Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

Title: Differences in sibilant perception between gender expansive and cisgender individuals

Corresponding author: Mx./Mr. Maxwell Hope, M.A.
Institution: University of Delaware
Department: Linguistics & Cognitive Science
Email: maxhope@udel.edu
Phone number: 215-983-8685

Co-author: Dr. Jason Lilley, PhD
Institution: The Nemours Foundation,
Department: Nemours Biomedical Research,
Nemours Speech Research Laboratory
Email: jason.lilley@nemours.org

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

Abstract

Acoustic cues of voice gender influence not only how people perceive the speaker's gender (e.g. whether that person is a man, woman, non-binary, etc.), but also how they perceive certain phonemes produced by that person. One such sociophonetic cue is the [s]/[ʃ] distinction in English; which phoneme is perceived depends on the perceived gender of the speaker. Recent research has shown that gender expansive people differ from cisgender people in their perception of voice gender and thus, this could be reflected in their categorization of sibilants. Despite this, there has been no research to date on how gender expansive people categorize sibilants. Further, while voice gender expression is often discussed within a biological context (e.g. vocal folds), voice extends to those who use other communication methods. The current study fills this gap by explicitly recruiting people of all genders and asking them to perform a sibilant categorization task using synthetic voices.

The results show that cisgender and gender expansive people perceive synthetic sibilants differently, especially from a "nonbinary" synthetic voice. These results have implications for developing more inclusive speech technology for gender expansive individuals, in particular for nonbinary people who use Speech Generating Devices (SGDs).

KEYWORDS: speech perception, gender, gender expansive, sibilant perception, speech generating devices

LEARNING OUTCOMES:

After reading this article, the learner will be able to:

- Summarize the current findings of gender's influence on sibilant production and perception
- Explain differences in sibilant perception between gender expansive and cisgender listeners
- Describe the relevant applications of gender expansive sibilant perception research

INTRODUCTION

Previous research on the sociophonetics of sibilant production and perception has largely been conducted using men's and women's voices[1,2,3]. Gender expansive (i.e. transgender and/or nonbinary) people have seldom been explicitly recruited and the experimental design has relied on stereotypical gender cues, which typically involve fundamental frequency (f_0) and formant frequencies in the "normal" men's and women's ranges[2,3], and when a neutral f_0 was used it was used in combination with neutral formant frequencies[4,5], obscuring potential effects of vocal tract parameters when pitch is neutral. Previous studies use the same sibilant continuum across vocal tract conditions, where the goal of the study is to elicit a compensation effect of perception from the listener; that is, if the listener hears an ambiguous sibilant in the middle of [s] and [ʃ], and is primed with a masculine gender cue, they are "compensating" in perception by shifting their perception to an [s] perception, for example. None of the previous studies investigated how a sibilant from each of these vocal tracts would be perceived itself, along its own continuum.

This study will consider the current literature on speech and gender perception while taking a novel approach to speech stimuli and perception ratings seeking to answer the question of how those who are gender expansive (GE) and those who are cisgender categorize sibilants when mean f_0 remains "gender neutral" and vocal tract characteristics, i.e. formant frequencies *and* spectral information, are taken from three different synthetic vocal tract configurations ("male", "female" and "nonbinary" vocal tracts). The applications are discussed in light of trans and nonbinary voice

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

therapy as well as the development of gender expansive synthetic voices for speech generating devices (SGDs).

GENDER AND SIBILANTS IN ENGLISH

Sibilant sounds, in particular [s] and [ʃ], are associated with gender differences and articulatory factors, such as how big the space is between the speaker's tongue and teeth. Women tend to produce [s] with the tongue closer to the teeth than men, leading to an increase in energy of the higher frequencies of the sibilant[6,7]. Flipsen and colleagues[6] found that women's peak frequency for [s] is around 6500-8100 Hz and men's peak frequency for [s] is around 4000-7100 Hz. Center of Gravity (COG), the average of the frequencies of a segment weighted by their amplitudes, also differs between genders, with men's COG of [s] being lower on average (around 5632 Hz, range of 4757-6167) than women's COG (around 6412 Hz, range 5727-6858 Hz)[7]. Both peak frequency and COG demonstrate a significant relationship with the articulatory measure of distance between teeth and tongue, however, Fuchs and Toda[7] found that the correlation was stronger for COG than for peak frequency. Specifically, as the distance between teeth and tongue increased for [s] production, the COG went down significantly. A study by Zimman[8] investigating sibilant production by trans masculine people found that COG varied significantly, and that those whose identity was more masculine (e.g. they identified strongly as a man) had lower COG than those whose identity was less masculine (e.g. those who identified as genderqueer or nonbinary). However, Zimman[8] 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, he noted that some queer trans men used low f_0 and high COG together as "stylistic bricolage" to signal their queer masculinity[8]. These studies demonstrate that people of various genders produce sibilants with varying COG and peak frequency. While there is notable variation within different social groups (men, women, genderqueer, etc.), COG tends to be different in systematic ways across groups. This has been shown to influence perception.

In a study by Munson[2], listeners were presented with ambiguous sibilant stimuli simultaneously with images of stereotypically male or female faces. Listeners who heard a man's or woman's voice with a vocal tract manipulated to be shorter were more likely to categorize a sibilant as [ʃ], and even more likely to categorize a sibilant as [ʃ] when the voice was paired with a woman's face. However, listeners were more likely to categorize a sibilant as [s] when a man's or woman's voice was manipulated to have a longer vocal tract regardless of the gender of the face the participants saw. This indicates that vocal tract length has an influence on sibilant categorization when the spectrum of the sibilants remains constant, and ambiguous, across vocal tract conditions. In the study by Winn and Moore[3], the authors manipulated f_0 and vocal tract length, finding a small and large effect on sibilant categorization respectively, but it should be noted that the only two f_0 conditions used were a mean f_0 of 104 Hz and a mean f_0 of 208 Hz, reflecting a "male" and "female" condition given that 104 Hz is well within the "male" pitch range and 208 Hz well within the "female" pitch range[9,10], and the vocal tract lengths were manipulated to reflect "masculine" and "feminine" vocal tract lengths. The authors did not use a "gender neutral" pitch or a vocal tract length in the middle of the "male" and "female" vocal tract lengths.

Thus, sibilant perception and categorization is reliant upon sociological factors such as perceived gender of the speaker, yet no one so far has investigated sibilant perceptions outside of a "male" or "female" framework. To date, studies into gender perception in speech still lack recruitment of gender diverse populations. Munson[2], for example, had only a total of only three men compared to

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

seventeen women in their pool of listeners. Therefore, even without explicitly considering people from between and beyond the gender binary, sibilant perception research may already be biased by a lack of participant diversity or consideration of effects of speaker or listener gender. More recent studies continue to lack in the recruitment of GE participants (e.g. Winn & Moore[3] only recruited male and female, presumably largely cisgender, participants). Finally, previous stimuli used natural productions of vowels as the basis for creating stimuli; sibilant categorization for a completely synthetic stimulus is yet unknown.

