2-15-2024

Auditory Free Classification of Gender Diverse Speakers

Brandon Merritt
bmmerritt@utep.edu

Tessa Bent
Indiana University - Bloomington, tbent@indiana.edu

Rowan Kilgore

Cameron Eads

Comments:

This open access articles was published in The Journal of the Acoustical Society of America. A popular media synopsis of this work was published in Scilight (https://doi.org/10.1063/10.0024759). The first author was interviewed about this work on Across Acoustics, the official podcast of the Acoustical Society of America (https://acrossacoustics.buzzsprout.com/1537384/14466368).

JASA ARTICLE

Auditory free classification of gender diverse speakers[a]

Brandon Merritt,[1,b] Tessa Bent,[2] Rowan Kilgore,[2] and Cameron Eads[2]
[1]*Department of Speech, Language, and Hearing Sciences, The University of Texas at El Paso, El Paso, Texas 79968, USA*
[2]*Department of Speech, Language and Hearing Sciences, Indiana University, Bloomington, Indiana 47408, USA*

ABSTRACT:
Auditory attribution of speaker gender has historically been assumed to operate within a binary framework. The prevalence of gender diversity and its associated sociophonetic variability motivates an examination of how listeners perceptually represent these diverse voices. Utterances from 30 transgender (1 agender individual, 15 non-binary individuals, 7 transgender men, and 7 transgender women) and 30 cisgender (15 men and 15 women) speakers were used in an auditory free classification paradigm, in which cisgender listeners classified the speakers on perceived general similarity and gender identity. Multidimensional scaling of listeners' classifications revealed two-dimensional solutions as the best fit for general similarity classifications. The first dimension was interpreted as masculinity/femininity, where listeners organized speakers from high to low fundamental frequency and first formant frequency. The second was interpreted as gender prototypicality, where listeners separated speakers with fundamental frequency and first formant frequency at upper and lower extreme values from more intermediate values. Listeners' classifications for gender identity collapsed into a one-dimensional space interpreted as masculinity/femininity. Results suggest that listeners engage in fine-grained analysis of speaker gender that cannot be adequately captured by a gender dichotomy. Further, varying terminology used in instructions may bias listeners' gender judgements.
https://doi.org/10.1121/10.0024521

(Received 28 July 2023; revised 5 January 2024; accepted 6 January 2024; published online 15 February 2024)

[Editor: Jody Kreiman] Pages: 1422–1436

I. INTRODUCTION

Investigations of auditory gender attribution (i.e., the gender label or category that a listener applies to a speaker) have historically operated within a gender binary framework.[1–7] Within this framework, speakers are assumed to represent cisgender identity (i.e., girls/women and boys/men whose gender identity is congruent with the sex assigned to them at birth) with speech characteristics associated with one of these two gender categories. However, contemporary understandings of gender acknowledge that it is a multifaceted construct. Gender encompasses gender identity, gender expression, and social gender roles, all of which are tied to cultural understandings and expectations of people who were assigned male or female at birth.[8] Estimates indicate that more than 1.4 million adults in the United States identify as transgender (i.e., individuals whose gender identity is incongruent with the sex assigned to them at birth), with this number expected to continue growing.[9] A comprehensive understanding of how listeners attribute gender to speakers must, therefore, be inclusive of gender diverse individuals.

Numerous studies suggest that the speech characteristics of transgender speakers may not align with those typical of cisgender men and women.[6,10–13] Transgender individuals may even purposefully adopt systematic ways of speaking to communicate a queer identity to listeners.[10,12,13] For example, transmasculine speakers (i.e., individuals assigned female at birth but who identify along the masculine spectrum) may manipulate speaking fundamental frequency (f_o) in tandem with /s/ center of gravity to convey their gender diversity.[13] Non-binary speakers (i.e., individuals whose self-concept of gender identity does not align with a binary system) may also use structured variation in f_o, vocal quality, and formant frequencies to communicate an identity distinct from cisgender and transgender men and women.[10,12] By combining acoustic-phonetic features associated with differing gender identities, these speakers may be asserting identities that intentionally challenge the diametric sociocultural norms associated with the speech of cisgender men and women.

Because individuals with diverse gender identities are increasingly visible in contemporary society, listeners may be exposed to more diverse social landscapes that introduce potentially novel acoustic-phonetic variability to which they must adapt. The stylistic nuance that is demonstrated in emerging literature on gender diversity and speech communication suggests a more complex construction and communication of gender than a binary framework allows or acknowledges. Thus, the "male/man" versus "female/woman" categorization task that is pervasive in research on gender

[a]Parts of this paper were presented in "Auditory free classification of gender diverse speakers," 183rd Meeting of the Acoustical Society of America, Nashville, TN in December 2022.
[b]Email: bmmerritt@utep.edu

https://doi.org/10.1121/10.0024521

attribution may poorly capture the production and perception of speech variability across speakers of diverse gender identities. Attempts to partition all speakers, and the gender attributed to them, into one of these two categories may, therefore, be reductionist and disregard potentially important acoustic-phonetic information that informs listeners' cognitive representation of speaker gender.

Although there is increasing representation of gender diversity in the scholarship of speech communication,[6,10,11,13–15] much of this work has focused on the role of the speaker in communicating gender. Less attention has been paid to listeners' perceptual representation of the acoustic-phonetic variability that exists among speakers of diverse gender identities. It is, however, important to examine the interpretative work that is required of listeners in constructing indexical meaning from the acoustic speech signal. The process of *enregisterment* (i.e., distinct forms of speech being recognized as belonging to specific social groups) requires both a speaker to produce acoustic-phonetic variation and a listener to attach social meaning to it.[16] This process aligns well with Azul and Hancock's socio-culturally mediated model of gender attribution,[17] where listeners take an active role in interpreting gender from the acoustic signal, and is the perspective from which this study approaches gender attribution.

