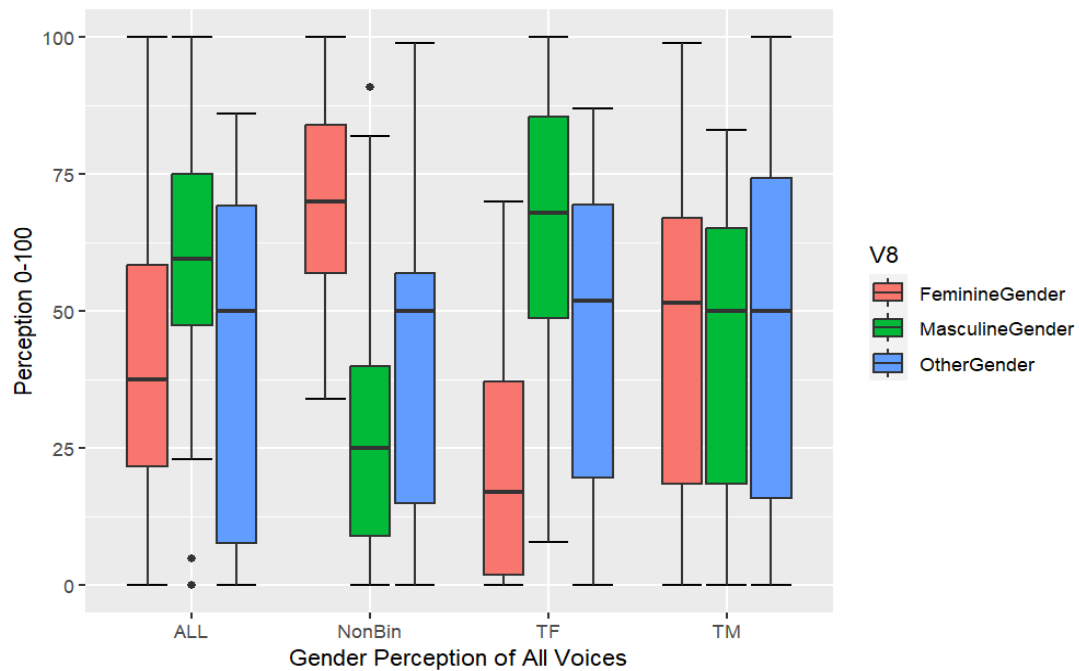


Figure 3.1. Perceptual ratings collapsed across gender expansive and cisgender groups. The x-axis shows the four different synthetic voices (ALL representing the voice created from all gender expansive talkers, NonBin representing the voice created from the nonbinary talkers, TF representing the voice created from the transfeminine talkers, and TM representing the voice created from the transmasculine talkers). The y-axis shows the perceptual gender rating on a scale of 0-100. Colors indicate gender of perception with *Feminine gender* in red, *Masculine gender* in the green and *Other gender* in the blue.

Figure 3.1 above shows that the four synthetic voices had large variabilities in gender perception scores across listeners represented by the relatively large boxes and lines of the boxplots. Both the ALL voice and the TF voice were perceived as more masculinely gendered compared to the NonBin and TM voices. The NonBin voice had the highest *Feminine gender perception*. The TM voice had very similar median scores across the three gender perception scales, all around 50 on a scale of 0 to 100.

Figure 3.2. below shows the gender perception (*Masculine*, *Feminine*, and *Other*) of the synthetic voice constructed from all 16 gender expansive talkers.

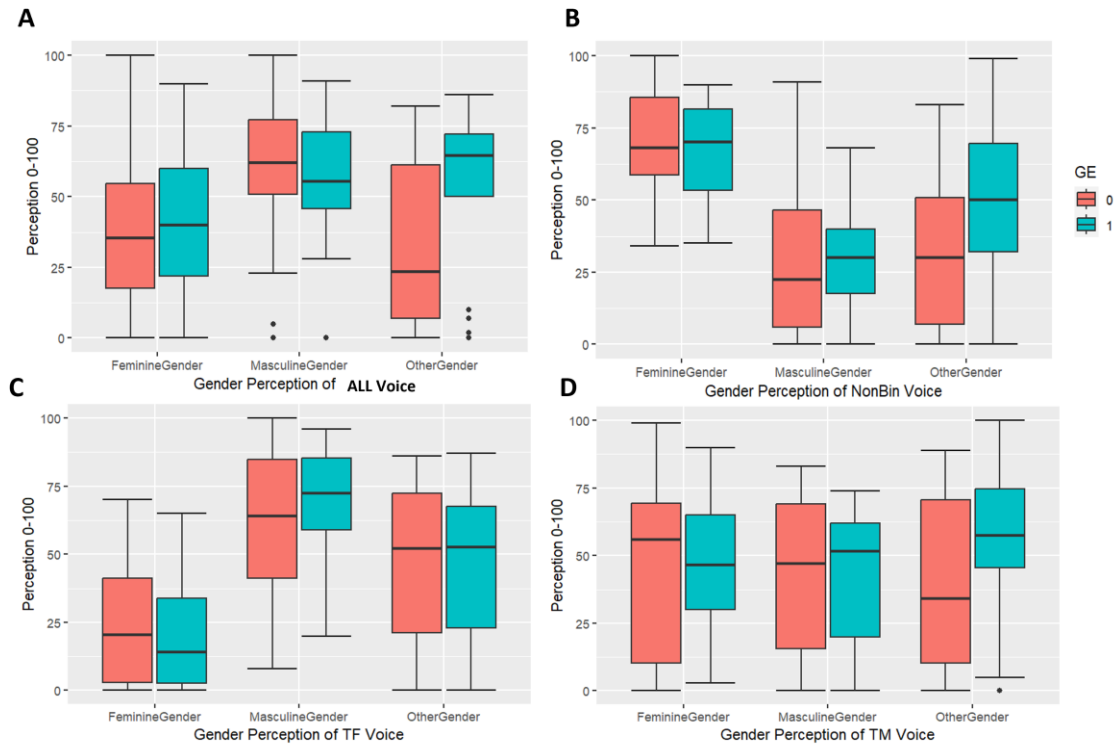


Figure 3.2. Boxplots illustrating differences in perception for the ALL voice made from all 16 GE talkers (A), the NonBin voice made from the eight nonbinary talkers (B), the TF voice made from the four transfeminine talkers (C), and TM voice made from the four transmasculine talkers (D), between groups where red boxes indicate the cisgender listener group and blue boxes indicate the gender expansive listener group.

The figure above illustrates that the cisgender and gender expansive listeners have similar *Feminine* and *Masculine* gender perception of the ALL synthetic voice while these groups differ in *Other* gender perception. The GE group shows

statistically significantly higher *Other gender perception* for the ALL voice. The GE and cisgender groups do not show statistically significant differences in *Feminine* or *Masculine gender perception* for the NonBin voice. However, there is a statistically significant difference between GE and cisgender listeners for *Other gender perception* of the NonBin synthetic voice. There were no statistically significant differences between groups for gender perception for the TF voice. Additionally, the TF voice had higher *Masculine gender perception* and lower *Feminine gender perception*. There were no statistically significant differences between groups for gender perception for the TM voice. The TM voice also had very similar median gender perception rating across the three scales, all roughly around 50 except for the other gender perception for cisgender group which is lower than the GE group, but not statistically significantly lower.

Three repeated-measures ANOVAs were used to analyze the effect of *Voice* (e.g. which of the four synthetic voices the participants heard) and *Group* (GE vs cis) on *Gender perception* (for *Other gender perception*, *Feminine gender perception* and *Masculine gender perception*). We generated Q-Q plots of the residuals for all 3 ANOVAs and found that all residuals fell roughly along a straight, diagonal line. We then conducted a Shapiro-Wilk normality test for each ANOVA with all p-values greater than 0.05. Finally, we conducted the Bartlett test of homogeneity of variances for all ANOVAs and again found all p-values to be greater than 0.05. We therefore concluded that assumptions of normality of distribution and homogeneity of variances were met. The assumption of independence of observations was met due to randomized design.

A repeated-measures ANOVA was performed to analyze the effect of *Voice* (e.g. which of the four synthetic voices the participants heard) and *Group* (GE vs cis) on *Other gender perception*. The ANOVA revealed that there was not a statistically significant interaction between the effects of *Voice* and *Group* ( $p = 0.121$ ). Simple main effects analysis showed that *Group* did have a statistically significant effect on *Other gender perception* ( $p = 0.055$ ). Simple main effects analysis showed that *Voice* did not have a statistically significant effect on *Other gender perception* ( $p = 0.374$ ).

A repeated-measures ANOVA was performed to analyze the effect of *Voice* and *Group* on *Feminine perception*. The ANOVA revealed that there was not statistically significant interaction between the effects of *Voice* and *Group* ( $p = 0.83$ ). Simple main effects analysis showed that *Group* did not have a statistically significant effect on *Feminine perception* ( $p = 0.947$ ). Simple main effects analysis showed that *Voice* did have a statistically significant effect on *Feminine perception* ( $p < 0.001$ ).

A repeated-measures ANOVA was performed to analyze the effect of *Voice* and *Group* on *Masculine perception*. The ANOVA revealed that there was not statistically significant interaction between the effects of *Voice* and *Group* ( $p = 0.797$ ). Simple main effects analysis showed that *Group* did not have a statistically significant effect on *Masculine perception* ( $p = 0.943$ ). Simple main effects analysis showed that *Voice* did have a statistically significant effect on *Masculine perception* ( $p < 0.001$ ).

### **3.3.2 Gradient gender perception of the voices**

There was a significant linear regression between *Other gender identity* and *Other perception*.

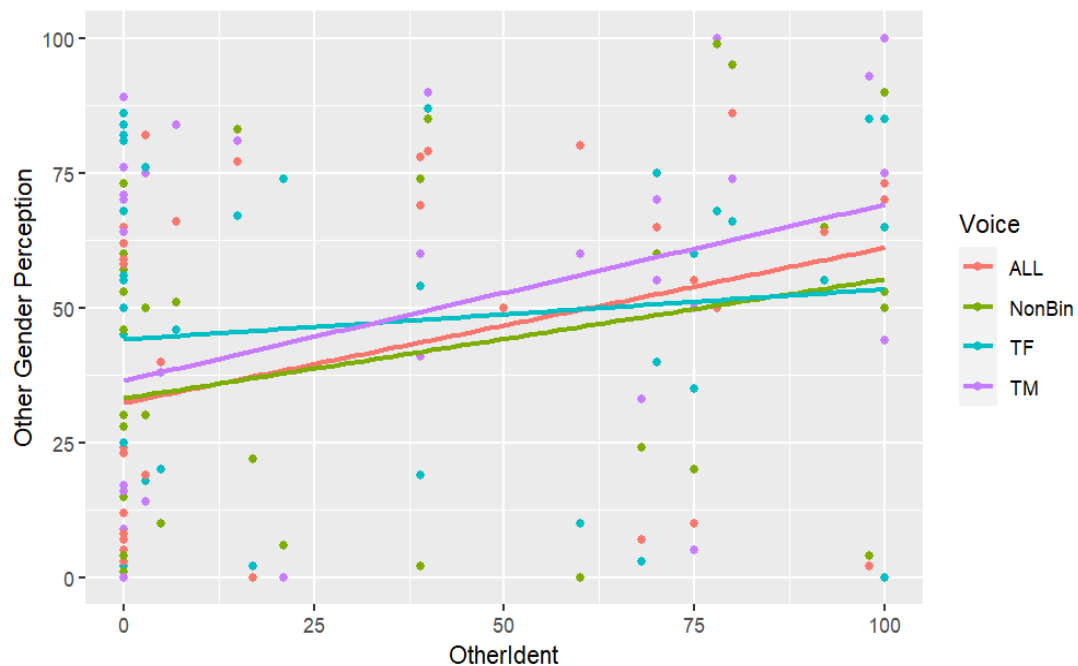


Figure 3.3 Linear regression demonstrating the relationship between *Other gender identity* of the listener (x-axis) and *Other gender perception* of the voices (y-axis). The red line indicates the voice made from all GE talkers, the green line represents the voice made from the eight nonbinary talkers, the blue line indicates the voice made from the four transfeminine talkers, and the purple line indicates the voice made from the four transmasculine talkers.

The figure above illustrates the relationship between *Other gender identity* and *Other gender perception*; this relationship was positive such that as the listener had greater *Other gender identity*, their perception of *Other gender* of the voices increased. The random effects analysis revealed two variance components: the intercept at the participant level (*Part*) demonstrated a variance of 381.5 ( $SD = 19.53$ ), while the residual variance was estimated at 465.0 ( $SD = 21.56$ ). The dataset comprised 160 observations nested within 40 groups (*Part*). Turning to fixed effects, the model indicated a statistically significant intercept at 36.51995 ( $t\ value = 7.629, p <$

0.001), and a significant effect for the variable *Other gender identity* with an estimate of 0.23431 ( $t$  value = 2.544,  $p$  = 0.0152). The correlation of fixed effects demonstrated a negative correlation ( $r$  = -0.676) between the intercept and *Other gender identity*.

### 3.4 Discussion and Conclusion

The study findings revealed significant distinctions in gender perception between cisgender listeners and listeners with gender expansive identities, particularly for the gender expansive (ALL) and nonbinary (NB) voices. While the transmasculine (TM) voice didn't show a statistically significant difference in gender perception between the listener groups, this difference approached statistical significance in other gender perceptions. Surprisingly, the transfeminine (TF) voice, contrary to initial hypotheses, demonstrated no significant listener group differences in any gender perception, despite having the highest masculinity ratings among the four voices. This implies that the way transfeminine talkers encoded gender in their speech did not align with prescriptive binary gender norms. The transmasculine voice, with more balanced ratings across gender scales, appeared to be perceived as more ambiguous, as it did not strongly lean towards masculinity or femininity on the 0-100 scale. However, there were no significant differences between listener groups for *Feminine* or *Masculine perception* of any of the four voices.

There was a statistically significant correlation between *Other gender* of the listener and perceived *Other gender* of the synthetic voices. Additionally, cisgender individuals exhibited a more limited use of the *Other* scale for their own identity, implying that separate modeling of gradient gender perception for cis and gender

expansive groups may be advisable. Ultimately, GE and cisgender individuals perceived the synthetic voices differently.

## Chapter 4

### CATEGORIZATION OF THE GENDER EXPANSIVE SYNTHETIC SIBILANTS: DECOUPLING OF [S] PERCEPTION AND GENDER

Part of this paper was published in *Seminars in Speech and Language* published by Thieme Group. The citation for the published paper is noted in Appendix A.

#### 4.1 Introduction

Much of the past research on the sociophonetics of sibilant perception has examined men's and women's voices, without specifying the sociodemographics of the men and women talkers (Strand & Johnson, 1996; Munson, 2011; Winn & Moore, 2020). Given the sociopolitical context in which these studies were conducted, it is likely that these were cisgender (adult) men and women. This supposition is supported by the fact that these articles characterized talkers using acoustic parameters that distinguish between cisgender (adult) men and women's voices robustly, such as fundamental frequency ( $f_0$ ) and average formant frequencies. For example, both Munson (2011) and Winn and Moore (2020) described stimuli as being in "typical" ranges. Munson et al. (2017) and Bouavichith et al. (2019) created stimuli with an ostensibly "neutral"  $f_0$  or "average" formant frequencies, however, these were defined by the values of (presumably) cisgender men and women. These studies used the same sibilant continuum across vocal tract conditions, with the goal of eliciting a compensation effect, that is, a shift in the boundary of the s-sh continuum when gender is suggested by a visual cue.