COMMUNICATING IDENTITY & GROUP MEMBERSHIP THROUGH SPEECH

Sociophonetic cues such as pitch[11], formant frequencies[11], prosody[12], voice quality[13], and spectral information of [s][14] can help to communicate identity and different group memberships. These group memberships may be based on age, race, gender, culture, or other important aspects of identity[15]. We may want to signal our group membership to affirm our identity, to show that we are not a part of another group, or both. People of various genders have multiple ways to signal their gender in speech, and gender expansive individuals may seek gender-affirming voice care if they do not feel they can accurately affirm their identity in speech. Not being able to signal group membership in speech may "out" a person as someone who is not an "in-group member" which can be undesirable, or dangerous, in certain scenarios.

> "As voice is one of the quickest ways we make judgments about people's genders, having a voice that is either 'too' conventionally feminine or masculine in relation to one's gender presentation could lead to these individuals being 'outed' as trans. Being 'outed' carries certain risks in that trans people experience higher levels of harassment, assault, or ridicule often resulting in further perpetuating a cycle of shame and fear."[16]

Individuals who use SGDs equally may be "outed" via their use of voice in undesirable, or dangerous, ways. Currently options on SGDs are limited, primarily coded as "male" or "female", and when not explicitly coded, still reproduce stereotypical feminine and masculine speech patterns. Some apps allow the shifting of pitch of the voice, and while this may alleviate some voice concerns, it does not address everything; just as with biological voice, there is more to voice gender than pitch; in fact pitch only accounts for about 41.6% of variation in voice gender perception ratings[11]. With limited options, nonbinary individuals who use SGDs may be unable to authentically express their gender. This may mean they are unable to effectively communicate their group membership with their voice to group members, they may not feel a congruence between their voice and their gender, or they may signal an undesirable identity to others with their voice. It is crucial that advances be made so that individuals who use SGDs can more accurately convey various aspects of their identity using their voice.

Nonbinary gender in speech generating devices: community information gathering

At the time of this writing, there has been no formal investigation into the intersections between nonbinary gender, speech, and the lived experiences of users of Speech Generating Devices (SGDs). In order to better frame this research, an investigation into the needs and concerns of nonbinary users of SGDs was conducted.

Ten nonbinary SGD users took part in an online survey which was IRB approved by the IRB of the University of Delaware. We asked separately about how well the voices on their SGDs captured their transition and their gender; "transition" in the survey question was indicated as

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

"transition (whatever this means to you, if it applies)". Half of the participants felt that the current voice gender options on their SGD do not capture well where they are in their transition and four said that they capture their transition somewhat well; none said that the options available to them captured their transition well (one indicated this was irrelevant to them). Seven of the 10 participants indicated that the current voice gender options do not capture their gender well, two indicated that the options capture their gender somewhat well, and none indicated that they captured their gender well (one indicated this was irrelevant to them). Several participants in the survey commented that while they had the option to shift the pitch of the synthetic voice, they did not feel this helped them capture their gender sufficiently.

Based on this investigation, it is clear that the current voice gender options on SGDs do not currently represent the voice genders of nonbinary users. Previous research has looked at how various communities use sociophonetic cues to signal in-group membership and convey gender identity, whether categorical or gradient, but as discussed above, the voices used in or explored in research almost always rely on male or female speech. Additionally, the research on voice gender perception has been extremely limited to the context of biological voice and has not yet considered synthetic voice a part of the discussion. Hope and Lilley[12,13] were the first to explore a multi-dimensional, nonbinary voice gender framework of perception which utilized synthetic voice as stimuli. However, that study was limited to subjective voice gender and did not examine how processing of phonemes themselves may be impacted by gender group and community membership.

THE CURRENT STUDY: HOW DO GENDER EXPANSIVE LISTENERS PERCEIVE SYNTHETIC SIBILANTS?

Given the differences in gender perception from voice, it is possible that sibilant perception and categorization is different for GE individuals than for cisgender individuals. This study seeks to examine how GE people categorize sibilants compared to cisgender people when the voices are kept in a "neutral" pitch range using synthetically made voices with varying vocal tract parameters (i.e. formant frequencies and spectral information). We use three different sets of vocal tract parameters based on the results from Hope and Lilley[12,13] which showed that GE people have a distinct "other" gender perception which is correlated with a nonbinary vocal tract created from averaging across male and female vocal tracts. A neutral pitch range was used for two reasons: 1) the results of Hope and Lilley[17,18] were based on stimuli with pitch held in the neutral range, and 2) this ensured that pitch would not sway results and obscure the detection of differences in categorization due to vocal tract. The f_o of the stimuli were also within the "neutral range" (145-175 Hz) according to Davies and colleagues[9].

First, we endeavor to identify differences in sibilant categorization between three sets of synthetic vocal tract parameters ("vocal tracts" or VTs for short), that is, 1) a "male" vocal tract with formant frequencies and spectral information taken from male speakers, 2) a "female" vocal tract with formant frequencies and spectral information taken from female speakers, and 3) a "nonbinary" vocal tract with formant frequencies and spectral information taken from a balanced combination of male and female speakers. Second, we seek to examine the differences between GE and cisgender listeners in terms of their categorization of sibilants across the three vocal tracts.

We hypothesized that, because GE individuals have shown a unique nonbinary voice gender perception in addition to female and male voice genders[18], the GE group would categorize sibilants differently between all three vocal tracts, whereas we hypothesized that cisgender listeners would only show a difference between the male and female vocal tracts, with their responses to the

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

nonbinary vocal tract being statistically indistinguishable from either of the other two. For listener group differences, we hypothesized that GE and cisgender listeners would categorize sibilants significantly differently from each other for the nonbinary vocal tract.

METHODS AND MATERIALS

Synthetic Voice Construction

A total of nine synthetic voices were created in Hope and Lilley[12] using speakers and recordings from the ModelTalker[19] database, via a speech synthesis system that modeled f_o-contour separately from the "vocal tract" (acoustic measures that model vocal tract characteristics such as formant frequencies and spectral information). Some voices were trained on speech from either 20 male or 20 female speakers, while others were trained on all 40 speakers to create "neutral" voices. In addition, some f_o-contour models were trained on modified f_o data such that the speaker mean f_o would match the global ("neutral") mean f_o, while preserving the relative f_o-contours. See Hope and Lilley[12] for more detailed explanation. Three of the resulting voices were used for this experiment:

1. a voice with a female vocal tract, female f_o-contour and average "neutral" f_o (FVT)
2. a voice with a male vocal tract, male f_o-contour and average "neutral" f_o (MVT)
3. a voice with a sample-averaged vocal tract, average f_o-contour, and average "neutral" f_o (NVT)

While we include the f_o-contour from the original voice here, note that the stimuli for this experiment were single words extracted from the same prosodic structure in the same sentence. Although the NVT voice was generated from the speech of 20 male and 20 female speakers – that is, none of them were identified as nonbinary – we refer to the NVT as the "Nonbinary" vocal tract below, because this voice was categorized as "nonbinary" by nonbinary listeners 100% of the time in the study by Hope and Lilley[13]; however, it should be noted that nonbinary people speak with a large range of voices.

Stimuli creation

For each synthetic voice, we extracted [s], [ʃ], and [i] from two Harvard sentences that had been generated as stimuli for Hope and Lilley[17,18]: "The birch canoe slid on the smooth planks" and "Glue the sheet to the dark blue background." We extracted the second [s] from the first sentence and the [ʃ] and [i] from the second sentence. Then, the sibilants were digitally mixed together using a modified script in PRAAT used in Phillips[20], creating a continuum from 0% [ʃ] to 100% [ʃ] at intervals of 10% (producing 11 sibilants) for each of the three voices. COGs (in Hz) for each step of the continuum for each vocal tract are listed in Table 1. The vowel was left unmodified from each base vocal tract and combined with the sibilant stimuli to create a "see" to "she" continuum. This resulted in 11 stimuli for each of the voices and therefore 33 stimuli total.