Additionally, listeners may condition speech perception on indexical categories attributed to the speaker, such as gender.[18–21] For example, Strand and Johnson[18] presented listeners with an identical fricative token paired with a video of either a cisgender man or woman speaker. When listeners heard the fricative token paired with a cisgender man, with an expected longer vocal tract and lower peak frequency, they were more likely to identify the token as /s/. However, when the same token was paired with a cisgender woman, with an expected shorter vocal tract and higher peak frequency, listeners were more likely to identify it as /ʃ/. In a follow-up study, listeners also tended to perceive different vowels both when primed with pictures of men and women and when asked to visualize a man or woman producing the token.[19] These results suggest that socially conditioned expectations for speech characteristics of speakers with varying gender identities shifted listeners' perceptual phoneme boundaries. Similar findings have been found based on attribution of a speaker's geographic origin, age, socioeconomic status, and sexual orientation.[22–25] Hence, multiple sources of indexical variation are embedded within the speech acoustic signal that require listeners to attribute variation to its correct source and recognize similarities and differences across speakers. Understanding listeners' perceptual organization of indexical categories, such as gender, would therefore clarify the cognitive mechanisms that underlie speech perception more broadly.

A. Limitations of traditional approaches to measure gender attribution

Through cumulative exposure to the acoustic-phonetic variability found across speakers of varying gender identities, listeners may develop models of speech characteristics typical for speakers with gender identities beyond a binary schematic organization.[20,21] Thus, dichotomous gender categorization tasks cannot capture listeners' perception and representation of the fine phonetic detail that underlies their attribution of gender, particularly for those speakers who do not identify as cisgender men or women.[10–13] For example, Mullennix *et al*.[3] tasked listeners with identification and ABX discrimination of speaker gender, among other perceptual tasks, from brief vowel tokens. Analyses of listeners' responses demonstrated a gradually sloping identification function and no sharp boundary between "male" and "female" categories. The authors concluded that the perceptual representation of speaker gender may therefore be gradient rather than categorical and based on fine phonetic detail that is stored in long-term memory. Hence, binary response options in many studies may prevent listeners from demonstrating their true representation of speaker gender and how different speakers are organized within that perceptual space.

The assumption that individuals conceptualize and represent speaker gender along a single bipolar dimension (e.g., "very male-like" or "very masculine" to "very female-like" or "very feminine") has also been questioned.[26,27] For example, Bem[26] found that individuals' self-concepts of masculinity and femininity were empirically independent of one another, suggesting that representation of gender may be multidimensional and that a single rating scale such as "very masculine" to "very feminine" may fail to adequately capture this complexity. In acknowledging this potentially more nuanced representation of gender, some investigations have sought to more finely examine listeners' gender attribution by providing additional response options in conjunction with or beyond discrete gender labels (e.g., confidence in categorical judgements or a separate rating scale for masculinity/femininity). Although these approaches acknowledge that gender attribution may lie along a continuum, they may conflate gender identity, masculinity, and femininity in the available response options they provide to listeners. For instance, some paradigms have measured masculinity and femininity by using visual analog or equal appearing interval scales as a proxy of gender identity.[28–31] Masculinity, femininity, and gender identity, however, are related yet distinct constructs. Numerous studies have shown listeners to have differing response patterns to gender identity and masculinity/femininity and, when measured, high intra- and inter-rater reliability.[14,32–34] Hope and Lilley[27] sought to separate masculinity, femininity, and gender identity by examining cisgender and transgender listeners' gender attribution of synthetic voice samples using three separate continuous scales of "masculinity," "femininity," and "other" and five categorical options of "man," "woman," "non-binary," "agender," and "genderfluid." Many listeners, particularly transgender, utilized the full range of the three rating scales and categorized voices outside of the traditional binary of "man" or "woman." Their findings suggest that listeners organize speaker gender within a multidimensional space.

However, their approach still assumed that listeners were able to segment attributed gender into discrete units along a visual analog scale.

Listeners' evaluations of speaker gender may also be influenced by the specific instructions or terminology provided by researchers. Studies that manipulate how listeners are instructed to evaluate speakers suggest that listeners' behavioral responses may be substantially altered by directing their attention to specific terms.[35–37] Houle *et al*.[36] found that varying anchor terms (e.g., "very female/very male" vs "feminine female/masculine male") along a visual analog scale systematically altered listeners' ratings. The authors concluded that the variability across scales may have been due to activation of varying degrees of stereotypes associated with gender that were triggered by the specific labels. Thus, the terms that investigators select for listener instructions or response options may lead to significant variability in how listeners interpret the task and respond. Although Hope and Lilley[27] provided their listeners with multiple experimenter-defined categorical response options when evaluating speakers, their paradigm assumed that the five category labels provided as response options accurately reflected listeners' perceptual organization and categorization of speakers.

In recognizing the numerous descriptive terms applicable to gender, a recent study allowed listeners to use their own terminology when attributing gender to speakers.[38] They found that participants provided an array of category labels, including "gender neutral," "genderqueer," "andro," "maleish," and "cismasc," emphasizing the broad spectrum of terms and their potentially varied semantic connotations that listeners may reference when attributing speaker gender. Thus, as Houle *et al*. suggest, different stereotypes may be activated depending on the terms that experimenters use, and different listeners may interpret the same terms or instructions differently. A recent meta-analysis indeed demonstrates the array of terms that experimenters have provided to listeners in measuring gender attribution.[39] The varying paradigms used to measure attribution of gender may have, thus, inadvertently introduced bias by experimenter-defined terminology and instructions. Such bias may have prevented an accurate assessment of listeners' gender attribution. Approaches that can more faithfully probe gender attribution with minimal bias are necessary to obtain a more complete picture of the cognitive processes that govern gender attribution and, more broadly, speech perception. Therefore, the purpose of this study was to examine how speaker gender was attributed when listeners were not restricted to experimenter-defined response options and audio samples solely from cisgender men and women. Further, we sought to determine how auditory-based gender attribution related to four factors:

- acoustic characteristics of the signal (e.g., f_o, formant frequencies, and vowel space area),
- auditory-perceptual ratings of speakers (e.g., definitely male/definitely female, very masculine/very feminine, and unforced/effortful),
- speakers' self-reported gender,
- task (e.g., stimulus set and instructions provided).