Munson, Winn and Moore, and others have not considered how gender-compensation effects might play out when the stimuli are produced by someone who does not meet these cisgender expectations. However, Munson et al. (2006) did investigate sibilant perception of talkers who varied in their adherence to cisgender norms; their talkers included both heterosexual individuals and those who belonged to the GLB (Gay Lesbian Bisexual) community, all of whom were cisgender. The findings showed that the listeners perceived [s] more often for the stimuli that came from women who were perceived as lesbian/bisexual than for the stimuli that came from heterosexually-perceived women. Thus, there exists a starting point for investigating this type of sibilant perception outside of a cisheteronormative framework, although transgender and nonbinary talkers have yet to be explicitly studied. The study presented here addressed this question by examining the perception of [s]-[ʃ] continua combined with vowels taken from synthetic voices created from gender expansive (GE) individuals. Because GE individuals often defy binary gender categorization and encode multiple different gender cues in speech, it is especially worthwhile to conduct such a study using stimuli from these individuals. Doing so will enable us to better understand how [s]-perception and gender intersect both when talkers decouple those cues, and how listeners perceive these decouple cues given that these cues are thought to be very coupled in the literature.

Studies such as those by Strand and Johnson (1999), Munson (2011) and Winn and Moore (2020) also neglected to examine how listeners' genders affect their perception of sibilant continua thereby obscuring possible effects of listener differences in categorization that arise from differences in how various listener groups encode gender in speech. Individuals' own gender is known to intersect with gender-

associated acoustic cues in speech perception (Smith et al., 2018, Hope & Lilley, 2022), whereby people's judgements of voice gender are influenced by their own gender identities. These gender differences in the listener have been thought to be motivated by the repeated adherence to rules associating certain acoustic cues with binary cisnormative categorization (e.g. low  $f_0$ s are associated continuously with male/masculine gender) or the repeated disassociation between these acoustic cues and binary cisnormative categorization (e.g. low  $f_0$ s are associated with a wide range of genders instead of just male/masculine genders) (Hope & Lilley, 2022).

Recent research we conducted has shown that a sibilant continuum taken from an "average" synthetic female voice, constructed from 20 presumably cisgender female talkers, and a sibilant continuum taken from a synthetic "nonbinary" voice, constructed from the same 20 female talkers and 20 presumably cisgender male talkers, elicit higher proportion [s] responses from GE listeners compared to cisgender listeners (Hope & Lilley, 2023). Figure 4.1 is taken from that experiment and shows the differences in sibilant categorization between the cisgender and the GE listeners. One potential reason for the higher proportion of [s] responses found was that the intermediate [s] sounds were already [s]-like "enough" for the GE listeners whereas the cisgender listeners were expecting [s] productions with much higher centers of gravity or peak frequencies and therefore categorized productions less often as [s].

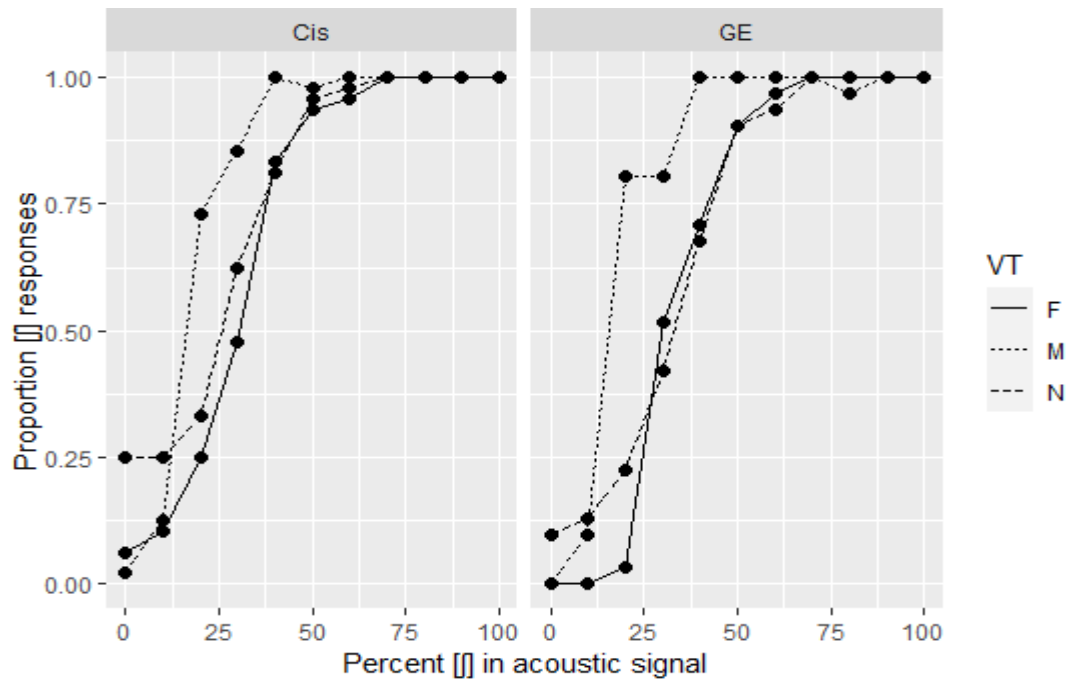


Figure 4.1 Sibilant categorization curves showing the proportion of [ʃ] responses by cisgender (Cis, left) and gender expansive (GE, right) listeners for three different synthetic voices (F: a female synthetic voice, M: a male synthetic voice, and N: a nonbinary synthetic voice).

In this experiment reflected in Figure 4.1 above, we did not use actual GE voices in the creation of our nonbinary stimuli but relied on the integration of data from female and male (presumably cisgender) talkers (Hope & Lilley, 2023). The present study aimed to examine how GE and cisgender individuals categorize synthetic sibilants across different sibilant continua and vocal tracts taken from four different synthetic voices created from GE talker data. GE talkers defy binary categorization; using stylistic bricolage as expounded by Zimman (2017), GE talkers do not couple [s] production and other vocal cues in cisheteronormative patterns, and this practice could influence perception. The four voices in this present study, which are the same as those in Chapter 3, have formant frequencies and spectral information

taken from four different synthetic voices (a "gender expansive" voice, a "nonbinary" voice, a "transmasculine" voice, and a "transfeminine" voice). Because the present study addressed differences in speech perception between gender expansive individuals and cisgender individuals and used synthetic voices that could be used in speech-generating devices (SGDs), background and applications of this research are discussed in the next paragraphs in light of trans and nonbinary voice therapy, as well as the development of more authentic GE synthetic voices than are currently available. By integrating considerations of language and gender, we can elucidate the intricate mechanisms that govern the process of speech perception.

#### **4.1.1 Gender and sibilant production in English**

Sibilant sound production, specifically [s] and [ʃ] differs across genders (Flipsen et al., 1999; Jongman, et al., 2000; Fuchs & Toda, 2010). In much of the previous research, it was found that women tend to produce [s] with the tongue closer to the teeth than men, resulting in higher sibilant frequencies (Flipsen et al., 1999; Jongman, et al., 2000; Fuchs & Toda, 2010). To elucidate this point, we present two figures below. Figure 4.2 below shows a production of the word "so" with the tongue first closer to the teeth and then Figure 4.3 shows the tongue more retracted with acoustic effects on the spectrogram and spectrum to highlight the relative differences in frequencies. One concern with the previous research in this regard is that the groups of women and men who were investigated in terms of sibilant production represented only a small minority of women and men and often represented those in dominant sociocultural groups, therefore generalizations about women's and men's speech as a whole based on these studies should be made with caution.

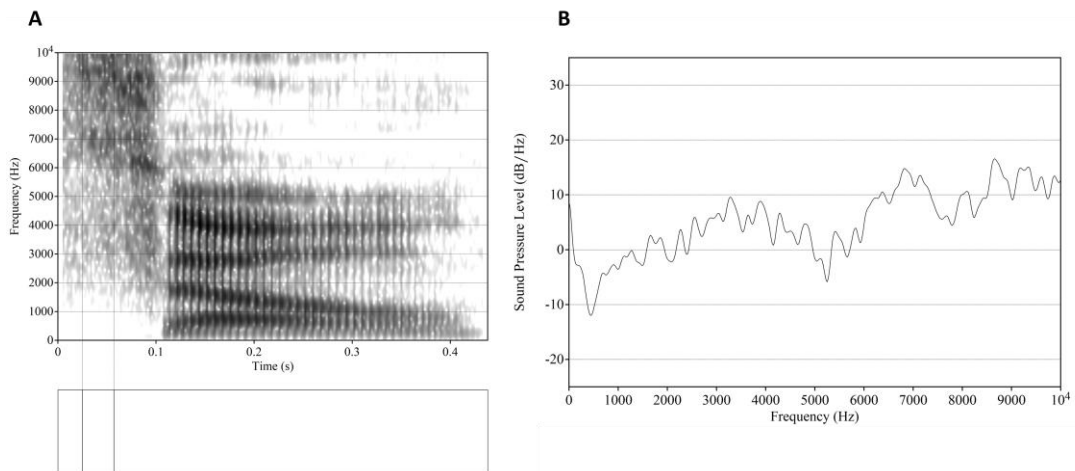


Figure 4.2 Production of the word “so” by the author with the tongue more forward in the mouth compared to that in figure 4.3. The A figure shows the spectrogram of the fricative and vowel and the B figure shows the spectrum of the fricative production.

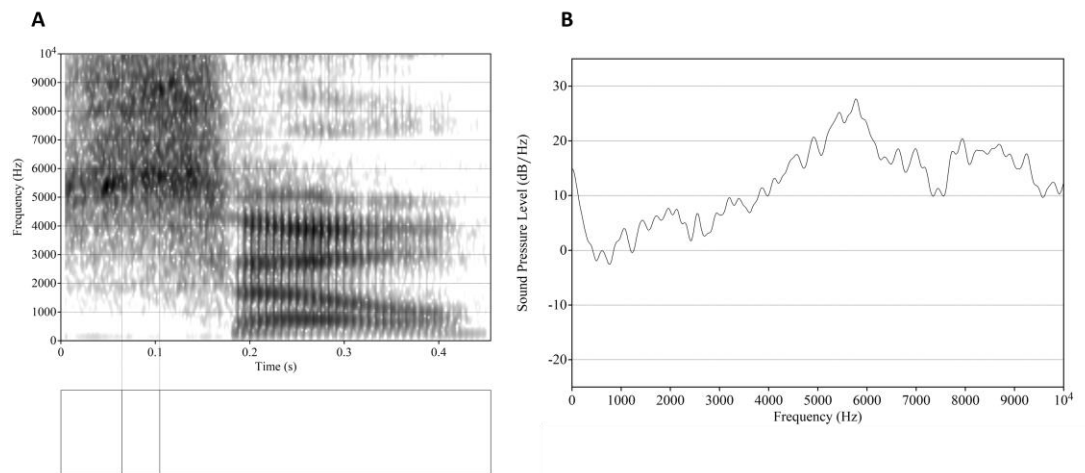


Figure 4.3 A production of the word “so” by the author with the tongue in a more retracted position compared to that in 4.2. The A figure shows the spectrogram of the fricative and vowel and the B figure shows the spectrum of the fricative productions.

As demonstrated in Figures 4.2 and 4.3, the placement of the tongue to the anterior teeth in [s] production influences the sibilants' frequencies. Both the average of the frequencies of a segment weighted by their amplitudes, known as Center of Gravity (COG) or spectral mean, and peak frequency demonstrate a significant relationship with the distance between the teeth and tongue, although the correlation is stronger for COG (Fuchs & Toda, 2010). As the distance between the teeth and tongue increases for [s] production, the COG decreases significantly (Fuchs & Toda, 2010). Men's COG for [s] is generally lower (around 5,632 Hz, range of 4,757–6,167) than women's COG (around 6,412 Hz, range: 5,727–6,858 Hz) (Fuchs & Toda, 2010). It is possible that because of these general trends, gender differences in [s]-perception reflect gender-specific places of articulation for [s] that are likely learned. However, as noted, one criticism with these studies is that they typically investigated socioculturally dominant groups and didn't represent a diverse range of talkers; therefore, the resulting generalizations about women's or men's speech with respect to [s] become shaky. It is important to continuously expand on previous research by collecting more inclusive data and re-assess old assumptions. A study by Zimman (2017) on trans masculine individuals found that COG varied significantly across participants, and that those who identified more strongly as men tended to have lower COG than those who identified as genderqueer or nonbinary. Zimman (2017) also noted that there was substantial variation in production both within and between subjects and that the relationship between masculinity and COG was more complex than it initially appeared on the surface. Specifically, some queer trans men used low  $f_0$  and high COG together as "stylistic bricolage" to signal their queer masculinity (Zimman, 2017).

Despite the variation observed within different social groups, COG tends to differ in systematic ways across genders, and this is especially true for those who fall into binary cisheteronormative categories. Moreover, vowel acoustics have been found to encode gender characteristics, and both spectral information for [s] and vowel acoustics, such as those derived from vocal tract length, have been shown to influence sibilant perception (Munson, 2011). However, in our previous study (Chapter 2), we analyzed the correlations between various acoustic parameters of 16 GE talkers including COG and peak frequency of [s] with self-reported gender on three independent gradient scales of feminine, masculine, and other gender identity, and three additional independent gradient scales of feminine, masculine and other gender expression. We found that there was no statistically significant correlation between gender (where gender was assessed by six gradient variables of identity and expression) and COG or peak frequency of [s]. We also found no significant differences in sibilant production between sub-groups of GE talkers; e.g., when we looked at GE women, GE men and GE nonbinary individuals, there were no statistically significant differences between groups for [s] COG or peak frequency. This suggests that the associations between gender and /s/ acoustics that underlie Strand and Johnson's findings do not apply to GE individuals. This finding leads us to hypothesize that gender compensation in sibilant perception might not occur for GE listeners. This possibility is especially likely for GE listeners if they have more lived experience perceiving GE talkers like those from Chapter 2.

#### **4.1.2 Gender and sibilant perception in English**

In a study by Munson (2011), which reconfirmed findings of Strand and Johnson (1996), listeners categorized ambiguous sibilant stimuli paired with images of

stereotypically male or female faces. Results showed that when listeners heard a natural man's or woman's voice manipulated to have a shorter vocal tract, they were more likely to categorize the sibilant as [ʃ], particularly when paired with a woman's face. However, when a man's or woman's voice was manipulated to have a longer vocal tract, listeners were more likely to categorize the sibilant as [s], regardless of the gender of the face they saw. The absence of face priming in that condition might reflect the failure of the faces to prime gender when the vocal-tract cues were unambiguous - a condition that only works when one has a cisheteronormative model of gender in one's mind. This suggests that vocal tract length influences sibilant categorization when the spectrum of the sibilants remains ambiguous across vocal tract conditions, but that this perceptual effect may be limited to those with cisheteronormative voice genders.

Similarly, in a study by Winn and Moore (2020), the authors manipulated  $f_0$  and vocal tract length to investigate their effects on sibilant categorization. However, the  $f_0$  conditions used were limited to a mean  $f_0$  of 104 Hz and a mean  $f_0$  of 208 Hz, reflecting "male" and "female" conditions, respectively. Additionally, the authors manipulated vocal tract length to reflect "masculine" and "feminine" vocal tract lengths, but did not use a "gender-neutral" pitch or a vocal tract length in the middle of the "male" and "female" vocal tract lengths. As such, sibilant perception and categorization may be influenced by social factors such as the perceived gender of the talker, but there has been little investigation of sibilant perception outside of a cisheteronormative framework.

Additionally, sibilant perception is reliant on factors such as listener characteristics which may help us to further understand the processes by which people encode and perceive gender in speech.

#### **4.1.3 Listener characteristics affect sibilant perception**

The talker can influence or sway sibilant perception based on their gender; however, listener characteristics also play a role in sibilant perception. Because sibilants in English sociophonetically encode gender, then gender may be one listener characteristic that affects sibilant perception. This idea also relates to social power, which is defined as “an individual’s relative capacity to modify others’ states by providing or withholding resources or administering punishments” (Keltner et al., 2003). Calloway (2021) demonstrated in a sibilant identification task the influence of auditory and visual cues on sibilant categorization and the effects of social power, e.g. the ability to influence others and make decisions due to their position in societal structure, on social information processing during lexical categorization. The findings indicated that listeners tended to choose /s/ more often when provided with cues suggesting that the talker is male (e.g. someone with higher social power). Additionally, the study suggested that listeners in the high-power condition process social information differently than those in the low-power condition. While both groups showed sensitivity to auditory cues, high-power listeners were less likely to categorize a sibilant as /s/ when a male voice was paired with a female face. These results align with the hypothesis that individuals in high-power conditions tend to rely on single, congruent cues for categorization, whereas individuals in low-power conditions are more sensitive to multiple incongruent cues.