Participants: listeners

Participants over the age of 18 who were native speakers of American English were recruited online via email and social media to partake in a speech perception experiment. All responses were anonymous. A total of 80 participants completed the online experiment with 32 of them identifying as being a part of the GE community. One such participant was excluded for having responses that were quite dissimilar to the rest, including responses of [s] to even 100%-[ʃ] stimuli. This left 31 GE

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

participants (Age: M = 27.7, SD = 5.77, range = 18-45). Forty-eight participants identified as cisgender (Age: M = 32.6, SD = 9.11, range 18-60).

Speech perception survey

A speech perception experiment was conducted which was approved by the IRB of The University of Delaware. Using Qualtrics[21], participants were first presented with several screening questions: 1) whether or not they were a native speaker of American English, 2) whether or not they were wearing headphones, and 3) in lieu of asking if they had any speech or hearing disorders, they were asked what word they heard when given a word produced by the one of the synthetic voices. If they answered yes to the first two questions and accurately heard the word, they then answered demographic questions including age, languages spoken other than American English, and questions about their gender identity, including whether or not they were part of the gender expansive community (e.g. transgender and/or nonbinary).

All participants proceeded to a two-alternative forced choice (2AFC) task. In the first listening trial, they were presented with a screen in which they could repeatedly listen to the stimuli from the nonbinary vocal tract synthetic voice, one at a time (presented in a pre-randomized order that was the same for all participants). For each stimulus, they were asked to indicate whether they heard the word “see” or “she” before proceeding to the next stimulus. Then the entire process was repeated for the female vocal tract and male vocal tract, in that order. In determining the order, we were aware that a contrast effect may occur such that, for example, the nonbinary vocal tract stimuli may be perceived as more feminine if presented directly after the male vocal tract stimuli. Since we were particularly interested in the perception of the nonbinary stimuli, we presented the nonbinary synthetic voice first to avoid such effects.

Of the 79 participants, 67 (40 cis, 27 GE) participated in an additional task after the 2AFC task. In this task, called the “goodness” task, they were first presented with a screen on which they could listen to any and all of the [ʃ] stimuli for the nonbinary vocal tract condition in any order and any number of times, allowing them to compare the stimuli. They were asked mark each stimulus as either sounding like “she”, sounding like “see” or “too hard to decide”. Next they were asked to mark which of those that they selected as sounding like “she” sounded *most* like “she”, then repeated this for “see” and for whichever was too hard to categorize, if applicable. These tasks were then repeated for the other two vocal tracts. Finally, their final three “she”-like selections were presented together (one from each vocal tract condition), and they were asked to pick which one of those was the most “she”-like. This was repeated for “see”, and then if they had any that were too hard to categorize, they were asked which one was the most difficult to categorize. This allowed us to examine if there was a particular vocal tract which the participants found was the “best” sounding [s] and [ʃ] and to examine group differences in perception of which vocal tract was “best”. “Best” means from the task itself that the listener thought it was the most “she” or “see” like and thus can mean “most accurate”, but it also can reflect a preference of the listener for that voice and thus could reflect “most preferred”. For the GE group, we expected a fairly even distribution between the three vocal tract conditions, as Hope and Lilley[18] found that GE listeners have a third distinct voice gender category anchored in a nonbinary vocal tract. In contrast, we anticipated the cisgender participants would largely choose either the MVT voice or the FVT voice, potentially with a slight bias toward the MVT voice. This bias has been found especially in synthetic voices, where listeners have been shown to prefer male synthetic voices over female synthetic voices[22].

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

Statistical analyses

Statistical analyses were computed in R[23]. Because we did not explicitly control for age-related hearing loss, we conducted a simple binomial regression for the whole group to see if age had any effect on overall sibilant categorization. For the first set of results, overall chi-square statistics were computed to discover main effects of vocal tract and group (e.g. cisgender vs. gender expansive) on overall sibilant categorization. Post-hoc pairwise chi-square tests were then used to find significantly different categorizations between vocal tracts within each group. Chi-square statistics were also computed to examine group differences for each vocal tract. The gradient responses were modeled with logistic generalized additive mixture models (GAMMs[24,25]). Finally, for the sorting of stimuli in the "goodness" task, chi-squares were conducted to look at overall differences between groups and differences within groups for the "best" [s] and [ʃ]. Not enough people responded that the stimuli were "too hard to decide", so inferential statistics on this condition were not computed.

RESULTS

The binomial regression which examined if Age had an effect on sibilant categorization showed that in our sample, Age did not have a main effect on sibilant categorization (p = 0.494).

Two-Alternative Forced Choice: Categorization of sibilants

A chi-square of overall categorization revealed that cisgender listeners had a statistically significantly larger proportion of [ʃ] responses compared to GE listeners (p = .02). There were also statistically significant differences between the vocal tract conditions for the two groups; pairwise post-hoc chi-squares were conducted to look for differences. For the cisgender group, only the FVT and MVT vocal tract conditions were statistically significantly different in overall categorization (p = .001); the GE group had a statistically significant difference between the FVT and MVT (p = .002) and the NVT and MVT (p = .002) conditions, but not the NVT and FVT conditions. Figure 1 shows the overall categorization between cisgender and gender expansive listeners for the three vocal tract conditions. A chi-square to analyze main effect of listener gender group (woman, man, nonbinary) on overall sibilant categorization was conducted. There was no statistically significant main effect of gender; however, the proportion [ʃ] responses from nonbinary individuals was lower than the proportions for both men and women for all three vocal tract conditions (Fig 2). Categorization curves with proportion [ʃ] responses for the three different vocal tracts per group are shown in Fig 3. For both groups, the MVT voice has a much greater percentage of [ʃ] responses compared to the other two vocal tract voices.

Two-Alternative Forced Choice: GAMM analysis

To model our data, we considered a standard logistic regression model, but inspection of the data suggested that it would not sufficiently meet the model's assumption of linearity in the logit domain. So we instead used generalized additive mixture models[24,25,], which are a form of nonlinear mixture model that models nonlinear curves (called smooths) as the sum of a set of simpler basis functions. The binomial data were modeled with logistic GAMMs (using the logit link function). We used the *bam* function of Wood[26] to generate the models. The fixed effects were Group (GE or Cis), Vocal Tract (VT), and Percent [ʃ]. In addition, Participant was modeled as a random smooth effect with a per-VT interaction. To measure the significance of each fixed effect, we ran a chi-squares test comparing the model with all fixed effects to a model with the fixed effect removed (using the *compareML* function of van Rij and colleagues[27]). Each of these chi-square tests indicated that the

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

full model was a significantly better fit than the simpler model (X^2(9.00) = 11.309, *p* = .007 for Group; X^2(12.00) = 56.708, *p* < .0001 for VT; X^2(12.00) = 293.946, *p* < .0001 for Percent).