B. Auditory free classification

In a free classification paradigm, participants sort stimulus items into two or more groups based on either general similarity or more explicit attributes defined by the experimenter.[40] In contrast to paired comparisons, in which listeners are presented with all possible pairings of stimuli and tasked with rating similarity for each, free classification presents listeners with the entire set of stimulus items visually. Perceived proximities of items are determined by counting how often items are sorted into the same group across listeners. Items that are grouped together more often are considered to have a closer proximity to one another and, therefore, a higher degree of perceptual similarity. Listeners' classifications can be submitted to cluster analysis and/or dimension reduction to identify the degree of dissimilarity among all items collectively and the most salient perceptual dimensions across the stimulus set.[40] Auditory-perceptual and acoustic-phonetic measurements can be taken from speech stimuli and used to interpret the cluster and dimension reduction solutions.

Free classification provides several benefits over traditional measures of gender attribution. First, free classification can avoid experimenter-defined *a priori* category labels and thus limits the possibility that listeners may interpret the same gender related terms in different ways. Second, free classification can be completed much faster than a traditional paired comparisons paradigm,[40] which may reduce listener fatigue. Third, free classification is flexible in that the researcher can define how many groups the participant should organize stimuli into or allow participants to choose the number of groups for themselves. Participants can also be instructed to group items on more specific criteria rather than general similarity, such as the speakers' native language background[37] or perceptual breathiness.[35]

Given the flexibility that free classification provides, the stimulus items chosen for comparison should be based on the goals of the research. For example, using the same utterance produced by all speakers provides similar phonetic content and permits comparisons of fine-grained acoustic and auditory-perceptual measures across speakers. However, using novel sentences for each speaker avoids having listeners focus on a small set of fixed linguistic features when making their classifications and requires them to make similarity judgements using abstract representations of speakers and social groups.[41] Thus, free classification tasks using identical versus unique sentences across speakers could reveal nuance in listeners' representations based on the stimulus set. In the present study, we sought to leverage the free classification paradigm to clarify listeners' perceptual organization of cisgender and transgender speakers in conditions that manipulated both stimulus set (speakers producing the same or different sentences) and instructions provided to listeners (group by general similarity or by gender identity). These conditions would permit a comparison of how stimuli and instructions influenced listeners' classification strategies.

C. Multidimensional scaling

Multidimensional scaling (MDS) is a non-linear dimension reduction technique used to quantify similarity judgements among a set of items.[42] Input for MDS analysis are estimates of similarity. For instance, a distance matrix can be computed from listeners' groupings in a free classification task that can be submitted to MDS analyses. MDS places each stimulus item into an N-dimensional space (where N is defined by the investigator) while preserving the between-item distances and visualizes the resulting points in a scatterplot. The number of dimensions that best represent the data are chosen based on the amount of variance accounted for by each dimension and their interpretability. Each dimension identified in the final MDS solution is subjectively interpreted by examining the organization of speakers along that dimension and by calculating correlations between stimuli coordinates in the MDS solution and pre-defined acoustic and auditory-perceptual measures of the stimulus set.[43] Although MDS analyses identify the most meaningful features of a data set, the goal is to reduce stimuli complexity. Consequently, some information is lost during the process. Interpretation of dimensions identified in the MDS solution may also be challenging, particularly for high dimensional solutions. Nevertheless, MDS analyses allow complex data sets (such as perception of speech) to be reduced to the primary dimensions along which stimulus items are perceived to differ and provide a quantitative measurement and visual representation of their relationship.[42]

Auditory free classification of speech stimuli paired with MDS analyses has been employed to investigate listeners' perceptual representations of a range of speech stimuli, including regional dialect,[40,44] second language accent,[37] and neurogenic speech disorders.[43] Across these differing speaker populations, attributed gender frequently emerges as a primary organizational strategy.[37,40,44,45] For example, dimensional solutions from MDS analyses are consistently interpreted by researchers as a gender dichotomy both for classifications based on general similarity and for classifications by more specific criteria (e.g., speaker place of origin[44] or regional dialect[40]) Although these studies provide important insight into the salience of attributed speaker gender during speech processing, they typically assume that all speakers represent cisgender men and women, and therefore do not account for speakers of diverse gender identities (e.g., transgender men and women, non-binary, and agender individuals). Understanding how such speakers are perceptually represented would clarify how listeners cognitively organize and resolve the potentially novel acoustic-phonetic variability that is represented by gender diversity. Therefore, to refine our understanding of the perceptual representation of speaker gender, this study evaluated listeners' organizational schema for speakers who represented a variety of gender identities using an auditory free classification paradigm. Specific research questions were as follows:

(1) When presented with speakers representing a range of gender identities in a free classification task, what acoustic and perceptual dimensions are most salient to listeners?
(2) To what extent do speakers' gender identities impact perceptual similarity?
(3) Does manipulating stimulus set (same vs. different sentences across speakers) and instructions (general similarity vs. attributed gender) influence listeners' classification behavior?

II. METHODS

A. Stimuli

Speakers were 30 transgender and 30 cisgender adults ages 18–67 (M = 25.5, SD = 10.8). Because gender identities within the transgender community can be quite diverse,[46] limiting speakers to transgender men and transgender women would have been too restrictive to account for the great variability in identity that exists within the transgender population. Speaker recruitment therefore included speakers who identified as transgender, broadly, and allowed participants to indicate more nuanced identities such as *man*, *woman*, *non-binary* (e.g., *genderfluid*, *genderqueer*), *agender*, and *another identity not listed* with a free text box to provide more specific terminology.[47] Transgender speakers were recruited by word of mouth and through advertisements placed on the Indiana University campus and posted to the Indiana University classified ads, local LGBTQIA+ support group newsletters, and social media. Transgender speakers included 1 agender individual, 15 non-binary individuals, 7 transgender men, and 7 transgender women. All speakers were monolingual speakers of American English with no presence (as judged by the first author, a speech-language pathologist of over 15 years) or reported history of voice, speech, or other communicative disorder. Eleven of the transgender speakers who identified as either men or non-binary reported taking exogenous testosterone therapy as part of their gender affirming care. Time on testosterone for these speakers at the time of recording ranged from 2.5 months to 4 years, and dosages, when reported, varied. Although some speakers anecdotally reported previous gender-affirming communication training, this information was not requested.

Transgender speakers were age- and dialect-matched with cisgender speakers, who included 15 cisgender men and 15 cisgender women. Cisgender speakers were primarily taken from the ALLSSTAR corpus.[48] Because the ALLSSTAR corpus was limited in the number of cisgender speakers from which to draw, 7 additional cisgender speakers (2 women and 5 men) were recruited for this study by word of mouth from the surrounding Bloomington, IN community and personal networks. Additional demographic information for speakers is presented in Table I.