These power dynamics may be one aspect at play in other group differences. Other group dynamics that intersect with relative societal power differentials may show differences in single versus multiple cue integration. For example, GE individuals actively defy binary categorization; this binary categorization was built on congruent gender cues, but many GE individuals do not fit into binary boxes or do not want to uphold pre-existing gender norms. Through their lived experiences, they challenge the relative power state. This challenging of gender norms plays out in sibilant production and perception. Because we have not found the same trends in sibilant production in GE participants as in cisgender studies in the past, perception of gender expansive sibilants may be affected by their lack of conformity to gender norms in speech. GE listeners likely decouple [s] and gender for GE talkers due to the difference in production of [s] and lack of systematic encoding of gender for GE individuals. So far, there has not been much investigation into differences in sibilant perception between cisgender and gender expansive listeners. However, in our previous study (Hope & Lilley, 2023), we created three different synthetic voices and used them to test differences in sibilant perception in a “see”-“she” categorization task. The voices were sourced from either (1) a “male” vocal tract, (2) a “female” vocal tract, or (3) a “nonbinary” vocal tract (created from 20 male and 20 female talkers, but which was categorized as nonbinary 100% of the time in work by Hope and Lilley [2022]). The results showed that GE participants heard “see” significantly more of the time compared to cisgender listeners, especially towards the [s] end of the continuum for both the “female” and “nonbinary” voices. This tells us that differences do indeed exist between GE and cisgender listeners.

Despite the findings of Hope and Lilley (2023), there were several issues with the experimental design. First, we used [i] as the vowel, which shifts perception towards a [ʃ] perception (Winn et al., 2013). Second, we used 20 male and 20 female talkers who were all presumably cisgender for the creation of the three voices; these voices therefore may not encode gender expansive sociophonetic cues of speech since they were averages of (presumably) cisgender talkers. Finally, we used sibilant continua and vocal tracts taken from the same voice and therefore did not isolate the effects of sibilant or vocal tract. Figures 4.4, 4.5, and 4.6 below show the spectrograms and spectra of the different stimuli used by Hope and Lilley (2023).

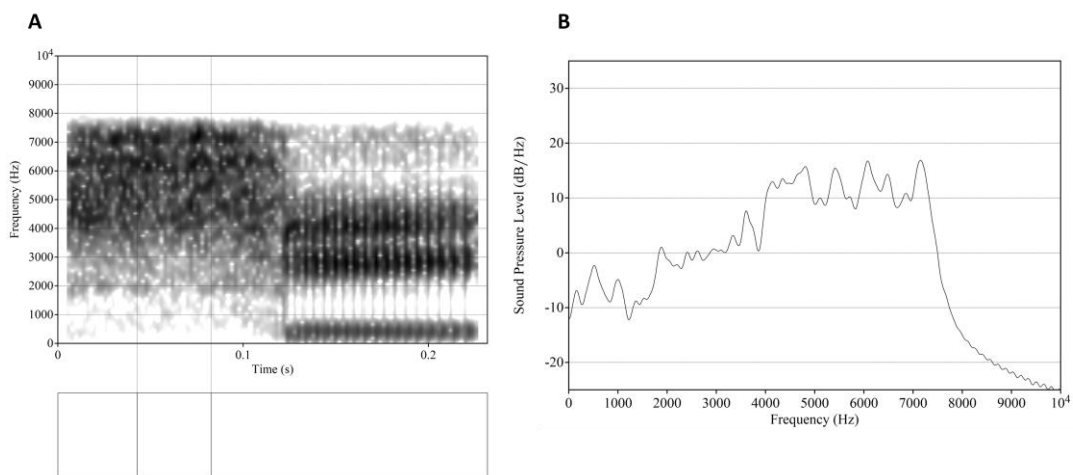


Figure 4.4 Spectrogram and spectrum from the “female” vocal tract “see” token from Hope and Lilley (2023)

Figure 4.4. above shows the spectrogram (A) of the word "see" produced by the “female” synthetic voice described by Hope and Lilley (2023) and spectrum (B) of the [s] from the word “see” on the left. The spectrum shows the middle 40 ms of the

[s] production. The peak frequency shown is around 7000 Hz and in general, the higher frequencies have higher sound pressure levels.

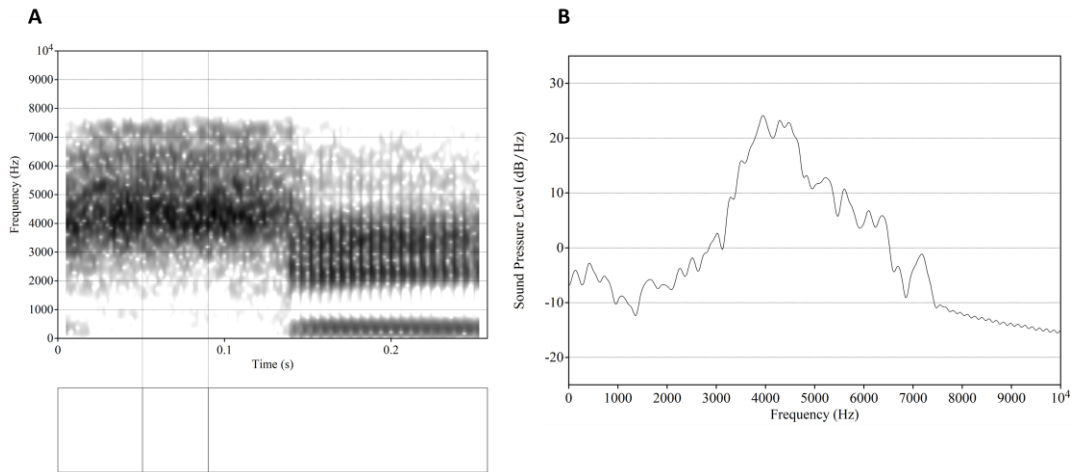


Figure 4.5 Spectrogram and spectrum from the “male” vocal tract “see” token from Hope and Lilley (2023)

Figure 4.5. above shows the spectrogram (A) of the word "see" produced by the “male” synthetic voice described in Hope and Lilley (2023) and spectrum (B) of the [s] from the word “see” on the left. The spectrum shows the middle 40 ms of the [s]. The peak frequency is around 4000 Hz and the highest sound pressure levels occur for the frequencies between 4000 and 5000 Hz.

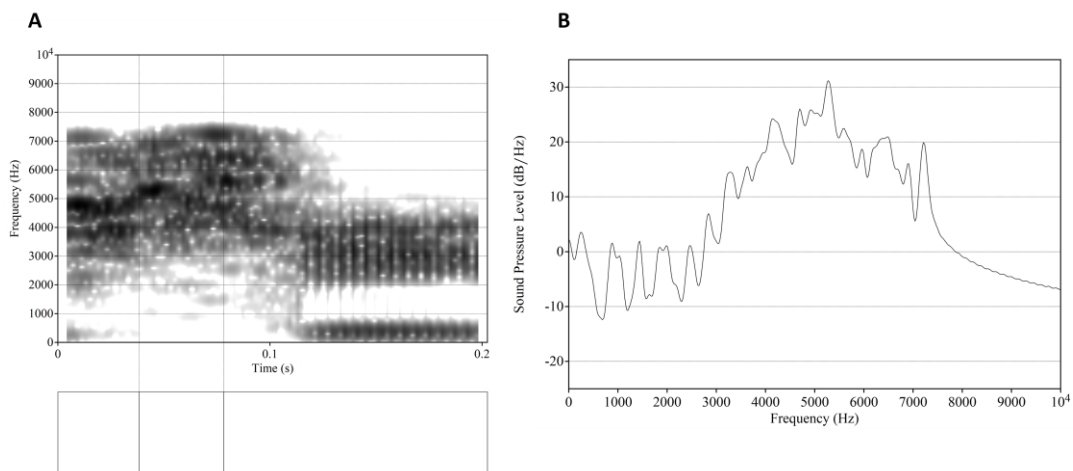


Figure 4.6 Spectrogram and spectrum from the “nonbinary” vocal tract “see” token from Hope and Lilley (2023)

Figure 4.6. above shows the spectrogram (A) of the word "see" produced by the “nonbinary” synthetic voice described in Hope and Lilley (2023) and spectrum (B) of the [s] from the word “see” on the left. The spectrum shows the middle 40 ms of the [s]. The peak frequency is around 5000 Hz and the highest sound pressure levels are between 4000 and 7000 Hz.

#### 4.1.4 Questions and Hypotheses

The broader goal of this project was to explore how cisgender and GE listeners categorize phonemes within an /s/-/sh/ continuum when the sibilants used to create the continuum and the associated vowels were sourced from gender expansive synthetic voices—in short, when the stimuli were modeled after the speech of gender expansive individuals. This project explored the perception of sibilants combined with different types of GE voices (e.g., a voice created from all gender expansive talkers, a voice

created from just nonbinary talkers, a voice created from just transfeminine talkers, and a voice created from just transmasculine talkers). Our research questions were:

(1) Do cisgender individuals differ from GE listeners in their categorization of stimuli along the /s/-/sh/ continua combined with vowels taken from GE synthetic voices? We hypothesized there would be differences and that the GE listeners would perceive [s] more of the time compared to cisgender listeners for all voices. This hypothesis derives from our previous findings (Hope & Lilley, 2023), which showed that GE listeners perceived [s] more often, especially for a “nonbinary” synthetic voice. This hypothesis is also supported by our more recent production findings that GE talkers did not systematically vary [s] production with gender (whether gender was measured along gradient scales of femininity, masculinity, or other gender or whether the GE talkers were split into sub-groups of self-reported categorical gender). The maximum average peak frequency found among all 16 talkers was 6743 Hz and the max average COG was 6502 Hz. These values were not as high as the values that were reported previously in terms of encoding of gender in [s]; with relatively low COG and peak frequency of [s], we anticipated that [s] productions with higher COGs and peak frequencies would be heard as even more [s]-like (see Chapter 2).

(2) Are there differences in sibilant categorization between the different voices (and further do these differences differ between GE and cisgender listeners and if so, how)? What are the effects of the vocal tract (e.g. vowels taken from different gender expansive synthetic voices) on sibilant categorization? We hypothesized that the cisgender group would show a vowel effect similar to effects reported by Strand and Johnson (1996), whereas we anticipated that there would be a different effect present in the GE group, especially for the gender expansive and nonbinary synthetic voices.

We made this hypothesis for the GE group, as their voices may not conform to cisheteronormative patterns as closely as the transmasculine and transfeminine voices do.

## **4.2 Methods and Materials**

### **4.2.1 Stimuli Creation**

#### **4.2.1.1 Synthetic Voice Creation**

The synthetic voices we used for this experiment are the same as in Chapter 3. For reference, the synthesizing process is summarized again:

We used Merlin DNN (Wu et al., 2016) and Theano (Al-Rfou et al. 2016) to train models that predict speech duration and acoustic features of speech. For synthesis, we used the WORLD vocoder (Morise et al., 2016) to generate natural-sounding speech based on pitch, spectral envelope, and aperiodic components.

#### **4.2.1.2 “Sack” to “Shack” Continuum Creation**

The four generated synthetic voices were made to produce the same carrier phrase with the word “sack” or “shack”: “please say the word [word] now.” The full word “sack” was extracted, and the [ʃ] from “shack” was extracted, then the [s] from “sack” was extracted, and the [s] and [ʃ] were blended using a sibilant blending Praat script by Matthew Winn to ensure that the amplitude was equalized across the continuum. This script produced nine steps. Each “ack” from each voice was combined with its own [s] to [ʃ] continuum, and then re-combined with the other voices’ [s] to [ʃ] continuum. In total, this created 144 stimuli. However, because the output of the WORLD vocoder when used in SGDs is typically sampled at 16k Hz, this caused a

problem for the script and cut the durations of the sibilants so that they may have sounded more affricate-like to some listeners. Therefore, we limited our analysis to only the stimuli which included the ALL sibilant as this was 0.60 seconds in duration and we could get a 0.40-second window to calculate the spectral means and peak frequencies; this resulted in 36 stimuli for the final analysis. Each stimulus was presented to each listener twice, in a random order and in different orders across listeners. The following table (Table 4.1) shows acoustic measures of the [ae] vowel from each of the four voices.

Table 4.1 Fundamental frequency ( $f_0$ ) averages and standard deviations for each of the sentences and passages for each synthetic voice.

<b>VOICE</b>	$f_0$	$f_1$	$f_2$
<b>ALL</b>	111	844	1719
<b>NB</b>	154	759	1770
<b>TF</b>	131	757	1571
<b>TM</b>	128	777	1732

Table 4.2 contains sibilant information (intensities and COGs) from the ALL sibilant spectrum from [s] to [ʃ].

Table 4.2 Sibilant spectrum measures of intensity in decibels (dB) and spectral mean in Hertz (Hz) for the ALL sibilant spectrum from step 1, which contained only the [s] signal to step 9 which contained none of the [s] e.g. contained only [ʃ]

<b>SIBILANT CONTINUUM STEP</b>	<b>INTENSITY (DB)</b>	<b>SPECTRAL MEAN (HZ)</b>
<b>STEP 1 (100% [S])</b>	50.2	7659
<b>STEP 2</b>	52.4	7783
<b>STEP 3</b>	53.7	7702
<b>STEP 4</b>	55.0	7456
<b>STEP 5</b>	56.2	7051
<b>STEP 6</b>	57.6	6240
<b>STEP 7</b>	59.0	5299
<b>STEP 8</b>	60.4	4450
<b>STEP 9 (0% [S])</b>	61.7	3814

#### 4.2.2 Participants

Forty participants recruited via the website Prolific ([www.prolific.com](http://www.prolific.com)) participated in an online speech perception experiment that was approved by the University of Delaware IR. There were two rounds of recruitment; the first was set to allow only 20 nonbinary individuals and the second was set to allow 20 cisgender and transgender men and women according to Prolific’s identity settings. Those who participated in the first round were excluded from participation in the second round. For both rounds, Prolific was set up to have a pre-screen that admitted only the participants who disclosed they were between the ages of 18 and 65 years and that the

place where they spent the most time before they turned 18 years old was in the United States. We used Prolific's "fluent language" screener which we set to English; the screening tool for identifying "fluent languages" enables participants to indicate the language(s) in which they consider themselves fluent. The participants' ages ranged from 18 to 61 years ( $mean = 29.7, SD = 9.6$ ) and the self-described (open-ended) racial breakdown was: 28 white, 2 Black, 2 Multicultural/mixed, 2 Latinx, 2 Hispanic, 2 Asian, 1 Native American, 1 Native American/mixed. Eighteen of the participants were gender expansive (the other 22 were cisgender). Participants were compensated \$15.49/hour for completing the experiment, as set by Prolific. The experiment took a median completion time of 30 minutes.

#### **4.2.3 Procedures**

The experiment was built and run through Gorilla. The first part of the experiment involved completing a comprehensive demographic survey that asked questions about the participants' age, race, and gender. Questions about gender were evaluated using six gender scales corresponding to three scales of gender identity (female, male, and other gender) and three scales of gender expression (feminine, masculine, and other gender).

Participants then read a screen with instructions informing them that they would be listening to sounds, and then answering whether they heard the word "sack" or "shack." Participants were instructed to be in a quiet space with minimal background noise and a pair of headphones, and then they answered a screening question to make sure that they could properly hear a word produced by one of these synthetic voices. Each stimulus was presented twice. There were four blocks with 72 trials in each block for a total of 288 trials. Blocks were organized so that they

contained only the same vocal tract (e.g. the same “ack”) within each block. Blocks were randomized and trials in each block were randomized.

After the listening task, participants then engaged in a subjective gender perception task with the four voices that formed the basis for the study presented in Chapter 3.