GAMMs are powerful – able to model and detect differences not only in overall effect means and slopes, but also curve shapes – but factor interactions are not easily computed from a full model. Since we were particularly interested in interactions between Group and Vocal Tract, we used GAMMs to model subsets of the data. For example, to measure the difference between GE and Cis responses to the Female VT, we modeled the data subset that excluded both Male and Nonbinary vocal tract stimuli, and used only Group and VT as main effects. To measure the effect of VT in this subset, we ran a chi-squares test comparing this model to a simpler model without the VT effect (using *compareML*). GAMMs have an advantage over binomial mixed effects regressions because they also allow us to examine the differences between groups and conditions at percentage ranges. We then inspected the curve modeling the difference between VT levels to determine at which values of Percent the two levels differ (using the *plot_diff* function from van Rij and colleagues[27]).

We found significant effects of Group overall, and with the Female and Nonbinary vocal tracts, but not for the Male vocal tract, as shown in Table 2.

Sibilant "goodness" task

For the sibilant "goodness" task in part two of the survey, there was a higher proportion of responses for the NVT as the best "see" or "she", recoded as best [s] or [ʃ], in the GE group (15%) than in the cisgender group (8%; see Fig 4). Furthermore, the responses in the GE group were exactly evenly split between the MVT and FVT, whereas in the cisgender group there were more responses for the MVT than the FVT. An overall chi-square on the unified responses across all stimuli showed no statistical significance between cisgender and gender expansive groups (X^2(2) = 2.03, *p* = .36). However, there were statistically significant differences within groups between vocal tracts chosen (for the cis group, *p* < .001; for the GE group, *p* = .02). The post-hoc pairwise chi-square tests showed that there was a statistically significant difference between the NVT and MVT, and between NVT and FVT, for best sibilant selection for the cisgender and gender expansive groups, but no statistically significant difference between the MVT and FVT stimuli for either group.

We further looked at whether there were differences between vocal tract choices for the "best" [s] and "best" [ʃ] stimuli separately for each group (see Fig 5). There were significantly more FVT choices for the stimuli categorized as "best" [ʃ] and significantly more MVT choices for the stimuli categorized as "best" [s]. This finding was significant for both the cisgender and GE groups.

DISCUSSION

Differences in sibilant categorization between groups

Overall categorization

The results from our vocal tract analysis showed that gender expansive listeners perceive synthetic [s] and [ʃ], sourced from three different vocal tract conditions, differently from cisgender listeners. Overall categorization showed that listener's gender itself (e.g. participant's identity as man, woman, nonbinary) did not have a statistically significant main effect on overall categorization; however, there were still notable trends. Nonbinary individuals had smaller proportions of [ʃ] responses for all three vocal tract conditions compared to men and women, and while women had the highest proportion [ʃ] responses for the Nonbinary vocal tract condition, men had the highest proportion [ʃ] responses for the Female vocal tract condition. Unlike gender itself, being part of the gender expansive community did show a statistically significant main effect on sibilant

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

categorization; gender expansive listeners perceived [s] more of the time, especially for the Nonbinary and Female vocal tract conditions. This means that there is a shared community experience among gender expansive individuals when it comes to sociophonetic cues embedded in sibilants.

Gradient sibilant categorization

For the Nonbinary vocal tract voice, cisgender listeners had less agreement on the [s] end of the continuum around what word they heard; even at 0% [ʃ], 25% of the responses were for [ʃ]. At 20 and 30% [ʃ] for the Nonbinary vocal tract, the gender expansive group had significantly more [s] responses than the cisgender group. This might indicate that gender expansive people are more sensitive to components of [s] in the acoustic signal. Furthermore, this result may be due to the lack of a distinct cognitive nonbinary voice gender category for cisgender listeners, leading to increased uncertainty in their perception. In both groups, more of the [ʃ] response proportions were in the middle of the scale (25-75%) for the Nonbinary vocal tract synthetic voice than for the Female and Male voices, indicating less agreement among listeners for the Nonbinary vocal tract stimuli. This highlights that although the gender expansive group perceived more [s] for the Nonbinary vocal tract, they still varied in their categorization as a group. This is important to emphasize since not all gender expansive people experience gender in the same way and perceptual differences in voice gender should be carefully considered in the context of voice therapy, for example.

Considering all the findings, the significant differences between groups hinge on the gender expansive listeners perceiving significantly more [s] than the cisgender listeners at the [s] end (i.e. 0% [ʃ]) of the sibilant spectrum from [s] to [ʃ]. This difference was evident in the Female and Nonbinary vocal tracts, but not for the Male vocal tract. A possible explanation for the group difference is that because pitch was held in the neutral range, the cisgender listeners no longer had that cue to rely on for categorization. In the study by Hope and Lilley[17], it was found that cisgender listeners perceived a Female vocal tract voice that had a higher pitch as significantly more feminine than a Female vocal tract voice with a neutral pitch, whereas this difference was not significant for the gender expansive listeners. Thus, with pitch neutral, the cisgender listeners perceived Female vocal tract voice in ways that were skewed towards a masculine perception of the sibilant (more [ʃ]), while the gender expansive listeners already perceived the Female vocal tract as feminine even with neutral pitch, and hence more likely to perceive [s] in the acoustic signal. Similarly, the greater percentage of [s] responses for the Nonbinary vocal tract condition for the gender expansive listeners may in part be due to a prior familiarity with this sort of vocal tract, because this vocal tract type is itself a distinct cognitive category for gender expansive people. Exposure to different types of voices (e.g. vocal tracts of varying length with neutral pitches) and association of these voices with various genders, rather than binning voices into binary categories, would explain how gender expansive people could perform more accurately compared to cisgender people. These findings also make sense in light of the lived experiences of gender expansive individuals who may keep their pitch neutral while using articulatory manipulations to convey aspects of their gender. Pitch may be a factor that could be difficult to manipulate for a given individual or it may not be desired to change pitch when a change in articulation or resonance is preferred. This may relate more broadly to a mixing and matching of sociophonetic cues that Zimman[8] refers to as “stylistic bricolage.” A key takeaway here is that gender expansive people mix and match auditory cues about speaker’s gender in ways that often look different from cisheteronormative standards that result in stereotypical feminine and masculine voice patterns.

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

Given the above explanation for the observed differences between cisgender and gender expansive listeners in perception at the [s] end of the continuum, one might question why we don't see a similar difference in perception at the [ʃ] end of the continuum, particularly with the Male vocal tract voice, where one might predict cisgender listeners to be more likely to perceive [s] due to the lack of pitch cue that would anchor them in a [ʃ] perception. A possible explanation lies in the observation that responses for all voices skewed towards a [ʃ] response, potentially obscuring differences in perception at the [ʃ] end of the continuum. This skew may be because our experimental design used synthetic voices instead of natural voices in combination with the fact that a high front vowel was used in the stimuli – [i] has been shown to shift sibilant perception to [ʃ][28], and our [i] was taken from the production of "sheet" so the vowel formants of [i] may have biased listeners toward [ʃ] overall. In the future, it will be essential for us to use a "sack" to "shack" continuum, with the vowel taken from "sack" as per Munson[4], as the vowel would not shift perception towards either end. Additionally, all of our stimuli had much lower COGs than previously used, especially towards the [s] end of the continuum, even for the Female vocal tract (see Table 1). It could be that compared to the averages found in many sociophonetic studies for the "typical female" [s] production, the speakers whose voices we used to train the synthetic voice models contained "outliers" who had lower COGs for their [s] production. However, this explanation does not negate the findings; in fact, it enhances the idea that gender expansive listeners have a broader range of what COGs may be considered "feminine" compared to cisgender listeners.