TABLE I. Demographic information for speaker participants.

Speaker group	Race/ethnicity	N	Sexual orientation (Ref. 49)	N	Dialect region	N
Transgender	Asian or Asian American	1	Asexual	2	Florida	1
	Black or African American	1	Asexual and panromantic	1	Mid-Atlantic	1
	Chose not to answer	1	Bisexual	9	Midland	18
	White	25	Demi-pansexual	1	New England	1
	White and Hispanic, Latinx, or Spanish Origin	1	Lesbian, gay, or homosexual	6	North	4
	White and Hispanic, Latinx, or Spanish origin and biracial or multiethnic	1	Pansexual	6	South	4
			Queer	4	West	1
			Straight or heterosexual	1		
Cisgender	Black or African American and White and biracial or multiethnic	1	Bisexual	1	Mid-Atlantic	7
	Chose not to answer	1	Chose not to answer	1	Midland	15
	Hispanic and Black or African American	1				
	Hispanic and more than one race	2	Lesbian, gay, or homosexual	2	Midland/North	1
	Not Hispanic and more than one race	1	Pansexual	1	North	2
	Non-Hispanic Asian	5	Straight or heterosexual	2	South	3
	Non-Hispanic other	1			West	2
	White	18				

All transgender speaker recordings and 7 cisgender speaker recordings for this study were conducted in-person in a sound-treated booth in the Speech Perception Lab in the Department of Speech, Language and Hearing Sciences at Indiana University after receiving approval from the Institutional Review Board. Speakers completed a demographic and language background questionnaire before completing the speech recordings. Speakers were then seated comfortably in a sound-treated booth and equipped with a head-mounted Shure Dynamic WH20XLR microphone that was positioned at approximately 5 cm mouth-to-microphone distance. Speakers participating at Indiana University were recorded using a Marantz PDM670 digital audio recorder connected to a Mac Mini at a 44.1 kHz sampling rate and 16-bit resolution. One dialect-matched cisgender speaker was recorded in a sound-treated booth in the Department of Communicative Disorders at the University of Alabama using a Steinberg audio interface and Electro Voice RE-20 dynamic cardioid mic at a 44.1 kHz sampling rate and 16-bit resolution. The remaining 22 cisgender speaker recordings were taken from the ALLSSTAR corpus and were recorded in-person at Northwestern University in a sound-treated booth using a Shure SM81 Condenser microphone connected to an Intel Core 2 Duo iMac at a 22.05 kHz sampling rate and 16-bit resolution.

Speakers recorded a variety of speech materials as part of a larger corpus development project, the details of which can be found in Merritt.[50] Stimuli for this study were meaningful, phonetically balanced, English read sentences from the Hearing In Noise Test (HINT).[51] Speakers were instructed to use their habitual speaking rate and most comfortable vocal presentation for recordings. All stimuli were intensity scaled to 70 dB. Breaks were provided as needed during the recording procedures. Speakers recorded specifically for this study were paid $10 per hour for their participation and the recording protocol lasted between 1 to 1.5 h.

B. Listeners

Listeners were 42 cisgender, monolingual, American English speakers (15 men and 27 women), ages 18–40 (mean = 23.7, SD = 5.5), who were recruited from the Indiana University campus and surrounding Bloomington community. Listeners were restricted to cisgender identity to avoid potential biasing effects of expertise with gender and voice/speech that individuals who identify as transgender may possess. Specifically, individuals who identify as transgender may be especially attentive to gender cues in speech because of their extensive lived experiences as a gender minority. Listeners older than 40 were ineligible to participate to limit potential bias from generational effects that may influence listeners' perceptual schemas. Additional listener demographic data is provided in Table II. Listeners had completed no more than two courses in communication sciences and disorders and no coursework in speech science or voice disorders. All listeners passed the hearing screening at 250, 500, 1000, 2000, 4000, and 8000 Hz presented at 20–25 dB. Listeners completed a demographic and social network questionnaire at the end of the experimental protocol. Participants completed all experimental tasks in a single session that lasted approximately 2.5 h. All were paid $25 for their participation. Along with the 42 listeners described, three additional listeners completed the experimental protocol, but their data were excluded from the analysis. Two participants' data were not saved following the auditory-perceptual ratings and the third identified as transgender.

C. Procedure

1. Free classification tasks

Listeners first completed three free classification tasks in which they grouped cisgender and transgender speakers by either general similarity or gender identity. Listeners were seated in front of a ViewSonic 20 in. VX2033wm LCD monitor in a sound booth that accommodated up to

TABLE II. Demographic information for listener participants.

Race/ethnicity	N	Sexual orientation	N	Dialect region	N
Asian or Asian American	7	Asexual	1	Florida	1
Asian or Asian American and White	2	Bisexual	6	Mid-Atlantic	1
Black or African American	3	Don't know	1	Midland	27
White	30	Lesbian, gay, or homosexual	8	New England	1
		Straight or heterosexual	26	North	6
				South	4
				West	2

two listeners at a time and wore Sennheiser HD 280 pro headphones. Following the methods by Clopper,[40] listeners were presented with a 12×12 grid on the right side of the screen and 60 square icons on the left side of the screen that were arranged in neat columns in alphabetical order using Microsoft PowerPoint. Each speaker was randomly assigned two-letter initials that were printed on each icon and, when clicked, played an utterance produced by that speaker. Listeners were instructed to drag each icon onto the grid and group icons so that speakers who sounded similar (tasks one and two) or who had the same gender identity (task three) were grouped together. Listeners were instructed to pay no attention to the meaning of the sentences when making groups. For each free classification task, listeners were told they could create as many or as few groups as they wished and there could be as many speakers in each group as they wanted. A 30-min time limit was imposed for each of the three tasks. They could click and hear each speaker's utterance as often as they liked while making groups. Listeners were not told they would be hearing cisgender and transgender speakers. They were provided scratch paper and a pencil with which to take notes about speakers while making groups if they wished but were not required to do so. Listeners were permitted breaks following each free classification task as described below.