#### 4.2.4 Statistical analyses

Statistical analyses were conducted in R. The lme4 package was used to compute logistic regressions using the glm and glmer functions (Bates et al., 2015). The lmerTest package was used to calculate likelihood ratio tests (LRTs) using the anova function with test=“LRT” (Kuznetsova et al., 2017). Visualizations were created using ggplot2 (Wickham, 2016). To answer the question if GE listeners perceived [s] more often than cisgender listeners, we aggregated all stimuli (e.g. the four different voices) and looked at overall “sack” response for cisgender and GE listeners. A mixed-effect logistic regression was used to test whether membership in the gender expansive community (GE), vocal tract, step or the interactions of step and GE with vocal tract significantly predicted “sack” response. The fitted regression model was:

$$Response \sim GE + Vocal\ Tract + Step + Step * Vocal\ Tract + \\ Vocal\ Tract * GE + (1|Participant) \quad (4.1)$$

where *Response* is a binary variable 1 vs. 0 (where 1 represents that they did hear “sack” and 0 represents that the listener did not hear “sack”); *GE* is a binary variable GE vs. Cis (encoded 1 vs. 0); *Vocal Tract* is a categorical variable with 4 categories (ALL representing the voice created from all gender expansive talkers, NonBin representing the voice created from the nonbinary talkers, TF representing the voice

created from the transfeminine talkers, and TM representing the voice created from the transmasculine talkers); and *Step* is encoded as a whole number between 1 and 9.

*Vocal Tract* was encoded in R as dummy variables with ALL as the reference.

As in Chapter 3, we will again adopt an alpha of 0.1 to reduce the potential for Type II error (Roy et al., 2004; Schumm et al., 2013).

### 4.3 Results

GE listeners were more likely to perceive “sack” across the entire sibilant spectrum. Figure 4.7 shows the sibilant categorization curves for both groups.

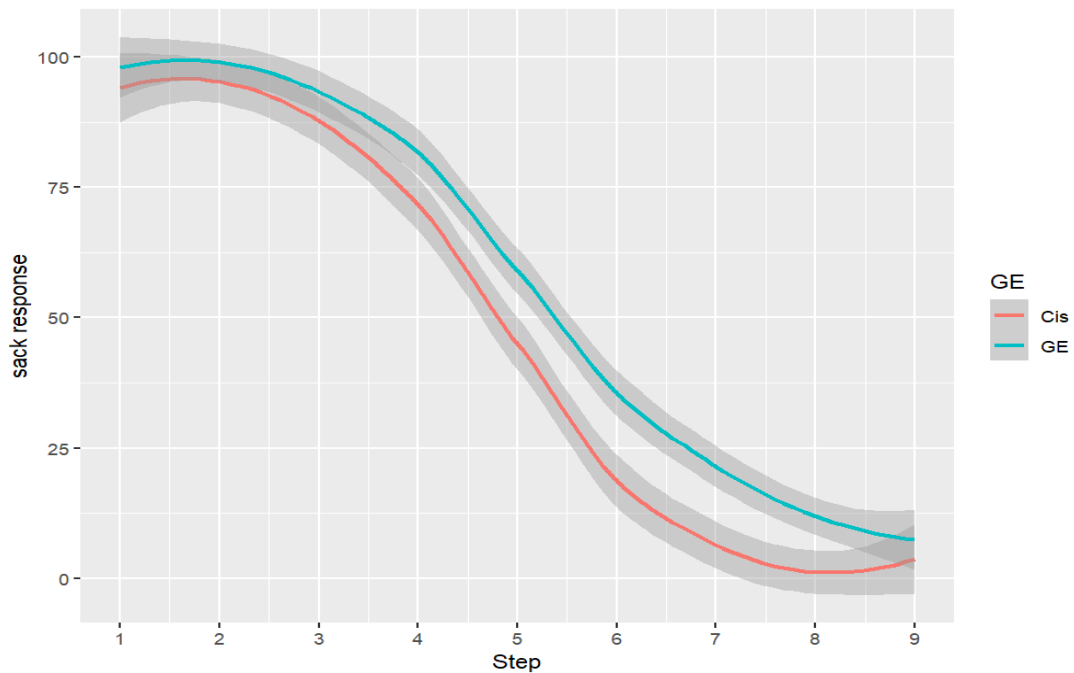


Figure 4.7 Overall sibilant categorization between gender expansive and cisgender listeners; the x-axis represents the “step” from 1 to 9 with 1 representing 100% [s] in the acoustic signal and 9 representing 0% [s] in the acoustic signal.

Figure 4.7 illustrates the higher “sack” responses for the GE group compared to the cisgender group for all stimuli combined.

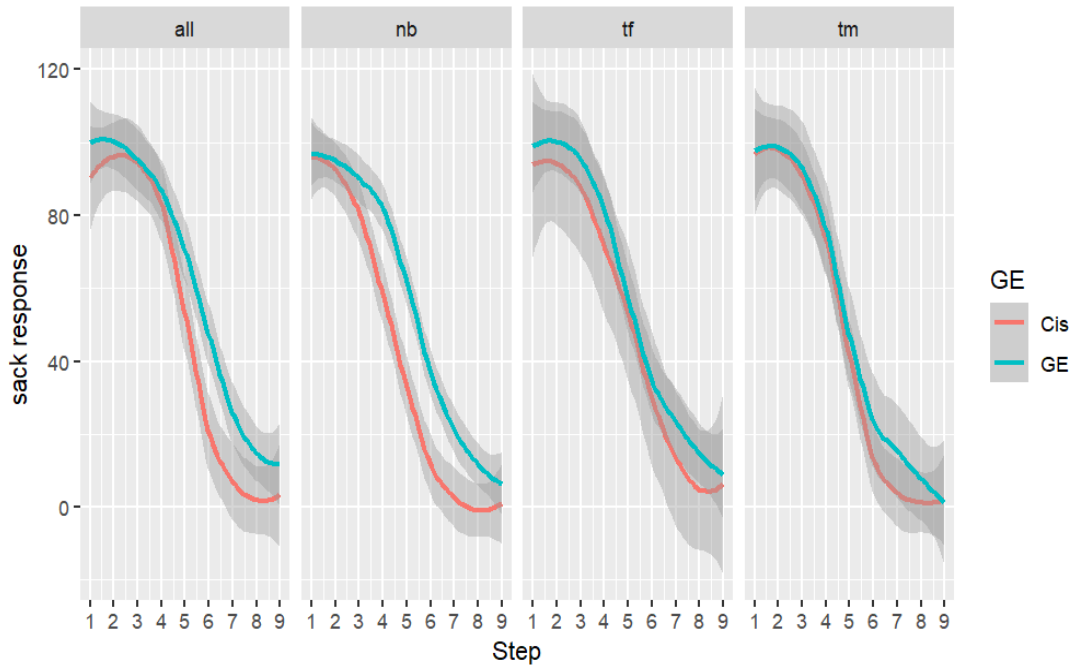


Figure 4.8 Sibilant categorization between gender expansive and cisgender listeners for the four different vocal tracts; the x-axis represents the “step” from 1 to 9 with 1 representing 100% [s] in the acoustic signal and 9 representing 0% [s] in the acoustic signal.

Figure 4.8 shows that for stimuli from each of the four vocal tracts, GE listeners perceived “sack” more often than cisgender listeners. Especially for the ALL and NB vocal tracts.

For our logistic regression, we first used the LRT to compare nested models against the full model (eq. 4.1) to assess the contribution of the two interaction terms to model fit. The first nested model we compared was the following:

$$Response \sim GE + Vocal\ Tract + Step + Vocal\ Tract * GE + (1|Participant) \quad (4.2)$$

An LRT comparing eq. 4.2 to 4.1 was not significant at  $p = 0.1045$ , implying that the *Vocal Tract \* Step* term did not contribute significantly to model fit. The second nested model we compared was:

$$\text{Response} \sim GE + \text{Vocal Tract} + \text{Step} + \text{Vocal Tract} * \text{Step} + \quad (4.3) \\ (1|\text{Participant})$$

An LRT comparing eq. 4.3 to 4.1 found a significant difference between model fits ( $p = 0.0286$ ), indicating that the *Vocal Tract \* GE* term was a significant contributor to model fit. Given that we might expect the effect of *Step* to not be affected by *Vocal Tract* and lack of significance in the LRT comparing eq. 4.2 to 4.1, we decided to use eq. 4.2 as our final model and not include the *Vocal Tract \* Step* interaction term (while still including the *Vocal Tract \* GE* interaction).

Table 4.3 shows the results of the logistic regression including coefficients, standard errors, z-values and p-values. Both the intercept and the effect of group (GE) showed significant effects on the response variable. The NB vs. ALL vocal tract comparison was highly significant as well. The variable Step had a highly significant impact on the response variable, with a strong negative coefficient, indicating its importance in the model. The interaction dummy term of the GE group and the TM vocal tract (compared to Cis group and ALL vocal tract) was also significant.

Table 4.3 Results of the logistic regression, including coefficients, standard errors (std. error), z-values and *p*-values (\* indicates that the *p* -value was < 0.1; \*\*\* indicates that the *p*-value was < 0.001).

VARIABLE	COEFFICIENT	STD. ERROR	Z-VALUE	P-VALUE
<b>(INTERCEPT)</b>	4.88064	0.22790	21.416	< 2e-16 ***
<b>GEGE</b>	1.03936	0.22812	4.556	5.21e-06 ***
<b>NB VOCAL TRACT (VT)</b>	-0.74019	0.21531	-3.438	5.86e-04 ***
<b>TF VT</b>	0.04711	0.21382	0.220	0.8256
<b>TM VT</b>	-0.27470	0.21337	-1.287	0.1979
<b>STEP</b>	-0.97739	0.03403	-28.719	< 2e-16 ***
<b>GE:NB VT</b>	0.23506	0.31930	0.736	0.4616
<b>GE:TF VT</b>	-0.43097	0.31992	-1.347	0.1779
<b>GE:TM VT</b>	-0.67213	0.31912	-2.106	0.0352 *

#### 4.4 Discussion & Conclusion

This study presents a novel approach to sibilant perception experiments; in particular, it is the first of its kind to use synthetic voices created from gender expansive individuals to assess sibilant perception. Our analysis revealed intriguing findings regarding the influence of vocal tract cues and gender identity on speech perception between gender expansive and cisgender listener groups. The intercept and the group variable exhibited significant effects on the response variable, emphasizing

the importance of gender expansive identity in speech perception. The GE group heard significantly more [s] overall than the cisgender group. This may be due to the fact that GE talkers do not tend to produce exceptionally high [s] (see Chapter 2) so when they are listening to GE voices, higher spectral means (along the [s]-[ʃ] continua) are perceived more often as [s]-like. The main effect of the NB vocal tract (compared to ALL) was significant, indicating higher “sack” responses for the NB voice. As predicted, the *Step* variable demonstrated a substantial impact on speech perception, highlighting the significance of the [ʃ] sound presence in the perception of the sibilant as [ʃ]. Our nested model comparison did not show that the interaction between vocal tract cues and *Step* significantly contributed to model fit, suggesting that the presence of [ʃ] sound influences perception consistently across different vocal tract configurations.

In conclusion, this study underscores the intricate interplay between listener gender identity, acoustic cues, and speech perception, providing valuable insights into the perceptual dynamics of nonbinary and gender expansive voices. Further research is warranted to explore the nuanced relationships between these factors and their implications for communication and identity.

## Chapter 5

### NONBINARY VOICE GENDER ENCODING IN SYNTHETIC VOICE

#### 5.1 Introduction

Humans can encode multiple facets of identity in speech and have the potential to construct various personas. Performance of gender is one of the areas in which we may construct a persona or personas through our use of speech and language. Construction of a persona related to gender in voice is not meant to imply that gender is not real, but rather that the substance of gender is less innate and more something that we do (and learn). Voice is one of the key ways that we encode gender and express our gender identity to others. From an early age, we are socialized to associate certain vocal features, such as pitch, with binary gender categories e.g. male or female (Patterson & Werker, 2002). As a result, our voices may be a significant source of dysphoria or discomfort for individuals whose gender does not align with these traditional binary categories. Voice that doesn't align with our identity can lead to misgendering and stigma, which can increase feelings of shame that has mental health repercussions (James et al., 2016; Grollman, 2017).

Synthetic voices have made significant strides in replicating human speech patterns, yet their state-of-the-art manifestations remain largely confined to cisgender, heterosexual male and female models (Ito et al., 2016; Zandie et al., 2021; Zen et al., 2019). Given this landscape, the integration of gender expansive synthetic voices represents a critical opportunity to redefine the boundaries of speech technology. By creating synthetic voices that reflect the diversity of gender identities and expressions,

we can help to ensure that individuals who do not identify as male or female have access to speech technology that accurately represents their identity. This can be particularly important for individuals who may feel marginalized or excluded by traditional binary gender norms, and who may experience external or internal harm as a result.

For example, nonbinary individuals may face significant challenges when using speech-generating devices or other forms of speech technology that are designed for binary gender categories. This can include feeling misgendered or dysphoric when hearing their own voice, as well as experiencing discrimination or misunderstanding from others who do not recognize their gender identity. By creating gender expansive synthetic voices that accurately reflect the speech patterns and characteristics of nonbinary individuals, we can help to protect against these harms and promote greater inclusivity and representation for nonbinary individuals in speech technology.

In addition to these practical benefits, gender expansive synthetic voices can also play an important role in promoting greater understanding and awareness of gender diversity. By creating synthetic voices that challenge traditional binary gender norms and highlight the diversity of gender identities and expressions, we can help to promote a more inclusive and equitable society. This inclusivity can be particularly important in contexts such as education, healthcare, and entertainment, where speech technology can play a significant role in shaping individuals' experiences and perceptions of gender. In conclusion, gender expansive synthetic voices are a crucial tool in promoting inclusivity, representation, and understanding of gender diversity. By accurately reflecting the speech patterns and characteristics of nonbinary individuals, and by challenging traditional binary gender norms, gender expansive

synthetic voices can help to protect against external and internal harm and promote greater inclusivity and representation for nonbinary individuals in speech technology and beyond.

The purpose of this last study is to explore the experiences of one nonbinary participant who used a synthetic voice created specifically for them over the course of a week. The participant recorded their experiences in a daily online journal, documenting how the synthetic voice impacted their communication and sense of gender affirmation. We chose a case study for this work due to the benefits of this particular methodology. Case studies are a research method that involves in-depth examination of a specific individual, group, or situation. These studies are often used in social sciences, business, and medical research. One major benefit of using a case study approach is the ability to collect rich, detailed data. By focusing on a single case, a wealth of information can be gathered that is not possible through other research methods (Paparini et al., 2020) and can include observations, interviews, and documentation of the case's history. Paparini et al. (2020) state that the case study method is widely acknowledged for its valuable role in helping us comprehend how context, which is dynamic and ever-changing, affects complex interventions at a systemic level. In addition, case studies allow for the exploration of complex phenomena that cannot be studied in a laboratory or experimental setting and provide an opportunity to examine how individuals or groups respond to unique situations or events. Case studies also provide a means of investigating relationships between variables in real-world settings. However, despite the advantages that case studies provide, they also have limitations such as a lack of ability to generalize to a broader population. Therefore, this case study should be viewed as an example to future

studies on these topics which should seek more participants in order to draw wider conclusions. Through this case study, we aim to shed light on the unique experiences of nonbinary individuals who use synthetic voices, and to identify areas for further research and development in this field. This will also serve as a feasibility study so that similar future studies can be conducted.

### **5.1.1 Synthetic voice and gender**

Synthetic voices, also known as text-to-speech (TTS) systems, are computer-generated speech that have been used in a variety of applications, such as virtual assistants, navigation systems, and communication aids for individuals with speech impairments. TTS systems work by converting text into speech, with a synthetic voice used to produce the sound of the spoken words.

Previous research has explored the relationship between synthetic voices and gender, with most synthetic voices created to sound either male or female. For example, a study by Mullennix et al. (2003) aimed to investigate whether people perceive computer-synthesized speech differently based on whether the voice is male or female. Participants listened to recordings of computer-synthesized speech, with half of the recordings using a male voice and the other half using a female voice. Participants then rated the voices on a number of characteristics related to social perception, including intelligence, friendliness, and attractiveness. Results showed that participants rated the female computer-synthesized voice as more friendly, attractive, and expressive than the male computer-synthesized voice. On the other hand, the male voice was rated as more intelligent, more persuasive, and dominant than the female voice. It was also found that the overall preference was for the male synthetic voices.