What's the "best" sibilant?

While the findings were not statistically significantly different, Fig. 4 shows that a larger percentage of the gender expansive group chose the Nonbinary vocal tract sibilant as their "best" sibilant in the sibilant "goodness" task. Additionally, while not statistically significant, more of the cisgender group chose the Male vocal tract sibilant over the Female vocal tract sibilant (50% versus 42%) as their best sibilant, while the gender expansive group did not (43% for both). When looking at the "best" sibilants for [s] and [ʃ] independently, there is a large preference for the Male vocal tract for "best" [s] and conversely a large preference for the Female vocal tract for "best" [ʃ]. While our stimuli were quite different from previous studies, and our overarching goal was also different, our sibilant "goodness" task reveals findings in line with past sibilant production research, namely that listeners compensate for gender in perception of sibilants. When comparing vocal tracts one against another, the Male vocal tract produces the "best" [s] because listeners know that male voices in general tend to produce sibilants with a relatively low COG (hence "more [ʃ]-like") and "compensate" in perception by more strongly associating a relatively high COG for the Male vocal tract (in this case 4263 Hz) with [s]. This is especially true when comparing this [s] side by side with, for example, the [s] from the Female vocal tract, which has a COG of 5601 Hz; this is not very high for a female vocal tract [s] since female voices tend to produce [s] with relatively high COGs. Conversely, when comparing stimuli side by side, listeners know female voices tend to produce [ʃ] at a relatively high COG (hence "more [s]-like"), so a sibilant with a COG at 3914 Hz is relatively low for a female vocal tract to produce, and thus very [ʃ]-like. In other words, it is possible that the listeners are subconsciously focusing on which production is most likely to be a *stereotypical* articulation of that sibilant based on the perception of the voice as a man or a woman. Using an example from above, the [s] from the Female vocal tract would not be an stereotypical articulation for most listeners when compared next to the [s] from the Male vocal tract because the [s] from the Female vocal tract had a relatively low COG for what women tend to produce for [s] – in an articulation, this would mean that the Female [s] would have been produced with the tongue further

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

back in the mouth than what is typically expected for this group. Thus, even though the COG is higher for the Female [s] than the Male [s], the Male [s] is closer to a stereotypical [s] articulation for male speakers, who tend to produce [s] with the tongue further back in the mouth, hence a relatively low COG, and represents a more stereotypical articulation of the sibilant given the inferred gender from the vocal tract. However, it is important to keep in mind that these stereotypical articulations based on perceptions of the voice as a man or a woman are often based on a perception of the voice as belonging to a *certain kind of man or woman* – e.g., one that adheres to a white, cisgender, and/or heteronormative way of speaking. Results may have been different if the voices we used to build the synthetic speech consisted of more racially diverse, gender expansive, or queer speech.

One potential reason that fairly few listeners chose a Nonbinary vocal tract stimulus as the "best" sibilant could be because the nonbinary synthetic voice was constructed using 40 natural voices, double the number of natural voices used for the other two synthetic voices. This means the model training had to deal with significantly more variation in speech, possibly resulting in worse voice quality such that when the voices were ranked against each other, the male and female voices were chosen more frequently as ideal because of this difference. Another reason that gender expansive listeners may not have chosen the Nonbinary vocal tract is that the natural voices used to train it were not actual nonbinary or gender expansive voices. In the end, these were averaged vocal tract voices sourced from 20 male and 20 female speakers. Therefore, while gender expansive people do show sensitivity to a synthetic average vocal tract, the voice we presented to them lacks various sociophonetic aspects of gender expansive speech. Going forward, it will be crucial to construct more authentic gender expansive and nonbinary synthetic voices for use in experimentation as well as to improve voice quality for gender expansive users of speech generating devices.

Clinical applications

One clear application of this study is the recognition of synthetic speech as a domain of voice which can situate itself in the realm of gender-affirming voice care. Speech scientists can learn from this study how to create new voices, and clinicians can learn how to broaden conception of voice to include synthetic speech. First is acknowledging the limitations that currently exist in synthetic speech. Most of the commercially available SGDs and apps have default voices that are, if not explicitly, then implicitly, coded as male or female. These options can be limited for nonbinary users of SGDs. Even if they are able to shift pitch, a service which is available on some devices and applications, this may not solve their voice concerns; vocal tract characteristics like formant frequencies and COG – properties that reflect vocal tract size and shape – are not often able to be manipulated in SGDs.

Even though this study was conducted with synthetic speech, it is possible that these results could have future clinical applications for gender-affirming voice care in the future. There are no studies that we know of which look at the manipulation of the frontal cavity (e.g. distance of tongue to teeth for [s]) for gender affirmation; but as explored in Fuchs & Toda[7], women regardless of palate size moved their tongues more forward to produce an [s] compared to men; that is, women all increased position of the tongue towards the teeth, decreasing the frontal cavity to produce a more feminine [s] (although notably, they did not find that men did not move their tongues more backward to produce a more masculine [s]). As noted in previous studies, a more forward position of the tongue toward the teeth during [s] production correlates with a higher COG[7], which itself correlates with higher perceived femininity[2,3,4,5]. Additionally, while there have been no specific studies on gender-affirming voice and sibilant production, some investigations have looked more

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

broadly into vocal tract manipulations including how "spreading the lips wider and bringing the tongue more forward" correlated with increased femininity[29]. Therefore, the results of this study could provide one additional tool for gender-affirming voice care, depending on the goals of the client. However, an important insight of this study is that the ultimate perception of [s] is bound to social factors including group membership and therefore, it will be pertinent for SLPs to ask their clients what sorts of voices they want to use and in what contexts (e.g. clients may want to explicitly signal GE group membership in some cases, or not in others).

Whether one is using synthetic voice or biological voice, we should all have the options to mix and match sociophonetic cues such as pitch and vocal tract characteristics. While this is becoming more accessible in voice care for certain individuals, more strides are needed in the realm of SGDs.

CONCLUSIONS

This study has shown that, in American English, gender expansive and cisgender participants perceive synthetic sibilants differently across three different vocal tract configurations when pitch is held in a neutral range. The results indicate that the two groups may use different perceptual strategies during sibilant categorization. These differences may provide insight into how experience with and exposure to diverse voice genders impact speech perception. Gender expansive individuals tend to be exposed to a variety of voices that are associated with a variety of genders, instead of grouping voices into men/women or masculine/feminine binary categories. Speech-Language Pathologists working with gender expansive individuals may wish to consider incorporating sociophonetic cues such as sibilant production, while keeping in mind that sibilant perception varies between gender expansive and cisgender individuals.

In the future, the results of this study could provide a basis for testing the effects of visual gender information (such as presenting faces with varying perceived genders) or written information (such as having listeners read a small paragraph about the speaker that contributes to inferred gender, such as in its use of pronouns) on sibilant perception for gender expansive individuals. Finally, this study has implications for the development of inclusive SGDs for gender expansive individuals. Because voice gender and sociophonetic cues are different for gender expansive people, offering a broader range of voice gender options for speech generating devices would increase a user's ability to convey their voice more authentically.

ACKNOWLEDGEMENTS

The authors would like to thank Mx. Benjamin Munson for help in interpretation of the data and Dr. Timothy Bunnell for his generous support in using the ModelTalker database and advice on technical aspects of the voice creation.