(1) *General similarity–Same sentence*. The goal of the general similarity-same sentence free classification task was to examine listeners' representations of speaker gender when instructed to group speakers producing the same utterance by general similarity. In this condition, each speaker produced the utterance "They had two empty bottles."
(2) *General similarity–Different sentence*. The goal of the general similarity-different sentence free classification task was to examine listeners' representations of speaker gender when instructed to group speakers producing different utterances by general similarity. Although listeners were instructed to pay no attention to the meaning of sentences, this condition allowed examination of a potential effect of stimulus set, whereby utterances containing varying phonemes and lexical items may contribute to variability in perceptual groupings. In this condition, each of the 60 speakers produced a unique sentence. The order of the two general similarity classification tasks (same vs different utterance) was counterbalanced across listeners.
(3) *Gender identity–Same sentence*. The goal of the third and final free classification task was to examine listeners' representations of speaker gender when instructed to group speakers producing the same utterance by gender identity. This condition allowed examination of a potential effect of instruction. Here, explicitly focusing listeners' attention to gender could reveal differences in organizational structure as compared to experimenter instructions to group by general similarity. Each speaker in this condition produced the same utterance, "They had two empty bottles."

2. Auditory perceptual ratings

Following the free classification tasks, listeners completed auditory-perceptual ratings of the sentence "They had two empty bottles" produced by each speaker. The order of speakers was randomized across listeners. Ten rating scales were used to aid in the interpretation of dimensions revealed by the MDS analyses. Auditory-perceptual ratings collected from listeners were adapted from perceptual rating scales developed by Gelfer[52] to measure perceptual attributes of voice, using 9-point equal appearing interval scales. The perceptual variables represented by the rating scales were as follows: "breathy voice" to "full voice," "clear" to "hoarse," "definitely male" to "definitely female," "insufficiently nasal" to "excessively nasal," "I like this voice" to "I do not like this voice," "monotonous" to "animated," "slow rate" to "rapid rate," "unforced" to "effortful," "very masculine" to "very feminine," and "young" to "old." Listeners were provided a printed list of the rating scales with simple explanations of the scale labels that they were able to reference while completing the task.

Before beginning the experimental task, listeners first completed practice ratings of three speakers, one from each group of cisgender women, cisgender men, and transgender, reading a different HINT sentence from that used in the experiment. Listeners were permitted to take breaks as needed throughout the rating scale task. To address potential confounds of scale order and anchor placement, the order of rating scale presentation and anchor labels were counterbalanced across listeners; half of the listeners rated a set order of scales and direction of anchor labels (e.g., masculinity–femininity), and the other half rated a different set order of scales and direction of anchor labels (e.g., femininity–masculinity). Listener responses were mutated in data analysis so that all responses were along the same direction for each rating scale. Stimuli randomization, presentation, and listener response recording were controlled through PsychoPy 2021.2.3.[53] After finishing the auditory-perceptual ratings, listeners completed a demographic and social network questionnaire. The social network questionnaire was adapted from Dhand *et al*.[54] to determine if participants' immediate social networks (top 10 closest relationships) were inclusive of transgender individuals, and whether the existence of such relationships may influence their responses. Because no listeners' immediate networks were inclusive of transgender individuals, further analyses of the composition of these networks and their relation to listeners' responses were not conducted.

3. Acoustic analyses

The goal of the acoustic analyses was to quantify speech parameters to use in interpretation of multidimensional scaling solutions for listeners' groupings. Acoustic-phonetic parameters were selected that have shown relations with auditory-perceptual evaluations of speaker gender. The specific acoustic-phonetic features selected were mean f_o across the duration of each speaker's utterance,[34,55,56] cepstral peak prominence-smoothed (CPPs),[57,58] the first three formant frequencies (F1, F2, and F3) of the four corner vowels /i, u, ae, ɑ/,[2,34,55,56] and vowel space area (VSA).[59] The utterance, "They had two empty bottles," was selected, because it contained each of the four corner vowels /i, u, ae, ɑ/. Acoustic analyses were conducted as described in Merritt[10] using PRAAT.[60] Additionally, unitless prototypicality indices for each speaker were computed following the methods of Merritt and Bent.[6] These indices equated speakers who had f_o and first formant frequency at either extreme end of the range of values across all speakers versus speakers with values more central. Separate prototypicality indices were created for f_o (PI:f_o) and the first three formant frequencies (PI:SF1, PI:SF2, and PI:SF3, respectively) for each speaker. Higher values for each index indicated that a speaker's frequencies were at the extreme ends of the range across all speakers (i.e., relatively low or high). Lower values indicated that a speaker's frequencies were closer to the median frequency and therefore more ambiguous. A summary of the ten auditory-perceptual and ten acoustic measures taken to aid in interpretation of MDS solutions is provided in Table III.

4. Statistical analyses

Data processing was conducted following the methods of Clopper.[40] For each free classification task, a PowerPoint macro[61] was used to extract summary data from each participant's completed classifications, including the number of groups created and the number of items within each group. Additionally, the macro produced for each listener a speaker-by-speaker similarity matrix, in which pairs of speakers placed within the same group were assigned a value of 1 and pairs of speakers not placed within the same group were assigned a value of 0. Thus, a 60×60 similarity matrix for each free classification task completed by each listener was created. Summary data were also stratified by listeners' reported sexual orientation on the demographic questionnaire. Those who reported their sexual orientation as straight or heterosexual were coded as straight (N = 26), and listeners who reported their sexual orientation as anything other than straight or heterosexual were coded as queer (N = 16).

TABLE III. Auditory-perceptual and acoustic measures taken to aid in interpretation of MDS solutions.

Measure	Unit of measurement
Auditory-perceptual measures	
Breathy voice/Full voice	9-point equal appearing interval scale
Clear/Hoarse	
Definitely male/Definitely female	
Insufficiently nasal/Excessively nasal	
I like this voice/I do not like this voice	
Monotonous/Animated	
Slow rate/Rapid rate	
Unforced/Effortful	
Very masculine/Very feminine	
Young/Old	
Acoustic Measures	
CPPs	dB
Mean speaking f_o	Hz
F1	Hz
F2	Hz
F3	Hz
PI:f_o	unitless
PI:SF1	unitless
PI:SF2	unitless
PI:SF3	unitless
VSA	Hz^2

Two sets of statistical analyses were conducted on the extracted data. The first analysis aimed to identify listeners' overall classification strategy, including the mean, minimum, and maximum number of groups produced. Free classification results were first evaluated for potential outliers that may spuriously influence results. The mean and standard deviation of the number of groups created across listeners for each free classification task were computed. Due to the exploratory nature of the study, it was decided that free classification responses with number of groupings greater than 3 standard deviations from the mean would be discarded. This decision resulted in one free classification from the general similarity-different sentence condition and one free classification from the gender identity-same sentence condition being removed from analyses. Upon inspection, it was noted that these listeners left many speakers ungrouped and may have had difficulty completing or misunderstood the directions for the task. Thus, there remained for analyses 42 listener free classification responses for the general similarity-same sentence condition and 41 for the general similarity-different sentence and gender identity-same sentence conditions.