While previous research has explored the relationship between synthetic voices and gender, there is a need for more diverse and inclusive synthetic voices that can accurately represent the diverse gender identities of individuals who use them. In particular, there is a growing recognition of the need for synthetic voices that can accurately represent nonbinary gender identities, which are not exclusively male or female (Danielescu et al., 2023; Hope & Lilley, 2023; Netzorg et al., 2024).

In recent years, notable advancements have been witnessed in the realm of synthetic voice generation, whereby computer-based technologies have increasingly demonstrated the capacity to emulate the vocal characteristics of human beings. Such artificially constructed voices find application across a multitude of domains, with a particular emphasis on speech-generating devices. These devices play a pivotal role in facilitating communication for individuals who are unable to engage in natural speech, whether this inability is permanent or temporary in nature. Empirical investigations in this area have unveiled that synthetic voices encode in-group and out-group memberships in terms of gender (Hope & Lilley, 2020; Hope & Lilley, 2022; Hope & Lilley, 2023). Thus, scholarly inquiries underscore the future potentialities of the empowering aspect of computer-generated voices for individuals who identify as nonbinary, affording them a heightened sense of autonomy and control in their conversational exchanges. If researchers take a community-informed approach to designing nonbinary synthetic voices, better user outcomes are hypothesized.

It is evident that the utilization of synthetic voices that transcend the binary constructs of gender holds promise in enabling nonbinary individuals to express their gender identity effectively. This, in turn, may foster greater visibility and recognition of nonbinary identities within the broader societal framework.

### **5.1.2 Nonbinary identity and synthetic voice**

Gender manifests in language through lexical choices, grammar, and vocal characteristics. Langman and Shi (2020) have illustrated that pronouns and the semantic content used to encode gender can be perceived variably, reflecting social relationships to the speaker. Additionally, gender cues differ across languages and cultures. This variability extends to phonetic features, as extensive research has shown that gender is a social construct influencing both the production and perception of these features (Zimman, 2017; Leung et al., 2018; Schmid & Bradley, 2019; Hope & Lilley, 2022).

The adoption of gender-inclusive language transcends mere politeness or political correctness; it carries profound implications for nonbinary individuals. Recent studies indicate that the persistent use of gendered language reinforces implicit biases and perpetuates stereotypical views of those who deviate from binary gender norms (Zimman, 2021). Moreover, language significantly shapes how nonbinary people perceive themselves and are perceived by others (Farrow, 2019). Nonbinary individuals, whose gender identities do not conform to traditional male or female categories, exist on a spectrum of gender identities. This demographic is frequently underrepresented in scholarly literature, resulting in a substantial gap in the discourse on their language use (Rechsteiner & Sneller, 2023). Preliminary research conducted has found that nonbinary talkers utilize features of speech that are traditionally likened to the speech of men and women, alongside features that are unique to nonbinary people (e.g. Schmid & Bradley, 2019; Hope et al., 2023). For instance, nonbinary people have been found to have an average fundamental frequency in the middle of the average fundamental frequencies for men and women (Schmid & Bradley, 2019), and do not show a correlation between acoustic detail in the production of sibilant

fricatives and gradient, multidimensional gender identity, unlike cisheterosexual men and women (Hope et al., 2023). However, these studies have been preliminary in nature, and larger, more comprehensive studies will be needed to have a fuller picture of nonbinary cues in speech production.

In general, scholars have observed a prevailing dearth of support within their respective communities for nonbinary individuals, who frequently grapple with the neglect of their identities. Research that includes transgender and nonbinary perspectives challenges the idea that sex and gender will always align in conventional ways, that one's gender identity can be determined solely by appearance, that sex and gender are strictly binary concepts, and that individuals within the same gender category will conform to norms for that category to the same degree (Zimman, 2021). Omitting nonbinary talkers from sociolinguistic investigation further marginalizes this community and the extant body of research provides compelling evidence that nonbinary individuals contend with societal pressures, compelling them to adhere to the confines of the traditional binary gender framework because their identities are frequently invalidated and neglected. Furthermore, empirical inquiry underscores the various forms of discrimination endured by nonbinary individuals on account of their gender expansive identities, experiences which often yield feelings of exclusion, marginalization, and invisibility within the societal fabric (James et al., 2016, Grollman, 2017).

It is noteworthy that nonbinary gender identities have made notable strides in achieving increased visibility and recognition, particularly within LGBTQ+ communities. Nonbinary individuals embrace a multifaceted spectrum of identities, encompassing genderqueer, agender, genderfluid, and a plethora of nonbinary

expressions. For many within this group, the presence of gender dysphoria, desire for gender euphoria, or the disquietude with their assigned gender underscores the quest for authentic gender expression. Conversely, the design of synthetic voices has, hitherto, predominantly adhered to a binary gender paradigm, with a pronounced orientation toward voices that align with either a male or female presentation, and specifically a certain type of male or female presentation e.g. often white, cisheterosexual. Current attempts at creating "gender-ambiguous" or "gender neutral" synthetic voices have relied on largely cisheterosexual, male and female speech corpora (Hope & Lilley, 2022; Székely et al., 2023). An emerging consensus underscores the need for the development of synthetic voices that are more diverse and inclusive, proficiently mirroring the multifarious gender identities and expressions characteristic of the individuals who rely on them for communication.

### **5.1.3 Questions and hypotheses**

Our question was broad, exploratory and descriptive: How do synthetic voices created from gender expansive talkers, which are identified by a nonbinary SGD user as gender-affirming, affirm that nonbinary SGD-user's gender? From this core question we have several other questions such as: how does the experience of this voice compare to their experience with previous, pre-set voices on their SGD? Were there differences in how affirmed they felt in different scenarios or settings?

We hypothesized that this voice would improve the participant's gender affirmation compared to previous preset voices on SGDs and that they would find their gender more affirmed in situations where they were speaking with in-group members, e.g., members of the gender expansive community, especially for those who

were also nonbinary, compared to situations where they were speaking to out-group members, e.g. individuals who are not part of the gender expansive community.

## **5.2 Methods and Materials**

### **5.2.1 Recruitment**

An initial survey was conducted in an online social media group that encouraged people to ask questions to users of Augmentative and Alternative Communication (AAC). This poll allowed the creation of a specialized list of nonbinary SGD users. From here, an IRB approved study was conducted, recruiting, via email, one participant from the aforementioned pool of nonbinary SGD users. The participant had to be a full time AAC user and have an SGD that was compatible with the ModelTalker (Bunnell et al., 2017) synthetic voices. This participant signed an electronic consent to participate. They were compensated \$350 in an electronic gift card for their participation. The participant will be referred to as TT.

### **5.2.2 Synthetic voice construction**

Because we wanted these synthetic voices to be affirming to the user in question, we sent an initial poll to TT asking some initial questions about what they wanted to get out of the voice, what their gender was on six gradient scales (0-100) for *Male, Female* and *Other gender identity* and *Masculine, Feminine*, and *Other gender expression*, and their categorical and open-ended gender descriptions.

Because TT is nonbinary and wanted a mix of feminine and masculine coded articulation traits, we decided to audition the synthetic voices that were created from all 16 gender expansive participants previously described (All) and the one created from the subset of 8 nonbinary participants (Nonbinary). Then, we created six new

voices by mixing and matching the acoustic and prosody models created from the Transmasculine and Transfeminine voices with the models from the nonbinary voice. We chose to keep the Nonbinary voice as a part of these next six voices due to the participant’s categorical gender identity as nonbinary (Table 5.1).

### 5.2.3 Voice audition

The participant heard each voice a total of three times, which were equal to three different Harvard Sentences. All stimuli were all randomized. They were asked to rate the voice on a scale of 0 to 100 for the question, “How gender-affirming does this voice sound to you?” Ratings were averaged to give the average affirmation rating per voice. We then took the top three voices and created installers for the participant to use to download onto their SGD. Information about the voices and their ratings is found in Table 5.1, with voices listed in order of highest to lowest ranking.

Table 5.1 Results of the demographic questionnaire.

<b>QUESTION</b>	<b>ANSWER</b>
<b>CATEGORICAL GENDER</b>	Nonbinary
<b>OPEN-ENDED GENDER</b>	The act of queering my neurodivergences (i.e. my gender itself is best described as a verb)
<b>GRADIENT MALE GENDER IDENTITY</b>	0
<b>GRADIENT FEMALE GENDER IDENTITY</b>	0

<b>GRADIENT OTHER GENDER IDENTITY</b>	100
<b>GRADIENT MASCULINE GENDER EXPRESSION</b>	60
<b>GRADIENT FEMININE GENDER EXPRESSION</b>	50
<b>GRADIENT OTHER GENDER EXPRESSION</b>	100
<b>WHAT TYPES OF VOICE QUALITIES DO YOU WANT IN A NONBINARY SYNTHETIC VOICE? DESCRIBE YOUR IDEAL SYNTHETIC VOICE.</b>	Pitch around "androgynous" range; overall mixed features for what would code feminine or masculine - in particular low breathiness, and mixed articulation.

This Table includes information provided by the nonbinary SGD-user about their gender and voice preferences. The user identifies categorically as nonbinary and has high other gender identity and expression while their male and female identities are 0 and their masculine and gender expression are roughly in the middle of the 0-100 scale. They preferred a voice that has feminine and masculine features and emphasized an “androgynous” pitch.

The table below shows the SGD-user’s average affirmation score for each voice and its corresponding prosody and acoustic models.

Table 5.2 Results from the auditioning process showing the voice, its prosody model, its acoustic model, and the average affirmation score.

<b>VOICE</b>	<b>PROSODY MODEL</b>	<b>ACOUSTIC MODEL</b>	<b>AVERAGE SCORE</b>
<b>VOICE1</b>	Nonbinary	Transmasculine	80
<b>VOICE2</b>	All	Nonbinary	62
<b>VOICE3</b>	Nonbinary	All	61
<b>VOICE4</b>	Transfeminine	Nonbinary	55
<b>VOICE5</b>	All	All	52
<b>VOICE6</b>	Nonbinary	Nonbinary	50
<b>VOICE7</b>	Transmasculine	Nonbinary	50
<b>VOICE8</b>	Nonbinary	Transfeminine	45

The top three voices were Voice1, which had the nonbinary prosody model and the transmasculine acoustic model; Voice2, which had the “all” prosody model and the nonbinary acoustic model; and Voice3, which had the nonbinary prosody model and the “all” acoustic model. Notably for this participant, the voices that had the transfeminine prosody or acoustic models were not among the top 3 voices in terms of average affirmation score.

#### **5.2.4 Acoustic profiles of the top three voices**

From the top three voices chosen, we analyzed the Grandfather Passage (Darley et al., 1975) produced by each of them for their vowel space dispersion (VSD), aVTL, and  $f_0$  mean, min, max, and range. VSD was calculated as the average Euclidean distance of the vowels to the center of the vowel space. Vowels used to

calculate VSD and aVTL were [i], [u], and [a]. There was a total of six [a], seven [u], and 21 [i] used. Text to acoustic alignment was done manually in PRAAT (Boersma & Weenink, 2021) by two independent researchers.

Table 5.3 Acoustic characteristics of Voice1, Voice2 and Voice3.

<i>Voice</i>	<i>VSD</i>	<i>aVTL</i>	<i>f<sub>0</sub>-Mean</i>	<i>f<sub>0</sub>-Min</i>	<i>f<sub>0</sub>-Max</i>	<i>f<sub>0</sub>-Range</i>	<i>Syllables/s</i>
1	453	15.1	158.8	137.2	179.4	42.2	3.20
2	496	14.5	147	129.6	166.8	37.1	3.21
3	433	14.7	158.8	138	179.1	41.1	3.19

### 5.2.5 Daily journal entries

Google Forms was used for the daily journal entries. The form included a date entry as the first question. The form then asked which voices (Voice1, Voice2, and/or Voice3) were used that day, in what settings those voices were used (at work, school, home, social, errands/chores, other) and with whom they spoke with those voices (friends, family, clients/customers, colleagues/coworkers, acquaintances/strangers, other). The form then asked an open-ended question with follow-up questions:

- *Describe one situation you used the voice in and what your communication partners were like. What voice(s) did you use in this setting?*

This question was followed by these questions:

- *In the situation described above, were/are you "out" to the communication partner(s) as nonbinary? (Yes/No/Other),*
- *During the situation described above, how well do you feel the voice affirmed your gender? (1-10, 1 = not at all, 10 = extremely well),*

- *During the situation described above, how strongly do you feel your communication partner(s) perceived your gender in an affirming way? (1-10, 1 = not at all, 10 = completely)*
- *Compared to a preset voice you've used on your SGD before, how much did this voice improve your authenticity in voice gender expression? (1-10, 1 = not at all, 10 = completely)*

Finally, the entry ended with two open-ended questions. The first was a response area for anything the participant wanted to talk about:

- *Anything you want to talk about with the voice(s)? This is an open-ended area for you to write about your experiences.*

The final question on the survey was used to compare external treatment based on the voices:

- *How do you feel those you are interacting with perceive you based on how they are treating you using this voice versus previous voices you have used in the past?*

### 5.2.6 Post-trial Questionnaire

TT completed the week-long journaling plus an additional day, cumulating in a total of eight entries. We then sent a follow-up survey the day after the trial to assess how they perceived the voices in terms of *Masculine*, *Feminine* and *Other gender perception*. We decided to do this post-trial because this gave them more time with the voices to feel out the gender of the voices.

The following table shows TT's *Feminine*, *Masculine*, and *Other gender perception* of the three voices.

Table 5.4 TT's *Feminine*, *Masculine*, and *Other gender perception* of Voice1, Voice2 and Voice3.

<b>VOICE</b>	<b>MASCULINE</b>	<b>FEMININE</b>	<b>OTHER</b>
--------------	------------------	-----------------	--------------

<b>VOICE1</b>	15	15	85
<b>VOICE2</b>	5	30	50
<b>VOICE3</b>	5	10	80

Voice1 and Voice3 have a higher degree of *Other gender perception* for the participant compared to Voice2 while Voice2 has a higher degree of *Feminine gender perception* compared to the other two voices.

### **5.2.7 Analyses**

To analyze the open-ended responses from the journal entries, we utilized the software Atlas.ti (ATLAS.ti, 2019) to generate initial codes, using an inductive thematic approach (Saldana, 2009; Jason & Glenwick, 2016). Then, from the 50 initial auto-generated codes, the author and an undergraduate research assistant reviewed the codes and condensed them into 15 total codes, or subthemes, within four major themes that reflected the aims of this study. The author and the undergraduate researcher then labeled the journal entry content using these new codes independently and afterwards, cross-checked for any discrepancies in labeling. Initial agreement was 86% in coding the entries. To analyze the quantitative data, we used Excel to calculate means.

## **5.3 Results**

### **5.3.1 Overview of journal entries**

Provided here is a chronological look at the open-ended journal entries. Summarized are the voice(s) the participant primarily used over the course of that day, who their conversations partners were, and what the main content of the conversation was about.

Day 1 - 02/18/2023

**Voice(s):** Voice1

**Situation of conversation:** recorded videos/ and share in AAC group

**Who they are talking to:** other AAC users

**Content of the conversation:** discussing wanting these options available + other users discussing wanted them for themselves, multiple of them said “MUCH GENDER” and similar in a good way

Day 2 - 02/19/2023

**Voice(s):** Voice2 / Voice3

**Situation of Conversation:** casual conversation at home

**Who were they talking to:** Partner

**Content of Conversation:** TT felt that the voices made them feel less human and difficult to understand. TT didn't mention how their partner felt about the conversation, since it was a casual conversation and not a big social interaction. Overall, TT said that they just need more time with the voices in order to get used to them. TT started out with Voice2 for experimentation but then switched to Voice3 for this conversation

Day 3 - 02/20/2023

**Voice(s):** Voice3

**Situation of Conversation:** Zoom Call with a support group

**Who were they talking to:** Acquaintances and strangers in the support group (some of them are transgender), their partner.