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

REFERENCES

[1] Strand EA, Johnson K. Gradient and visual speaker normalization in the perception of fricatives. In: Gibbon D, ed. *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996.* Boston, MA: De Gruyter Mouton; 1996:14-26. doi:10.1515/9783110821895-003

[2] Munson B. The influence of actual and imputed talker gender on fricative perception, revisited (L). *J Acoust Soc Am.* 2011;130(5):2631-2634. doi:10.1121/1.3641410

[3] Winn MB, Moore AN. Perceptual weighting of acoustic cues for accommodating gender-related talker differences heard by listeners with normal hearing and with cochlear implants. *J Acoust Soc Am.* 2020;148(2):496-510. doi:10.1121/10.0001672

[4] Munson B, Ryherd K, Kemper S. Implicit and explicit gender priming in English lingual sibilant fricative perception. *Linguistics.* 2017;55(5):1073-1107. doi:10.1515/ling-2017-0021

[5] Bouavichith DA, Beddor PS, Tobin SJ, Hildebrandt T, Craft JT, Calloway I. Perceptual influences of social and linguistic priming are bidirectional. In: Escudero P, Warren P, Tabain M, Calhoun S, eds. *Proc Int Conf Phon Sci.* Melbourne, Australia; 2019:1039-1043.

[6] Flipsen P, Shriberg L, Weismer G, Karlsson H, McSweeny J. Acoustic characteristics of /s/ in adolescents. *J Speech Lang Hear Res.* 1999;42(3):663-677. doi:10.1044/jslhr.4203.663

[7] Fuchs S, Toda M. Do differences in male versus female /s/ reflect biological or sociophonetic factors? In: Fuchs S, Toda M, Zygis M, eds. *Turbulent Sounds: An Interdisciplinary Guide.* New York, NY: De Gruyter Mouton; 2010:281-302. doi:10.1515/9783110226584.281

[8] Zimman L. Variability in /s/ among transgender speakers: Evidence for a socially grounded account of gender and sibilants. *Linguistics.* 2017;55(5):993-1019. doi:10.1515/ling-2017-0018

[9] Davies S, Papp VG, Antoni C. Voice and communication change for gender nonconforming individuals: Giving voice to the person inside. *Int J Transgend*. 2015;16(3):117-159. doi:10.1080/15532739.2015.1075931

[10] Schneider S, Courey M. Transgender voice and communication – vocal health and considerations. UCSF Gender-affirming Health Program. Published June 17, 2016. https://transcare.ucsf.edu/guidelines/vocal-health

[11] Leung Y, Oates J, Chan SP. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. Journal of Speech, Language, and Hearing Research. 2018;61(2):266-297. doi:10.1044/2017_jslhr-s-17-0067

[12] Hancock A, Colton L, Douglas F. Intonation and gender perception: Applications for transgender speakers. *Journal of Voice*. 2014;28(2):203-209. doi:10.1016/j.jvoice.2013.08.009

[13] Podesva RJ, Callier P. Voice quality and identity. Annual Review of Applied Linguistics. 2015;35:173-194. doi:10.1017/s0267190514000270

[14] Mack S, Munson B. The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'Gay Lisp' Stereotype. Journal of Phonetics. 2012;40(1):198-212. doi:10.1016/j.wocn.2011.10.002

[15] Hall-Lew L, Moore E, Podesva R. Social Meaning and Linguistic Variation: Theorizing the Third Wave. Cambridge, United Kingdom: Cambridge University Press; 2021.

[16] Venier C. Voice Feminization Therapy and Quality of Life in Transgender Women: A Critical Review and Case Study. 2017. Accessed from: https://www.uwo.ca/fhs/lwm/teaching/EBP/2016-17/Venier.pdf

[17] Hope M, Lilley J. Cues for perception of gender in synthetic voices and the role of identity. *Interspeech.* 2020 Oct;2020:4143-4147.

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

[18] Hope M, Lilley J. Gender expansive listeners utilize a non-binary, multidimensional conception of gender to inform voice gender perception. *Brain Lang.* 2022;224:105049. doi:10.1016/j.bandl.2021.105049

[19] Bunnell HT, Lilley J, McGrath K. The ModelTalker project: A web-based voice banking pipeline for ALS/MND patients. *Interspeech.* 2017 Aug;2017:4032-4033.

[20] Phillips JB. *Sibilant Categorization, Convergence, and Change: The Case of /s/-Retraction in American English.* Dissertation. University of Chicago; 2020.

[21] *Qualtrics Survey Software.* Qualtrics XM. https://www.qualtrics.com/core-xm/survey-software/

[22] Mullennix JW, Stern SE, Wilson SJ, Dyson C. Social perception of male and female computer synthesized speech. *Comput Human Behav.* 2003;19(4):407-424. doi:10.1016/s0747-5632(02)00081-x

[23] R Core Team. *R: A language and environment for statistical computing.* (Version 4.0.3, 2020). R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[24] Hastie TJ, Tibshirani RJ. *Generalized Additive Models (Monographs on Statistics and Applied Probability 43).* London, UK: Chapman & Hall/CRC; 1990.

[25] Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Series B Stat Methodol.* 2011;73(1):3–36.

[26] Wood SN. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.* R package version 1.8-33 (2020).

[27] van Rij J, Wieling M, Baayen RH, van Rijn H. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* R package version 2.4 (2020).

[28] Mann VA, Repp BH. Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Percept Psychophys.* 1980;28(3):213–228. doi:10.3758/bf03204377

[29] Carew L, Dacakis G, Oates J. The effectiveness of oral resonance therapy on the perception of femininity of voice in male-to-female transsexuals. *Journal of Voice*. 2007;21(5):591-603. doi:10.1016/j.jvoice.2006.05.005

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

FIGURES AND TABLES

Box figures

Gaps in Current Practice and Research

- Previous research based largely on male and female speakers and listeners; unclear how gender expansive people perceive sibilants
- No investigation into perception of sibilants and speaker's gender using completely synthetic stimuli
- Current practice of gender-affirming voice has largely ignored needs of nonbinary speakers (including SGD users) and listeners

Key Findings

- When the acoustic signal is more like [s] than [ʃ] for a Female and Nonbinary synthetic voice, GE listeners will perceive it as [s] more often than cisgender listeners will
- GE listeners chose a Nonbinary synthetic sibilant as a "best" sibilant more frequently and had a more equal preference between the Male and Female synthetic voices