The second set of analyses used MDS to identify the most salient perceptual dimensions of similarity among the speakers. MDS analyses were conducted to determine the perceptual similarity of speakers across all listeners combined and between straight and queer listeners. To examine the perceptual similarity of speakers across all listeners, the speaker-by-speaker similarity matrices were summed to create pooled similarity matrices, one for each free classification task. These matrices contained gradient values of similarity between 0, indicating speakers were never grouped together, and 42 (or 41 for the general similarity-different sentence and gender identity-same sentence conditions), indicating speakers were always grouped together.[40] Each pooled matrix was separately submitted to MDS using the Euclidean distance algorithm with ordinal similarity data in SPSS 25.0.[62] Five dimensions were first entered into the model for each similarity matrix as a general heuristic to evaluate the spread of solution stress from a low to high dimensional space.[42] Stress values produced by an MDS solution are visually represented as Scree plots. In these plots, lower values of stress for a given number of dimensions indicate greater variance accounted for and, hence, a better model fit. A common approach is to select the number of dimensions at which an "elbow" forms in the scree plot, which indicates that additional dimensions no longer substantially improve the model fit.[42] The number of dimensions that best explained the similarity judgments in the present study was determined based on analysis of stress values from scree plots and considerations of solution interpretability.

Individual differences scaling (INDSCAL), a subset of MDS, was used to examine potential differences in the perceptual similarity of speakers between straight and queer listeners. INDSCAL was conducted with the weighted Euclidean distance algorithm and ordinal similarity data in SPSS. INDSCAL determines the similarity space for the entire set of listeners and provides dimension weights for each listener that can then be averaged for each group of listeners. Statistical comparisons of these dimension weights were made between straight and queer listeners to determine if each group attended more to one dimension than others. Additional discussion of these statistical techniques can be found in Clopper[40] and Carroll and Chang.[63]

Interpretations of MDS solutions for each free classification task were made by first examining the organization of speakers along the dimensions that best fit the perceptual similarity matrices. Next, we calculated correlations between stimuli coordinates in each of the three MDS solutions and the acoustic and auditory-perceptual measures described in Table III using SPSS. Those variables that showed the strongest correlations with the stimuli coordinates were used to interpret and label each dimension identified in the MDS solutions.

III. RESULTS

A. Free classification groupings

Listeners on average created 16.1 (range = 2–37, SD = 7.5) and 15.6 (range = 2–33, SD = 6.6) groups in the general similarity-same sentence and general similarity-different sentence conditions, respectively. For the gender identity-same sentence condition, listeners on average created 2.9 (range = 2–7, SD = 1.3) groups. A one-way analysis of variance (ANOVA) revealed significant differences in average number of groups [$F(2,121) = 66.56$, $p < 0.001$] and average number of speakers per group [$F(2,121) = 150.98$, $p < 0.001$] among conditions. Tukey's HSD Test for multiple comparisons found that number of groups and number of speakers per group were not significantly different between the general similarity-same sentence and general similarity-different sentence conditions. However, the number of groups in the gender identity-same sentence condition was significantly lower than both the general similarity-same sentence ($p < 0.001$, 95% C.I. = [–16.2, –10.13]) and general similarity-different sentence ($p < 0.001$, 95% C.I. = [–15.49, –9.39]) conditions. Correspondingly, the number of speakers per group in the gender identity-same sentence condition was significantly higher than the general similarity-same sentence ($p < 0.001$, 95% C.I. = [15.56, 21.33]) and general similarity-different sentence ($p < 0.001$, 95% C.I. = [15.38, 21.18]) conditions.

We next examined whether listeners' sexual orientation may have impacted classification strategy by comparing the summary data for straight versus queer listeners. Excluding the previously described free classifications with number of groupings greater than 3 standard deviations above the mean, a two-way multivariate analysis of variance (MANOVA) was conducted to evaluate differences across conditions in the number of groups or the number of speakers per group based on listeners' sexual orientation. Results, although trending toward a greater number of groups created by queer listeners, revealed no significant differences across conditions in the number of groups or the number of speakers per group between straight and queer listeners, $F(4, 234) = 2.3$, $p = 0.06$; Wilks' $\Lambda = 0.93$.

B. Multidimensional scaling solutions

Scree plots for the general similarity-same sentence, general similarity-different sentence, and gender identity-same sentence conditions are presented in Fig. 1.

For the general similarity-same sentence and general similarity-different sentence conditions, a clear elbow formed at 2 dimensions, and so 2-dimensional solutions were chosen as the best fit for the data. The MDS analyses were then re-run using the same analysis parameters except for specifying only 2 dimensions. These 2-dimensional solutions resulted in an S-Stress value of 0.06 and Dispersion Accounted For of 0.97 for both the general similarity-same sentence and general similarity-different sentence conditions. For the gender identity-same sentence condition, the addition of dimensions beyond 1 increased stress and, hence, variance unaccounted for. Thus, a 1-dimensional solution was selected as the best fit for the data. This solution produced an S-Stress value of 0.001 and Dispersion