**Content of Conversation:** TT had interacted with this group a couple of times. They are the only AAC user and had a hard time feeling included in the conversations. Their gender did feel affirmed though

and they stated how they prefer Voice3 over the others. They did find the conversation difficult however, since the voice wasn't identical to a natural human voice and wished that the voice had more character to make conversation easier.

Day 4 - 02/22/2023

**Voice(s):** Voice1 / Voice2

**Situation of Conversation:** Weekly voice call with online friends

**Who were they talking to:** Friends, family, and partner

**Content of Conversation:** TT felt affirmed and included in the conversation, using Voice1. They successfully felt neither male nor female and that they will use the voices (mainly Voice1) when they want to make a point about their gender, because that voice had the most positive responses. TT experimented with Voice2 but decided to use Voice1 for this conversation as it affirmed their gender more than Voice2.

Day 5 02/23/2023

**Voice(s):** Voice1

**Situation of Conversation:** Casual conversation at home

**Who were they talking to:** Partner

**Content of Conversation:** The conversation was difficult for TT because their partner was having a hard time understanding the voice and was getting frustrated with it. It was only when they were both sitting down that it got easier. TT stated how they felt limited using the voice because they felt like they couldn't move around with it. TT did feel affirmed however, stating that the voice is successfully neither male nor female, but being understood in public is a concern.

Day 6 02/24/2023

**Voice(s):** Voice1

**Situation of Conversation:** Checking in to go pick up a CSA

**Who were they talking to:** Acquaintances, strangers

**Content of Conversation:** Some of the people TT talked to knew they were nonbinary. The people were confused, which TT said was good because it affirms that their voice is neither male nor female. TT did feel that it was dehumanizing when people they talk to don't know how to interact with them because of their presentation.

Day 7 02/26/23

**Voice(s):** Voice3

**Situation of Conversation:** Testing AAC methods and access for fun

**Who were they talking to:** Partner

**Content of Conversation:** TT felt like they were “fighting” with the voice and that it didn't affirm their gender (nonbinary). The voice gave them dysphoria because of the limitations they felt as a consequence of the limitations of the technology.

Day 8 02/27/23

**Voice(s):** Voice3

**Situation of Conversation:** Zoom call with only AAC users

**Who were they talking to:** Friends, acquaintances

**Content of Conversation:** TT felt a little more pressure than everyone else in the conversations because they were much slower using the voice. As a result, they felt some exclusion. The voice said their name correctly, which they said is very important when having conversations, because most generative voices have trouble with names.

### 5.3.2 Survey results

For the question of how much the voices affirmed the TT’s gender (on a scale of 1 to 10), the average across all situations was a score of 6.13. The situation with the lowest score for how affirming the voice was occurred on Day 7 when TT was talking with their queerplatonic partner who was also an AAC user and was using Voice3. The situation which evoked the highest score was on Day 4, when TT was on a voice call with friends and was using Voice1. This situation also scored the highest for how strongly they were perceived in an affirming way by their communication partners (on a scale of 0-10, this scored a 10). The average for that question was 8.38. For the question of how much did the voice improve their authenticity in voice gender expression, the average score (from 0-10) was 5.5. The highest scoring days were Day 3, 4, 5 and 6 with a score of 7 and the lowest score was from Day 7. Daily responses are shown in Figure 5.1.

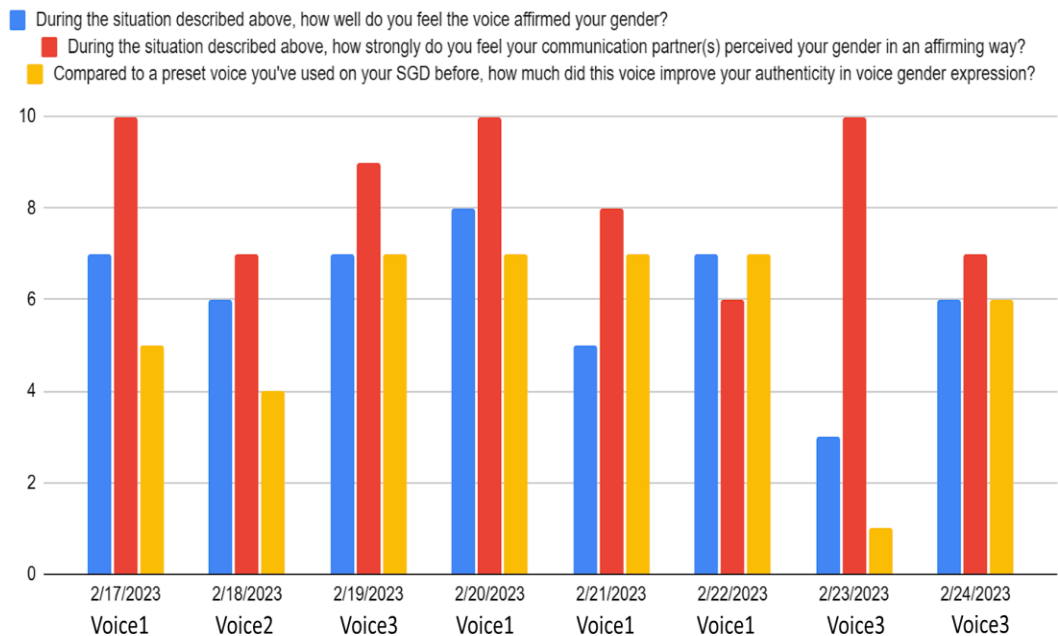


Figure 5.1 Responses from each of the days the voices were used for the 3 quantitative questions. The bottom voice label represents the voice that was primarily used that day.

Voice2 was only documented for one situation; although TT reported having tried it out on their own, they were not satisfied with this voice so we will not report on further quantitative aspects of it due to a lack of quantitative data.

Averaging by voice, Voice1 had an average score of 6.75 across four days for the question of how much it affirmed TT's gender, an average score of 8.5 for how strongly TT felt their communications partners perceived their voice in an affirming way, and an average score of 6.5 for how much this voice improved TT's authenticity in voice gender expression compared to previous preset voices on SGDs. Voice3 had an average score of 5.3 across three days for the question of how much it affirmed TT's gender, an average score of 8.67 for how strongly TT felt their communications partners perceived their voice in an affirming way, and an average score of 4.6 for how much this voice improved TT's authenticity in voice gender expression compared to previous preset voices on SGDs.

### **5.3.3 Qualitative Data Results**

The total number of instances across subthemes was 79. Numbers for the various themes discovered were: feelings and sentiments (n=19), acoustic cues (n=19), actions and social interactions (n=23), and technology (n=18). Each theme, subtheme, and supporting segments from participants' interviews are discussed below. Table 5.5 displays the themes and subthemes as constructed via ATLAS.ti after they were independently processed by two different researchers, the author and their research assistant, and then reconciled together.

Table 5.5 Overall themes and subthemes of the codes used and the total instances of each subtheme (n) used in all journal entries.

<b>THEME</b>	<b>SUBTHEME</b>	<b>INSTANCES (N)</b>
<b>FEELINGS/SENTIMENTS</b>	Gender Affirmation (+/-)	12
	Frustration	1
	Excitement	2
	Uncertainty/Insecurity	4
	Preference	7
<b>ACOUSTIC CUES</b>	Speed	5
	Pitch	2
	Prosody	10
	Voice Quality	2
<b>SOCIAL INTERACTIONS</b>	Gender Identity (Perceived)	10
	Community	5
	External Emotions (Perceived)	1
<b>TECHNOLOGY</b>	Artificiality	7
	Language/Communication	4
	Barrier	
	Inclusion	7

Table 5.5 shows that the themes of Gender Affirmation, Prosody, and Gender Identity (Perceived) have the highest number of instances of mention in the analyzed journal entries whereas the journal entries touched less on TT's other emotions (such as Frustration or Excitement), pitch and voice quality, and how other people felt during their interactions (External Emotions (Perceived))

Table 5.6 shows the total instances (n) and the percentage of each voice that accounted for the codes.

Table 5.6 Percentage of instances of each code attributed to Voice1, Voice2, and Voice3.

THEME	SUBCATEGORY	N	%	%		
				VOICE1	VOICE2	VOICE3
<b>FEELINGS / SENTIMENTS</b>	Gender Affirmation (+/-)	12	75	16.67	8.33	
<b>FEELINGS / SENTIMENTS</b>	Frustration	1	100	0	0	
<b>FEELINGS / SENTIMENTS</b>	Excitement	2	100	0	0	
<b>FEELINGS / SENTIMENTS</b>	Uncertainty / Insecurity	4	75	0	25	
<b>FEELINGS / SENTIMENTS</b>	Preference	7	57.14	14.29	28.57	
<b>ACOUSTIC CUES</b>	Speed	5	80	0	20	
<b>ACOUSTIC CUES</b>	Pitch	2	50	0	50	
<b>ACOUSTIC CUES</b>	Prosody	10	60	10	30	
<b>ACOUSTIC CUES</b>	Voice Quality	2	50	0	50	
<b>SOCIAL INTERACTIONS</b>	Gender Identity (Perceived)	10	70	10	20	

<b>SOCIAL INTERACTIONS</b>	Community	5	40	0	60
<b>SOCIAL INTERACTIONS</b>	External Emotions (Perceived)	1	100	0	0
<b>TECHNOLOGY</b>	Artificiality	5	40	40	20
<b>TECHNOLOGY</b>	Language / Communication	4	50	25	25
<b>TECHNOLOGY</b>	Inclusion	7	42.86	0	57.14

As demonstrated in the table above, Voice1 shows the majority of codes because it was the voice which TT used the most and recorded their experiences about. Voice2 had the least percentage of codes because it was used the least by TT.

Finally, there was a subanalysis into the Gender Affirmation codes because these were associated with both positive and negative sentiments (which are shown in Table 5.7).

Table 5.7 The subtheme “gender affirmation” broken down into its positive and negative instances with percent of those instances attributed to each voice.

<b>SUBCATEGORY</b>	<b>INSTANCES</b>	<b>% VOICE1</b>	<b>% VOICE2</b>	<b>% VOICE3</b>
<b>GENDER AFFIRMATION (+)</b>	8	100	0	0

<b>GENDER</b>	4	25	25	50
<b>AFFIRMATION</b>				
<b>(-)</b>				

Based on the information in Table 5.7, Voice1 accounted for all journal entries coded as positive gender affirmation while the negative instances of gender affirmation were overall much less and divided more evenly amongst the three voices.

In the following subsections, we will illuminate a few key quotes from each of the major themes.

### 5.3.3.1 Feelings and sentiments

An overwhelming majority of the feelings/sentiments codes pertained to Voice1. TT recognizes a sense of positive gender affirmation with this voice as shown in the following quote:

*“I would happily use a voice similar to [Voice1] gender-wise. It feels like people aren't wrong it isn't making me male it isn't making me female it is just, neither; not correct in terms of what my gender \*is\* but not \*wrong\* like voices are generally.”*

This was coded under the subtheme “gender affirmation” and was one of the eight positive instances of gender affirmation.

However, Voice1 also caused frustration; this was due to the difficulty others had in understanding the voice.

*“I've found there's a lot more of the frustration like mentioned above but once you get past that it's really helpful to have the auditory reminder that no I'm not male or female.”*

This was coded as the subtheme “frustration”.

### 5.3.3.2 Acoustic cues

TT was very attuned to different acoustic cues in the voice. They found that the average pitch of these voices was a good fit. This is highlighted in the quote below:

*“Doing a pitch analysis on some phrases I keep saved, [Voice1] came out as the average of 166Hz (minimum average 156Hz maximum average 174Hz). This pitch works really well for me.”*

This was coded under the subtheme “pitch”.

One of the primary reasons TT did not continue to use Voice2 was highlighted by the monotony of the voice and the overall voice quality as highlighted in the following quote:

*“My immediate instinct to voice 2 was it reminded me of the charlie brown teacher of it just felt like it was droning on and like I got lost in it.”*

This was coded as the subtheme “prosody”.

### 5.3.3.3 Social Interactions

TT recognized how others were viewing them in terms of gender. In the following quote, they highlight that Voice1 made them perceived as not male or female, but that this was also informed by the fact that their communication partners in this instance were nonbinary themselves, suggesting in-group sociophonetic cues using this voice:

*“I feel like they saw me similarly in terms of not being male or female, which is what the things these voices are doing most, but that was because of them being both nonbinary themselves.”*

This was coded as “Gender identity (perceived)” as well as “Community” because it related to the nonbinary community in this case.

TT felt that Voice3 was better than Voice1 in some ways, especially in the beginning of the week. However, they noticed some difficulty with others adapting to this voice that made it harder for others to affirm their gender.

*“Voice 3 felt like it might feel even better than voice 1 but like without any adjustment it was harder without people actively being affirming (because of only interacting with someone who has a hard time adjusting to voices.)”*

This was coded under “gender identity (perceived)”.

#### **5.3.3.4 Technology**

The primary challenges with the voices stemmed from the technological aspects of the voices and how the voices sounded “artificial.” These technology related problems often created language and communication barriers. These challenges are especially highlighted in the following quotes:

*“It really confirmed to me that the things I don't like [about Voice1] are in the ease of understanding and how people treat your area as he was getting repeatedly frustrated about not being able to understand what I was saying with even the slightest bit of background noise. That's always a challenge with all AAC voices but it's been harder with these.”*

This was coded as both the technology subtheme “language barriers” and as “emotions (perceived)” in the social interactions theme.

*“Overall I've been liking voice 3 best as I expected from my quick trying of each but that is making it more disappointing how it's feeling like it's not a person's voice. Not meaning not identical to any particular voice just it feels like something is missing to give it character.”*

This was coded as the subtheme “artificiality”.

## 5.4 Discussion

This study represented a feasibility study for future researchers to investigate nonbinary synthetic voice creation and use. In particular, the study developed a community-informed and directed approach to: (1) developing nonbinary synthetic voices for a nonbinary SGD-user and (2) assessing gender-affirmation in voice through synthetic voice use. Our participant, TT, found that our voices affirmed their gender more than previous, preset voices on SGDs. In the journal entries, the individual explored their experiences using different voices on their SGD device to express their nonbinary gender identity. They engaged in various conversations with different people and in different settings. Voice1 seemed to be the preferred choice, as it helped TT feel neither male nor female and garnered positive responses. However, Voice3 was also preferred for gender affirmation but had limitations in terms of sounding natural. TT encountered challenges in feeling included in conversations, especially in groups where they were the only AAC user. They also expressed concerns about being understood in public. Overall, there was a trend of seeking a voice that affirmed their nonbinary gender identity while facing challenges related to voice limitations and social interactions.

Voice1, which was assessed in the initial survey as having the highest gender affirmation score by TT, was the one that had the most affirming experience based on analysis of qualitative data. TT also felt more included by their conversation partners using this voice. TT did not use Voice2, which had the lowest “Other” gender perception rating by TT (see Table 5.4), nearly as much and largely cited the gender mismatch and monotony of this voice. However, they did use Voice3 much more consistently and found that it outperformed Voice2. This makes sense in light of Voice3’s higher “Other” gender rating compared to Voice2 (80 compared to 50). This

is interesting because Voice3 was rated as slightly lower on the initial survey as being gender affirming; despite that, the values for Voice2 and Voice3 were very similar on the initial survey. The findings from Table 5.6 illustrate a multifaceted perception of three distinct voices (Voice1, Voice2, and Voice3) across different themes. Voice1 stood out as strongly associated with gender affirmation, suggesting that it played a crucial role in affirming the participant's gender. Out of all mentions of gender affirmation, Voice1 was associated with all of the positive codes, and only once was associated with a negative feeling of gender affirmation. That voice was also linked to feelings of uncertainty and insecurity, indicating a complex emotional response. Furthermore, these voices were linked with a degree of monotony and a lack of adequate speed that impaired authentic prosody expression. Voice3 was initially linked to a sense of community, preference, and inclusion, indicating its potential to foster social connections before Voice1 became the preferred voice. However, it is important to note that Voice3 was less consistently associated with other themes, highlighting the diversity of perceptions and sentiments associated with these voices. These findings underscore the intricate relationship between voice, identity, and communication and suggest that individuals perceive and evaluate voices in a nuanced manner based on various contextual factors.