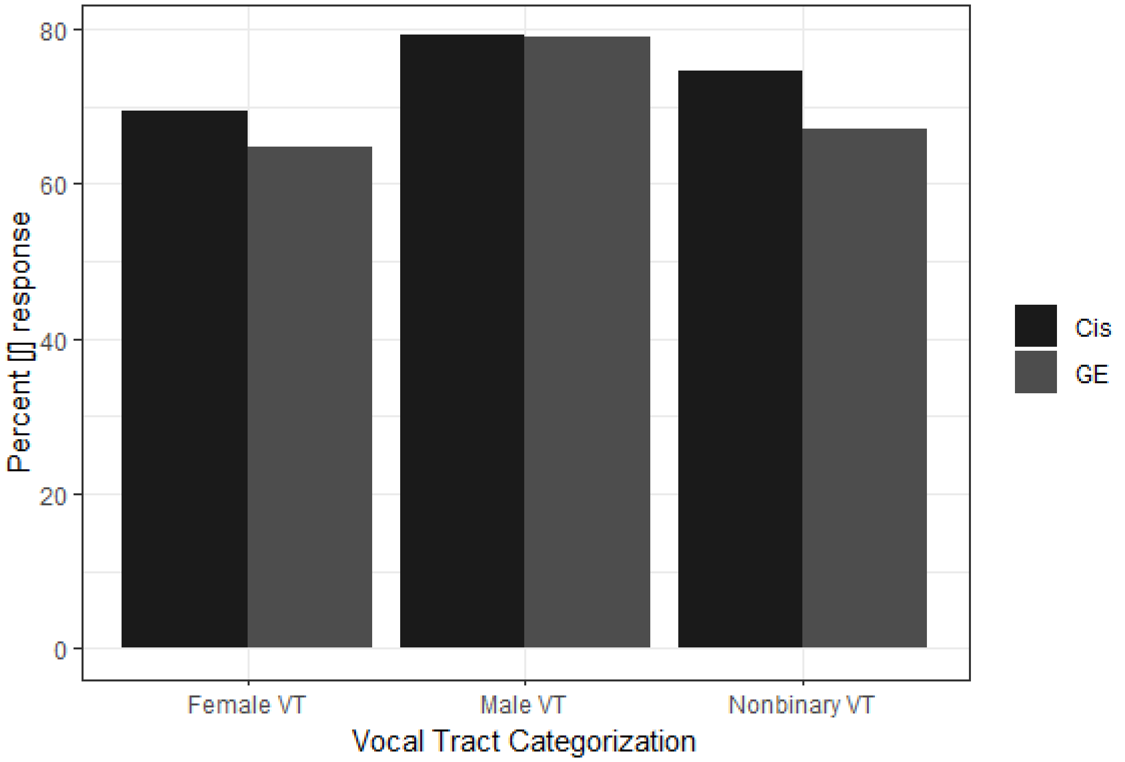


Fig 1: *Overall categorization across vocal tract conditions between gender expansive (GE) and cisgender (Cis) listeners. The y-axis represents percent [ʃ] responses.*

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

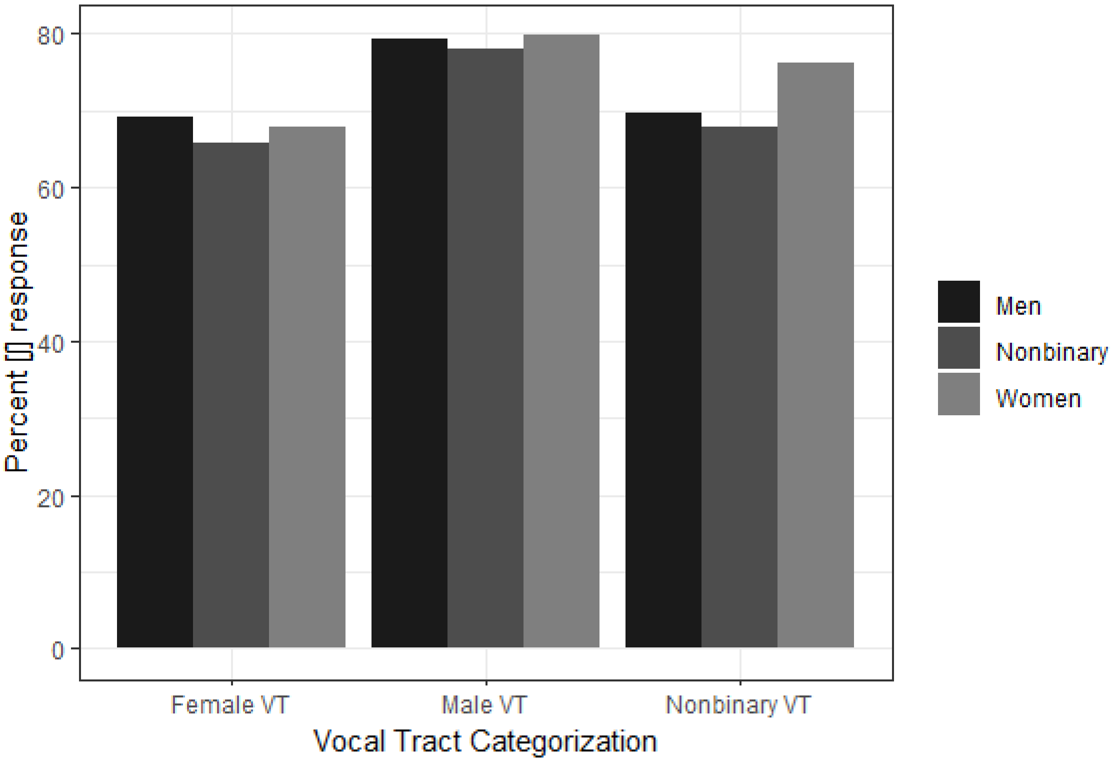
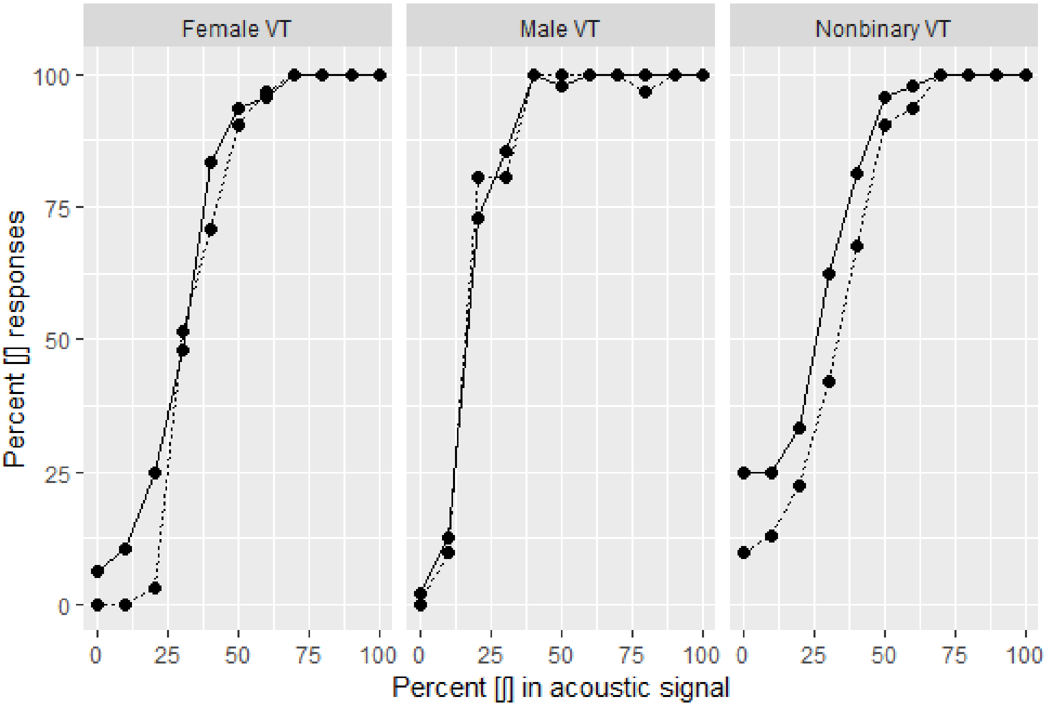


Fig 2: *Overall categorization across vocal tract conditions between listener gender groups. The y-axis represents percent [ʃ] responses.*



Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

Fig 3: *Cis (solid line) and GE (dotted line) overall sibilant categorization for the three vocal tract conditions. The x-axis represents the percent-[ʃ] in the acoustic signal and the y-axis shows percent [ʃ] responses.*

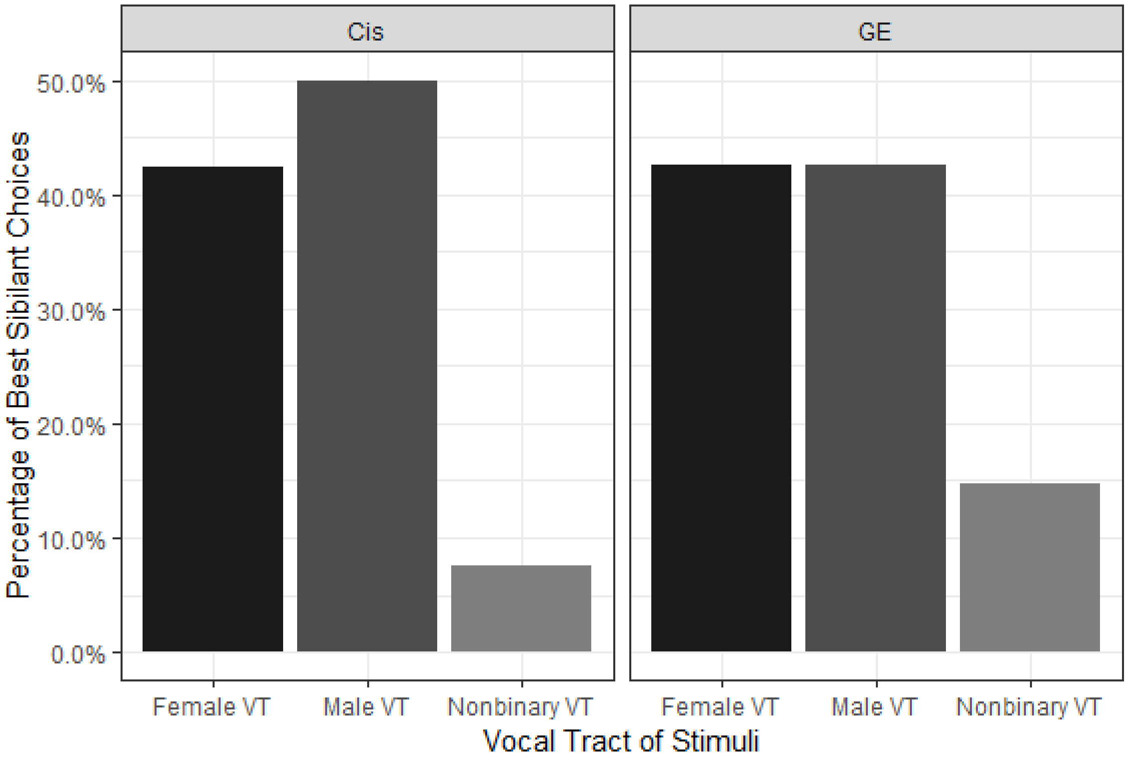


Fig 4. Percentage (%) responses of the vocal tract choices for the "best" sibilant stimuli derived from the categorizations of all the stimuli from the sibilant "goodness" task, split between the two groups (cisgender on the left and GE on the right). The percentages in each box add up to 100%.

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

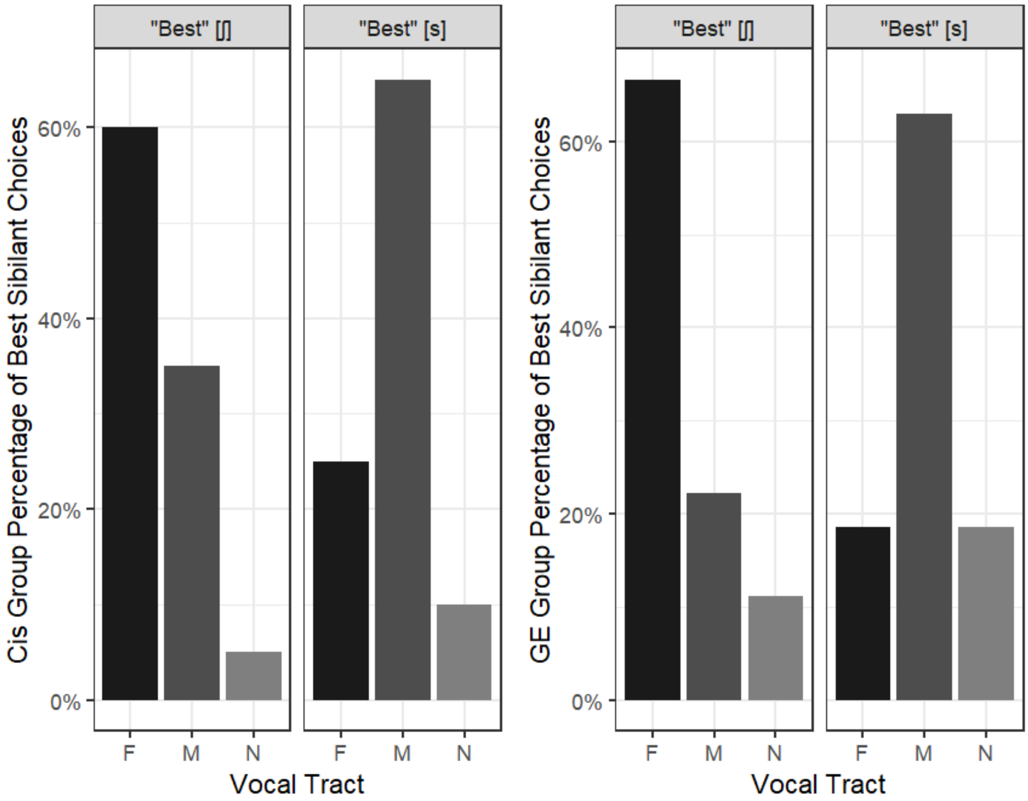


Fig 5. Percentage (%) responses of vocal tract choices for “best” [s] and “best” [ʃ] stimuli in the sibilant “goodness” task. The cisgender group is shown in the two leftmost bar graphs and the gender expansive (GE) group is shown in the two rightmost.

Table 1. *Centers of Gravity (COGs) in hertz for the three different vocal tracts for the 11-step continuum from [s] to [ʃ]. Numbers in the column headings represent percent [ʃ] in the acoustic signal.*

VT	*[s]-0*	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*[ʃ]-100*
Male	4263	4218	4070	3871	3705	3601	3546	3520	3511	3510	3468
Nonbinary	4654	4565	4380	4066	3958	3858	3804	3778	3767	3764	3766
Female	5601	5344	4810	4392	4158	4039	3979	3949	3934	3928	3914

Table 2. *GAMM comparison results. Column 4 indicates the percent-[ʃ] measurement points at which the GE and cisgender groups were significantly different.*

Comparison	***X^2***	***p-value***	***diff. points***

Accepted Manuscript
Version of Record at: https://doi.org/10.1055/s-0043-1761950

GE vs Cis	5.050	.018 *	10, 20
FVT subset	6.662	.004 **	0, 10, 20
NVT subset	5.115	.016 *	20, 30
MVT subset	3.005	.111	--