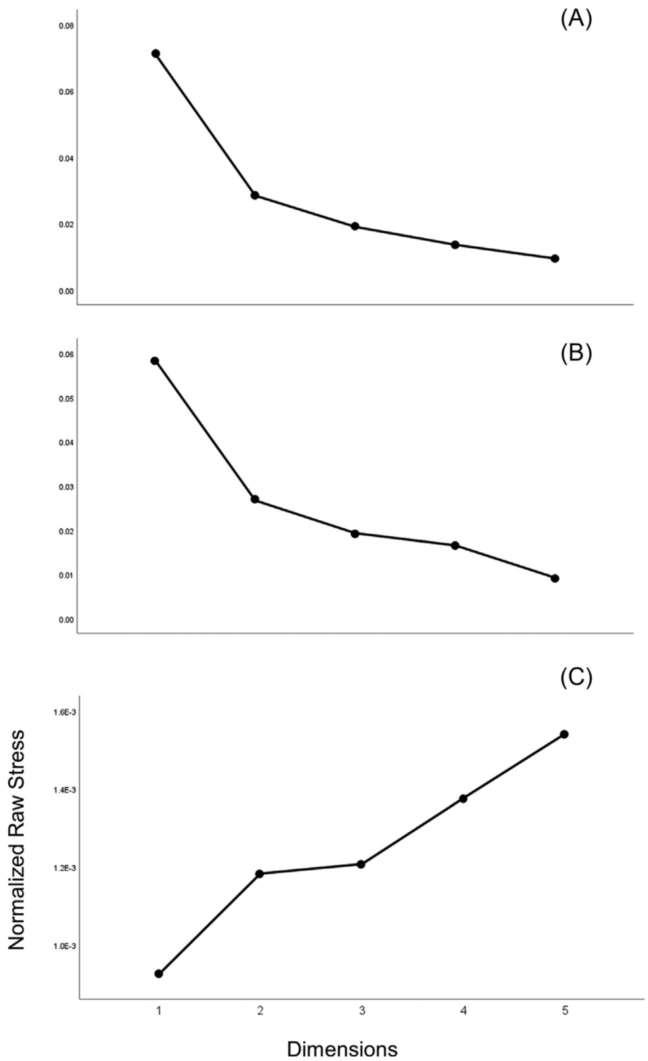

FIG. 1. Scree plots of multidimensional scaling solutions for the (A) general similarity-same sentence, (B) general similarity-different sentence, and (C) gender identity-same sentence conditions.

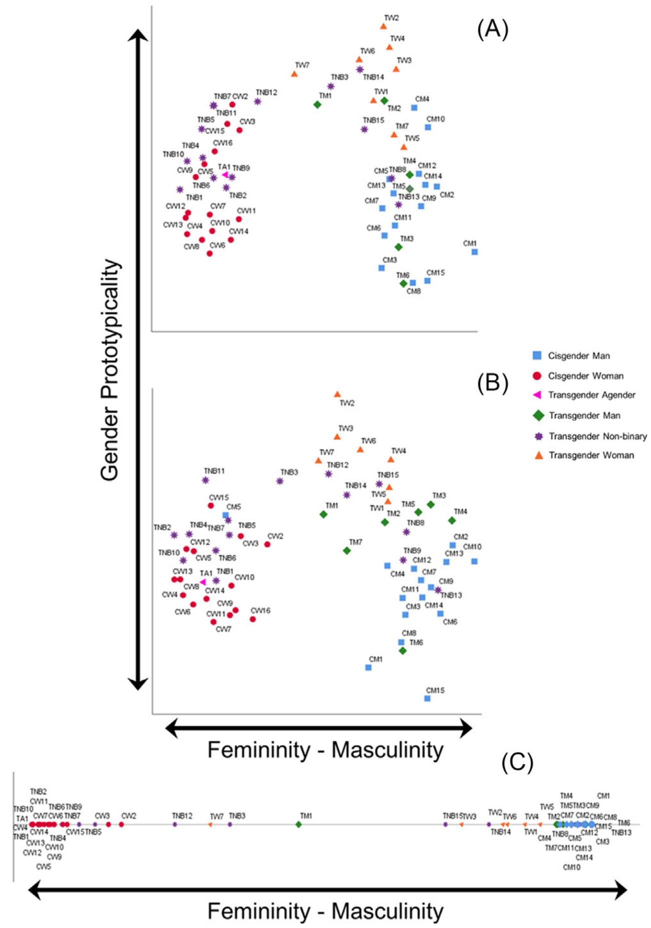

FIG. 2. (Color online) Multidimensional scaling solutions for the (A) general similarity-same sentence, (B) general similarity-different sentence, and (C) gender identity-same sentence conditions. CM: Cisgender Man, CW: Cisgender Woman, TA: Transgender Agender, TM: Transgender Man, TNB: Transgender Non-binary, TW: Transgender Woman. A color image is available online.

Accounted For of 0.99. The MDS solutions for each free classification condition are shown in Fig. 2.

C. Interpretation of identified dimensions

Examination of Fig. 2 demonstrates a gradient rather than discrete and binary organization of speakers. For each MDS solution presented in Fig. 2, Pearson correlation coefficients were used to assess relations between each speaker's coordinates and the acoustic-phonetic measures of f_o, CPPs, formant frequencies, and prototypicality indices. Spearman rank correlation coefficients were used to assess relations between MDS coordinates and the auditory-perceptual measures of Breathy voice/Full voice, Clear/Hoarse, Definitely male/Definitely female, Insufficiently nasal/Excessively nasal, I like this voice/I do not like this voice, Monotonous/Animated, Slow rate/Rapid rate, Unforced/Effortful, Very masculine/Very feminine, and Young/Old. Second and third formant frequencies demonstrated high multicollinearity with first formant frequency as determined by demonstrating variance inflation factors ≥ 5.[64] They were, therefore, removed from analyses and only first formant frequency and its prototypicality index (PI:SF1) was used in the interpretation of MDS solutions. Similarly, auditory-perceptual ratings of Definitely male/Definitely female demonstrated high multicollinearity (≥ 5) with auditory-perceptual ratings of Very masculine/Very feminine. Definitely male/Definitely female was, therefore, also removed from analyses and only Very masculine/Very feminine was used in interpretation of MDS solutions. All other acoustic and auditory-perceptual measures demonstrated variance inflation factors of < 5. Significant correlations between dimension coordinates and acoustic-phonetic, auditory-perceptual, and prototypicality indices measures for each free classification condition are reported in Tables IV–VI. All correlations between these measures are provided in the supplementary material.[78]

Dimension 1 was similar for the general similarity-same sentence, general similarity-different sentence, and gender identity-same sentence conditions. Fundamental frequency and first formant frequency were both strongly correlated with Dimension 1 across conditions. Very

TABLE IV. Significant correlations between multidimensional scaling coordinates and acoustic-phonetic, auditory-perceptual, and prototypicality measures for the general similarity-same sentence condition. For measures that correlated with both dimensions, the *r* value for the dimension with the greater magnitude is bolded. f_o: fundamental frequency, PI:f_o: prototypicality index for f_o, PI:SF1: prototypicality index for first formant frequency, SF1: first formant frequency, VSA: vowel space area.