## **5.5 Conclusions**

Prior to this work, there has been little to no research, as far as we are aware, on the construction of synthetic voices specifically designed for nonbinary users using gender expansive talkers as the basis for synthetic voice creation. Synthetic voices created from gender expansive individuals are better than preset SGD voices coded explicitly or implicitly as male or female. In this study, we showed both quantitative

and qualitative data in the form of a feasibility study that paves the road for future studies to be conducted on how gender is encoded in nonbinary speech-generating device users. Future studies may want to consider using updated vocoders and making a concentrated effort towards preserving emotional cadence in voice.

## Chapter 6

### CONCLUSION

This dissertation has explored the production of gender expansive speech, the development of and perception of gender expansive synthetic voices, and the use of gender expansive synthetic voices by a nonbinary, speech-generating device user. It is the first series of studies, to our knowledge, aimed at using community-informed and -directed approaches to synthetic voice creation in the gender expansive community.

In the first study (Chapter 2), 16 gender expansive talkers participated in recording 400 English utterances, with the subsequent analysis revealing significant correlations between specific acoustic features and multidimensional gender variables. Notably, the study uncovered a strong link between the  $f_2$  formant of the vowel [o] and gradient male identity.  $f_2$  is strongly representative of front-to-backness in the oral cavity, suggesting the use of a back tongue positioning to signal masculinity. These findings have broad implications for sociophonetics and speech language pathology, potentially guiding future research on vocal techniques and therapies for gender expansive individuals.

The second study (Chapter 3) addressed the perception of gender in synthetic voices, highlighting marked distinctions between cisgender and gender expansive listeners, especially in the evaluation of gender expansive and nonbinary voices. Surprisingly, the transfeminine voice, despite being perceived as highly masculine, showed no significant differences in gender perception between the groups, challenging conventional expectations about gender encoding in speech. Meanwhile,

the transmasculine voice appeared to be perceived as more ambiguous, with balanced ratings across gender scales, and no significant differences were observed in feminine or masculine perception for the that voice. Additionally, there was a significant correlation between the listener's gradient other gender and their rating of other gender along a gradient in the synthetic voices. However, there was a difference in the usage of the other gender scales between cisgender and gender expansive listener groups, indicating the need to investigate separate models for cis and gender expansive listener groups in the future.

In the third study (Chapter 4), the analysis explored the intricate interplay between vocal tract cues, gender identity, and speech perception between gender expansive and cisgender listeners using a sibilant perception task. While group membership (GE or Cisgender) significantly influenced speech perception of [s], vocal tract cues did not independently drive perceptual differences, except for the Nonbinary vocal tract condition which had significantly higher [s] responses. One difficulty in interpreting these results was that the sibilant stimuli were unusually short for typical sibilants, making them sound almost affricate-like. Thus, the results should be taken cautiously, and future researchers should repeat this experiment with more representative stimuli. This research opens the door for further investigation into the nuances of sibilant perception and their implications for communication and identity.

The fourth study (Chapter 5) focused on the development of nonbinary synthetic voices, with a participant (TT) reporting that these voices affirmed their gender more effectively than preset voices on speech-generating devices (SGDs). Voice1, which had the highest rating in the voice audition, emerged as the preferred choice, affirming TT's nonbinary identity. While Voice3 was initially rated lower in

gender affirmation, it was used more consistently and outperformed Voice2 in terms of gender affirmation. The study uncovered the multifaceted perception of these voices, with Voice1 strongly associated with gender affirmation, both personal and external, albeit sometimes perceived as slow and monotonous. Similarly, Voice3 was also linked to acoustic cues of monotony. This study revealed concerns about voice monotony and speed, highlighting the need for improved technology in encoding emotion and more naturalistic prosody in synthetic voices.

Overall, the findings from these four studies emphasize the intricate relationship between voice, identity, and communication, demonstrating that individuals produce and evaluate voices based on contextual factors, group membership, and individual identity. The implications from these findings will strongly augment the fields of sociophonetics, speech synthesis, and communication sciences and disorders by highlighting the need for community-informed and -developed synthetic voices, inclusive and affirming of all.

## REFERENCES

- Al-Rfou, R., Alain, G., Almahairi, A., Angermueller C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., et al. (2016). Theano: a Python framework for fast computation of mathematical expressions, arXiv e-prints arXiv-1605. <https://doi.org/10.48550/arXiv.1605.02688>
- Albuquerque, L., Oliveira, C., Teixeira, A., Sa-Couto, P., Freitas, J., & Dias, M. S. (2014). Impact of age in the production of European Portuguese vowels. *Interspeech 2014*. <https://doi.org/10.21437/interspeech.2014-244>
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52(1), 388-407. <https://doi.org/10.3758/s13428-019-01237-x>
- Avery, J. D., & Liss, J. M. (1996). Acoustic characteristics of less-masculine-sounding male speech. *The Journal of the Acoustical Society of America*, 99(6), 3738–3748. <https://doi.org/10.1121/1.414970>
- Bates D., Mächler M., Bolker B., & Walker S. (2015). “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.
- Bauer, G. R., Scheim, A. I., Pyne, J., Travers, R., & Hammond, R. (2015). Intervenable factors associated with suicide risk in transgender persons: A respondent driven sampling study in Ontario, Canada. *BMC Public Health*, 15(1), 1-13.
- Berry, D. A. (1990). Subgroup Analyses. *Biometrics*, 46(4), 1227–1230. <https://www.jstor.org/stable/2532464>
- Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.48, retrieved 18 February 2021 from <http://www.praat.org/>
- Brown, L., & Pillot-Loiseau, C. (2022). Bright Voice Quality and Fundamental Frequency Variation in Nonbinary Talkers. *Journal of Voice*. <https://doi.org/10.1016/j.jvoice.2022.08.001>

- Bunnell, H. T., Lilley, J., & McGrath, K. (2017). The ModelTalker Project: A Web-Based Voice Banking Pipeline for ALS/MND Patients. *Interspeech*, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden. [https://www.isca-speech.org/archive\\_v0/Interspeech\\_2017/pdfs/2054.PDF](https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/2054.PDF)
- Carew, L., Dacakis, G., & Oates, J. (2007). The Effectiveness of Oral Resonance Therapy on the Perception of Femininity of Voice in Male-to-Female Transsexuals. *Journal of Voice*, 21(5), 591–603. <https://doi.org/10.1016/j.jvoice.2006.05.005>
- Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult talkers. *PLoS ONE*, 7(2). <https://doi.org/10.1371/journal.pone.0031353>
- Cartei, V. & Reby, D. (2013). Effect of formant frequency spacing on perceived gender in pre-pubertal children’s voices. *PLoS ONE*, 8(12). <https://doi.org/10.1371/journal.pone.0081022> Dabbs and Mallinger, 1999
- Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., & Hwang, Y. (2012). Cross-linguistic studies of children’s and adults’ vowel spaces. *The Journal of the Acoustical Society of America*, 131(1), 442–454. <https://doi.org/10.1121/1.3651823>
- Danielescu, A., Horowitz-Hendler, S. A., Pabst, A., Stewart, K. M., Gallo, E. M., & Aylett, M. P. (2023, April). Creating inclusive voices for the 21st century: A non-binary text-to-speech for conversational assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).
- Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders* (3rd ed.). Philadelphia, PA: W.B. Saunders Company
- Davies, S., & Goldberg, J. M. (2006). Clinical Aspects of Transgender Speech Feminization and Masculinization. *International Journal of Transgenderism*, 9(3–4), 167–196. [https://doi.org/10.1300/J485v09n03\\_08](https://doi.org/10.1300/J485v09n03_08)
- Dolquist, D. V. (2023). *Project POV: A Palette of Voices for Transmasculine Individuals* (thesis).
- Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4), 453-476. doi: 10.1111/j.1467-9841.2008.00374.x

- Evans, S., Neave, N., Wakelin, D., & Hamilton, C. (2008). The relationship between testosterone and vocal frequencies in human males. *Physiology & Behavior*, 93(4–5), 783–788. <https://doi.org/10.1016/j.physbeh.2007.11.033>
- Fairbanks, G. (1960). *Voice and articulation drillbook*, 2nd edn. New York: Harper & Row, 124-139.
- Farrow, M. (2019). *Transgender Vulnerabilities: State-Issued Identity Documents and Third Gender Options* (Doctoral dissertation).
- 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](https://doi.org/10.1016/S0892-1997(96)80009-9)
- Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeney, 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>
- Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Toda, & M. Zygis (Eds.), *Turbulent Sounds* (Vol. 21, pp. 281–302). De Gruyter Mouton. <https://doi.org/10.1515/9783110226584.281>
- Glaser, R., York, A., & Dimitrakakis, C. (2016). Effect of testosterone therapy on the female voice. *Climacteric*, 19(2), 198–203. <https://doi.org/10.3109/13697137.2015.1136925>
- Grollman, E. A. (2017). Multiple disadvantaged statuses and health: The role of multiple forms of discrimination. *Journal of Health and Social Behavior*, 58(3), 291-308.
- Hancock, A., Colton, L., & Douglas, F. (2014). Intonation and gender perception: Applications for transgender talkers. *Journal of Voice*, 28(2), 203–209. <https://doi.org/10.1016/j.jvoice.2013.08.009>
- Hancock, A. B., & Pool, S. F. (2017). Influence of listener characteristics on perceptions of sex and gender. *Journal of Language and Social Psychology*, 36(5), 599–610. <https://doi.org/10.1177/0261927x17704460>
- Havenhill, J. (2024). Articulatory and acoustic dynamics of fronted back vowels in American English. *Journal of the Acoustical Society of America*, 155(4): 2285-2301. <https://doi.org/10.1121/10.0025461>

- Heffernan, K. (2004). Evidence from HNR that /s/ is a social marker of gender. *Toronto Working Papers in Linguistics*, 23. <https://twpl.library.utoronto.ca/index.php/twpl/article/view/6208>
- Hope, M. & Lilley, J. (2020). Cues for Perception of Gender in Synthetic Voices and the Role of Identity. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020)*, (pp. 4143–4147).
- Hope, M. & Lilley, J. (2022). Gender expansive listeners utilize a nonbinary, multidimensional conception of gender to inform voice gender perception. *Brain and Language*, 224, 105049. <https://doi.org/10.1016/j.bandl.2021.105049>
- Hope, M. & Lilley, J. (2023). Differences in sibilant perception between gender expansive and cisgender listeners. *Seminars in Speech and Language*, 44(02). <https://doi.org/10.1055/s-0043-1761950>
- Hope, M., Ward, C., Lilley, J. (2023). Nonbinary American English speakers encode gender in vowel acoustics. *Proc. INTERSPEECH 2023*, 4713-4717, doi: 10.21437/Interspeech.2023-1772.
- IEEE. (1969). Harvard Sentences. Subcommittee on Subjective Measurements: IEEE Recommended Practices for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, 17, 227–46.
- Ito, K. & Johnson, L. (2017). The LJ Speech Dataset. <https://keithito.com/LJ-Speech-Dataset>
- James, S. E., Herman, J. L., Rankin, S., Keisling, M., Mottet, L., & Anafi, M. (2016). *The Report of the 2015 US Transgender Survey*. National Center for Transgender Equality.
- Jason, L., & Glenwick, D. (2016). *Handbook of Methodological Approaches to Community-Based Research : Qualitative, Quantitative, and Mixed Methods* (pp. 33–41). Oxford University Press.
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3), 1252–1263. <https://doi.org/10.1121/1.1288413>

- Junger, J., Habel, U., Bröhr, S., Neulen, J., Neuschaefer-Rube, C., Birkholz, P., Kohler, C., Schneider, F., Derntl, B., & Pauly, K. (2014). More than just two sexes: The neural correlates of voice gender perception in gender dysphoria. *PLoS ONE*, 9(11), e111672. <https://doi.org/10.1371/journal.pone.0111672>
- Junger, J., Pauly, K., Bröhr, S., Birkholz, P., Neuschaefer-Rube, C., Kohler, C., Schneider, F., Derntl, B., & Habel, U. (2013). Sex matters: Neural correlates of voice gender perception. *NeuroImage*, 79, 275–287. <https://doi.org/10.1016/j.neuroimage.2013.04.105>
- Henton, C. (1999). Where is female synthetic speech? *Journal of the International Phonetic Association*, 29(1), 51-61. <https://doi.org/10.1017/s0025100300006411>
- Kawitzky, D., & McAllister, T. (2020). The effect of formant biofeedback on the feminization of voice in transgender women. *Journal of Voice*, 34(1), 53–67. <https://doi.org/10.1016/j.jvoice.2018.07.017>
- Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, 110(2), 265-284. <https://doi.org/10.1037/0033-295X.110.2.265>
- Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, 74, 74–97. <https://doi.org/10.1016/j.jcomdis.2018.05.004>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13). <https://doi.org/10.18637/jss.v082.i13>
- Lammert, A. C., & Narayanan, S. S. (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLOS ONE*, 10(7), e0132193. <https://doi.org/10.1371/journal.pone.0132193>
- Langman, J., & Shi, X. (2020). Gender, language, identity, and intercultural communication. In *The Routledge Handbook of Language and Intercultural Communication* (pp. 219-233). Routledge.
- LeAnn, B., & Claire, P. L. (2022). Bright Voice Quality and Fundamental Frequency Variation in Nonbinary Talkers. *Journal of voice : official journal of the Voice Foundation*, S0892-1997(22)00234-X. Advance online publication. <https://doi.org/10.1016/j.jvoice.2022.08.001>

- Leung, Y., Oates, J., & Chan, S. P. (2018). Voice, Articulation, and Prosody Contribute to Listener Perceptions of Speaker Gender: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research: JSLHR*, 61(2), 266–297. [https://doi.org/10.1044/2017\\_JSLHR-S-17-0067](https://doi.org/10.1044/2017_JSLHR-S-17-0067)
- Levy, Tamaya. (2023). The Phonetics of Prejudice: Exploring Emotional and Racial Perceptions of African American Language and The 'Angry Black Woman' Stereotype. 10.13140/RG.2.2.30676.48002.
- Lieberman, M. D. (2007). The X-and-C-systems. *Social neuroscience: Integrating biological and psychological explanations of social behavior*, 290-315.
- Merritt, B., & Levi, S. V. (2023). Incorporating a gender expansive perspective into speech science pedagogy. 153(3\_supplement), A211–A211. <https://doi.org/10.1121/10.0018685>
- Mullennix, J. W., Stern, S. E., Wilson, S. J., & Dyson, C. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, 19(4), 407–424. [https://doi.org/10.1016/s0747-5632\(02\)00081-x](https://doi.org/10.1016/s0747-5632(02)00081-x)
- Munson, B. (2011). The influence of actual and imputed talker gender on fricative perception, revisited (L). *J Acoust Soc Am*;130(05):2631–2634
- Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., & Başkent, D. (2020). Development of voice perception is dissociated across gender cues in school-age children. *Scientific Reports*, 10(1), 1–11. <https://doi.org/10.1038/s41598-020-61732-6>
- Netzorg, Robin & Yu, Bohan & Guzman, Andrea & Wu, Peter & McNulty, Luna & Anumanchipalli, Gopala. (2024). Towards an Interpretable Representation of Speaker Identity via Perceptual Voice Qualities. 12391-12395. 10.1109/ICASSP48485.2024.10446197.
- Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *The Journal of the Acoustical Society of America*, 97(1), 520–530. <https://doi.org/10.1121/1.412278>
- Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The Emergence of Phonetic Segments: Evidence from the Spectral Structure of Fricative-Vowel Syllables Spoken by Children and Adults. *Journal of Speech, Language, and Hearing Research*, 32(1), 120–121. <https://doi.org/10.1044/jshr.3201.120>

- Paparini, S., Green, J., Papoutsis, C., Murdoch, J., Petticrew, M., Greenhalgh, T., Hanckel, B., & Shaw, S. (2020). Case study research for better evaluations of complex interventions: Rationale and challenges. *BMC Medicine*, 18(1). <https://doi.org/10.1186/s12916-020-01777-6>
- Patterson, M. L., & Werker, J. F. (2002). Infants' Ability to Match Dynamic Phonetic and Gender Information in the Face and Voice. *Journal of Experimental Child Psychology*, 81(1), 93–115. <https://doi.org/10.1006/jecp.2001.2644>
- Pisanski, K., Oleszkiewicz, A., Plachetka, J., Gmiterek, M., & Reby, D. (2018). Voice pitch modulation in human mate choice. *Proceedings of the Royal Society B: Biological Sciences*, 285(1893), 20181634. <https://doi.org/10.1098/rspb.2018.1634>
- Podesva, R. J. (2011). The California vowel shift and gay identity. *American Speech*, 86(1), 32-51. doi: 10.1215/00031283-1272339
- Prolific. (2023). Quickly Find Research Participants You Can Trust | Prolific. Version January 2023 - March 2023. <https://www.prolific.com/>
- Pribil, J., Pribilova, A., & Matousek, J. (2016). GMM-based speaker gender and age classification after voice conversion. 2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE). <https://doi.org/10.1109/splim.2016.7528391>
- Přibil, J., Přibilová, A., & Matoušek, J. (2018). Evaluation of speaker de-identification based on voice gender and age conversion. *Journal of Electrical Engineering*, 69(2), 138–147. <https://doi.org/10.2478/jee-2018-0017>
- R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Raphael, L. J., Borden, G. J., & Harris, K. S. (2012). *Speech science primer: Physiology, acoustics, and perception of speech*. LWW.
- Rechsteiner, J., & Sneller, B. (2023). The effects of topic and part of speech on nonbinary talkers' use of (ING). *University of Pennsylvania Working Papers in Linguistics*, 29(1), 19.
- Roy, N., Weinrich, B., Tanner, K., Corbin-Lewis, K., & Stemple, J. (2004). Replication, Randomization, and Clinical Relevance. *Journal of Speech, Language, and Hearing Research*, 47(2), 358–365. [https://doi.org/10.1044/1092-4388\(2004/029\)](https://doi.org/10.1044/1092-4388(2004/029))

- Rubin, M. (2017). Do p Values Lose Their Meaning in Exploratory Analyses? It Depends How You Define the Familywise Error Rate. *Review of General Psychology*, 21(3), 269–275. <https://doi.org/10.1037/gpr0000123>
- Saldana, J. (2009). *The Coding Manual for Qualitative Researchers*. United Kingdom: SAGE Publications.
- Schmid, M., & Bradley, E. (2019). Vocal pitch and intonation characteristics of those who are gender nonbinary. *Proceedings from the 19th International Congress of Phonetic Sciences*, 2685–2689.
- Schneider, F., Derntl, B., & Pauly, K. (2014). More than just two sexes: The neural correlates of voice gender perception in gender dysphoria. *PLoS ONE*, 9(11), e111672. <https://doi.org/10.1371/journal.pone.0111672>.
- Schwartz, M. F. (1968). Identification of a Speaker's Sex: A Fricative Study. *Journal of the Acoustical Society of America*, 43(5), 1178–1179. <https://doi.org/10.1121/1.1910954>.
- Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research*, 57(1), 285–296. [https://doi.org/10.1044/1092-4388\(2013\)12-0314](https://doi.org/10.1044/1092-4388(2013)12-0314)
- Smith, E., Junger, J., Pauly, K., Kellermann, T., Neulen, J., Neuschaefer-Rube, C., Derntl, B., & Habel, U. (2018). Gender incongruence and the brain – behavioral and neural correlates of voice gender perception in transgender people. *Hormones and Behavior*, 105, 11–21. <https://doi.org/10.1016/j.yhbeh.2018.07.001>
- Strand E.A. & Johnson K. (1996) Gradient and visual speaker normalization in the perception of fricatives. In: Gibbon D, ed. *Natural Language Processing and Speech Technology: Results of the 3rd KON- VENS Conference*, Bielefeld, October 1996. Boston, MA: De Gruyter Mouton; 14–26
- Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /S/ in Glaswegian (J. Cole & J. I. Hualde, Eds.; pp. 65–86). Mouton de Gruyter. <http://eprints.gla.ac.uk/8985/>
- Székely, É., Gustafson, J., Torre, I. (2023) Prosody-controllable Gender-ambiguous Speech Synthesis: A Tool for Investigating Implicit Bias in Speech Perception. *Proc. INTERSPEECH 2023*, 1234-1238, doi: 10.21437/Interspeech.2023-2086

- Tjaden, K., & Turner, G. S. (1997). Spectral Properties of Fricatives in Amyotrophic Lateral Sclerosis. *Journal of Speech, Language, and Hearing Research*, 40(6), 1358–1372. <https://doi.org/10.1044/jslhr.4006.1358>
- Tripp, A., & Munson, B. (2021). Perceiving gender while perceiving language: Integrating psycholinguistics and gender theory. *WIREs Cognitive Science*. <https://doi.org/10.1002/wcs.1583>
- Trans Student Educational Resources (TSER). (2015). “The Gender Unicorn.” <http://www.transstudent.org/gender>.
- van Bezooijen, R. (1995). Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women. *Language and Speech*, 38(3), 253–265. <https://doi.org/10.1177/002383099503800303>
- Weirich, M., & Simpson, A. P. (2018). Gender identity is indexed and perceived in speech. *PLoS ONE*, 13. <https://doi.org/10.1371/journal.pone.0209226>
- Wickham H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.
- Winn M.B. & Moore A.N. (2020) Perceptual weighting of acoustic cues for accommodating gender-related talker differences heard by listeners with normal hearing and with cochlear implants. *J Acoust SocAm*;148(02):496–510
- Wu, Z., Watts, O., & King, S. (2016). Merlin: an Open Source Neural Network Speech Synthesis System. In *SSW*, p. 202-207.
- Yuasa, I. P. (2008). *Culture and Gender of Voice Pitch: A Sociophonetic Comparison of the Japanese and Americans*. Equinox.
- Zandie, R., Mahoor, M. H., Madsen, J., & Emamian, Eshrat S. (2021). RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2106.08468>
- Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., & Wu, Y. (2019). LibriTTS: A corpus derived from Librispeech for Text-to-Speech. <https://ar5iv.labs.arxiv.org/html/1904.02882>
- Zimman, L. (2017). Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/. *Language in Society*, 46(3), 339–370. <https://doi.org/10.1017/S0047404517000070>

Zimman, L. (2021). Gender diversity and the voice. In *The Routledge handbook of language, gender, and sexuality* (pp. 69-90). Routledge.

## Appendix A

### PERMISSIONS

Hope, M., Ward, C., Lilley, J. (2023). Nonbinary American English speakers encode gender in vowel acoustics. *Proc. INTERSPEECH 2023*, 4713-4717, doi: 10.21437/Interspeech.2023-1772.

**“Can I place a copy of my paper in an institutional or other repository?**

Yes. For any paper published in the proceedings of INTERSPEECH or other ISCA sponsored events whose copyright is transferred to ISCA, ISCA grants each author permission to use the paper in that author's dissertation or in institutional and public repositories such as arXiv, provided that the paper is correctly cited, with DOI if available. This permission applies to all authors of the article. Authors are required to refer to the paper as specified in the ISCA Archive rather than refer to the institutional or public repository copy of the paper.” (<https://www.isca-archive.org/#about>)

Hope, M. & Lilley, J. (2023). Differences in sibilant perception between gender expansive and cisgender listeners. *Seminars in Speech and Language*, 44(02). <https://doi.org/10.1055/s-0043-1761950>

**“Permissions for a Thesis or Dissertation:** In non-open access papers published in a subscription journal, after assigning copyright, authors still retain the right to include their article, including the Version of Record without embargo, in their thesis, and permission from Thieme is not needed for this use as long as it remains strictly non-commercial.” (<https://www.thieme.com/en-us/journal-policies>)

## Appendix B

### IRB/HUMAN SUBJECTS APPROVAL



Institutional Review Board  
210H Hulihan Hall  
Newark, DE 19716  
Phone: 302-831-2137  
Fax: 302-831-2828

DATE: September 22, 2022

TO: Maxwell Hope, M.A.  
FROM: University of Delaware IRB

STUDY TITLE: [1866932-1] Creating and using gender expansive synthetic voices  
SUBMISSION TYPE: New Project

ACTION: APPROVED  
EFFECTIVE DATE: September 21, 2022  
NEXT REPORT DUE: September 20, 2023

REVIEW TYPE: Expedited Review  
REVIEW CATEGORY: Expedited review category # (6,7)

Thank you for your New Project submission to the University of Delaware Institutional Review Board (UD IRB). The UD IRB has reviewed and APPROVED the proposed research and submitted documents via Expedited Review in compliance with the pertinent federal regulations.

As the Principal Investigator for this study, you are responsible for, and agree that:

- All research must be conducted in accordance with the protocol and all other study forms as approved in this submission. Any revisions to the approved study procedures or documents must be reviewed and approved by the IRB prior to their implementation. Please use the UD amendment form to request the review of any changes to approved study procedures or documents.
- Informed consent is a process that must allow prospective participants sufficient opportunity to discuss and consider whether to participate. IRB-approved and stamped consent documents must be used when enrolling participants and a written copy shall be given to the person signing the informed consent form.
- Unanticipated problems, serious adverse events involving risk to participants, and all non-compliance issues must be reported to this office in a timely fashion according with the UD requirements for reportable events. All sponsor reporting requirements must also be followed.

The UD IRB REQUIRES the submission of a PROGRESS REPORT DUE ON September 20, 2022. A continuing review/progress report form must be submitted to the UD IRB at least 45 days prior to the due date to allow for the review of that report.

If you have any questions, please contact the UD IRB Office at (302) 831-2137 or via email at [hsrb-research@udel.edu](mailto:hsrb-research@udel.edu). Please include the study title and reference number in all correspondence with this office.

**INSTITUTIONAL REVIEW BOARD**

[www.udel.edu](http://www.udel.edu)

## Appendix C

### SUPPLEMENTAL INFORMATION

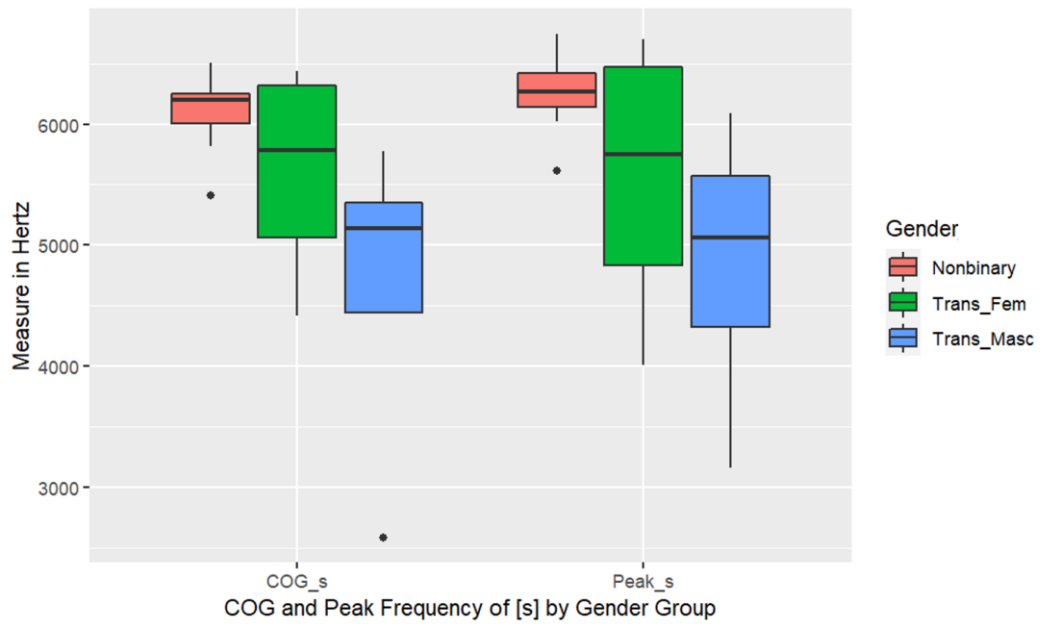


Figure S1: Boxplots showing the center of gravity (COG\_s) and peak frequency (Peak\_s) of [s] for the three gender groups.

Table S1. Correlation (Pearson's  $r$ ) results which differed in terms of statistical significance with and without the first author. (\* represents that the  $p$ -value of that correlation was statistically significant at the level of  $p < 0.05$ )

VOWEL (FORMANT)	GENDER VARIABLE	INCLUDING	EXCLUDING
[a] ( $f_1$ )	Other Gender	0.54 *	0.51
	Identity		
[o] ( $f_2$ )	Female Gender	0.47	0.53*
	Identity		
[o] ( $f_3$ )	Other Gender	0.50*	0.47
	Identity		
[e] ( $f_2$ )	Other Gender	0.55*	0.51
	Expression		

Table S2. Correlations per normalized formant per vowel, using the Johnson (2020) method. Bolded values represent Pearson's  $r$  correlations where  $r > 0.50$

		FEM	MALE	OTHER	FEM	MASC	OTHER
		GENIDENT	GENIDENT	GENIDENT	GENEXP	GENEXP	GENEXP
[i]	$f_1$	-0.17	-0.07	-0.18	0.22	-0.15	-0.31
	$f_2$	<b>-0.54</b>	0.43	0.36	<b>-0.64</b>	<b>0.54</b>	0.25
	$f_3$	-0.28	0.22	0.07	-0.47	0.32	-0.003
[a]	$f_1$	<b>-0.52</b>	<b>0.57</b>	0.32	-0.41	0.57	0.14
	$f_2$	0.25	0.16	-0.21	-0.25	0.22	-0.15
	$f_3$	-0.02	0.13	-0.16	-0.01	0.05	0.02
[u]	$f_1$	0.10	0.03	-0.41	0.27	-0.10	-0.46
	$f_2$	0.01	-0.07	0.01	-0.03	-0.07	-0.10
	$f_3$	0.11	-0.18	-0.01	-0.01	-0.10	0.05
[o]	$f_1$	0.10	-0.13	-0.08	<b>0.57</b>	-0.26	-0.23
	$f_2$	<b>0.80</b>	<b>-0.75</b>	-0.01	0.50	<b>-0.78</b>	0.17
	$f_3$	-0.36	0.36	0.02	0.32	0.34	0.09
[e]	$f_1$	0.37	-0.46	-0.31	<b>0.64</b>	<b>-0.57</b>	-0.19
	$f_2$	-0.34	0.27	0.33	-0.38	0.35	0.34
	$f_3$	-0.33	0.11	0.11	-0.44	0.26	0.11

## Demographic Survey from Chapter 2

1. What is your age? [write in]
2. What is your race? (use as many words as you'd like) [write in]
3. What is your gender (check all that apply)?
  - i. Man
  - ii. Woman
  - iii. Nonbinary
  - iv. Agender
  - v. Other [write in]
4. What is your gender (open ended version)? [write in]
5. On a scale of 0-100, how much do you feel your gender identity is female aligned? (0 being not female aligned at all and 100 being completely female aligned)
6. On a scale of 0-100, how much do you feel your gender identity is male aligned? (0 being not male aligned at all and 100 being completely male aligned)
7. On a scale of 0-100, how much do you feel your gender identity is aligned with something outside of "female" or "male"? (0 being not "other" aligned at all and 100 being completely "other" aligned)
8. On a scale of 0-100, how much do you feel your gender expression is feminine? (0 being not feminine at all and 100 being completely feminine)
9. On a scale of 0-100, how much do you feel your gender expression is masculine? (0 being not masculine at all and 100 being completely masculine)
10. On a scale of 0-100, how much do you feel your gender expression is something other than "feminine" or "masculine"? (0 being not "something other" at all and 100 being completely "something other")

11. What comes to mind when you think about speech and gender? What other parts of your identity intersect with your speech (if any)? (write as little or as much as you'd like) [write in]