Dimension 1		Dimension 2	
Variable	*r*	Variable	*r*
Very masculine/ very feminine	−0.88[a]	PI:f_o	−0.56[a]
f_o	**−0.76**[a]	PI:SF1	−0.56[a]
SF1	−0.66[a]	Monotonous/animated	0.42[a]
Young/Old	**0.35**[a]	VSA	**0.36**[a]
VSA	−0.29[b]	Young/Old	0.3[b]
		I like this voice/ I do not like this voice	−0.3[b]
		f_o	0.27[b]

[a]Correlation is significant at the 0.01 level.
[b]Correlation is significant at the 0.05 level.

Masculine/Very Feminine, a perceptual feature strongly tied to these acoustic measures,[39] also strongly related to this dimension. Moreover, strong correlations were found for the MDS coordinates of dimension 1 between the general similarity-same sentence and general similarity-different sentence conditions ($r = 0.9$, $p < 0.001$), between the general similarity-same sentence and gender identity-same sentence conditions ($r = 0.98$, $p < 0.001$) and between the general similarity-different sentence and gender identity-same sentence conditions ($r = 0.9$, $p < 0.001$). Therefore, dimension 1 was interpreted as perceived masculinity/femininity for all conditions. This dimension appears to progress from high to low frequencies, with speakers tending to have f_o and first formant frequency within an intermediate range clustering toward the central portion of this dimension.

Dimension 2 in the general similarity-same sentence and general-similarity-different conditions was also similar. Dimension 2 coordinates of both conditions moderately correlated with f_o and first formant prototypicality indices.

TABLE V. Significant correlations between multidimensional scaling coordinates and acoustic-phonetic, auditory-perceptual, and prototypicality measures for the general similarity-different sentence condition. For measures that correlated with both dimensions, the *r* value for the dimension with the greater magnitude is bolded. f_o: fundamental frequency, PI:f_o: prototypicality index for f_o, PI:SF1: prototypicality index for first formant frequency, SF1: first formant frequency, VSA: vowel space area.

Dimension 1		Dimension 2	
Variable	*r*	Variable	*r*
Very Masculine/ Very Feminine	−0.84[a]	PI:SF1	−0.55[a]
f_o	**−0.73**[a]	PI:f_o	−0.37[a]
SF1	−0.59[a]	VSA	0.28[b]
VSA	**−0.3**[b]		

[a]Correlation is significant at the 0.01 level.
[b]Correlation is significant at the 0.05 level.

TABLE VI. Significant correlations between multidimensional scaling coordinates and acoustic-phonetic, auditory-perceptual, and prototypicality measures for the gender identity-same sentence condition. f_o: fundamental frequency, SF1: first formant frequency, VSA: vowel space area.

Dimension 1	
Variable	*r*
Very masculine/very feminine	−0.97[a]
f_o	−0.79[a]
SF1	−0.67[a]
Young/Old	0.38[a]
VSA	−0.3[b]

[a]Correlation is significant at the 0.01 level.
[b]Correlation is significant at the 0.05 level.

As with Dimension 1, speakers appear to be organized within dimension 2 based on f_o and first formant frequency. However, the manner with which f_o and first formant frequency was used as an organizational strategy differed between the two dimensions. Speakers with f_o and first formant values at the lower and upper extremes of the stimulus set (i.e., those who had larger prototypicality indices values) tended to separate along the *y* axis from speakers with f_o and first formant values at intermediate values (i.e., those who had smaller prototypicality indices values). Similar to dimension 1, a strong correlation was also found between the MDS coordinates of dimension 2 for the two conditions ($r = 0.84$, $p < 0.001$). Dimension 2 was, therefore, interpreted as gender prototypicality for both conditions. Notably, both dimensions 1 and 2 demonstrated a gradient organization of speakers that spanned the two dimensions.

D. INDSCAL analyses

The relative weightings of dimensions 1 and 2 for the general similarity-same sentence and general similarity-different sentence free classification conditions for all listeners combined were calculated using INDSCAL. Independent samples *t* tests were conducted to compare the means of the weightings for dimensions 1 and 2 in each of the two conditions. There were no significant differences between dimension 1 and dimension 2 weightings for both the general similarity-same sentence, $t(82) = 1.8$, $p = 0.08$, and general similarity-different sentence, $t(78) = 0.75$, $p = 0.46$, conditions.

We also compared MDS dimension weightings for listeners who were coded as straight versus listeners who were coded as queer using INDSCAL. There were no differences between straight and queer listeners in the number of dimensions identified or in the relative weightings of dimensions 1 and 2 for the general similarity-same sentence and general similarity-different sentence free classification tasks [$t(40) = 0.32$, $p = 0.75$, and $t(39) = 1.3$, $p = 0.2$, respectively].

IV. DISCUSSION

This study used auditory free classification to investigate listeners' organizational schema for speakers who represented a variety of gender identities. We found that, based on the distribution of speakers along identified perceptual dimensions and acoustic and auditory-perceptual correlates, listeners used gradient representations of masculinity/femininity and gender prototypicality as the primary organizing factors in classifying speakers. These representations emerged even without specifically referencing gender in the instructions provided to them. These findings align with prior free classification studies that have found a salient "gender" dimension when listeners were presented with stimuli from cisgender speakers representing a range of regional dialects or second language accents and tasked with grouping them by general similarity or dialect.[37,41,44,45] Previous studies have generally described the gender dimension in a binary fashion (e.g., women on one half of a MDS solution and men on the other half). Our data, however, indicate that these binary interpretations of gender may have been reductionist and overlooked listeners' finer discrimination that we present here. Nonetheless, the fact that listeners use gender-related features as a key organizational strategy demonstrates the perceptual salience of gender information in speech. In line with exemplar or hybrid models, listeners may condition retrieval of linguistic units on structured variation seen across differing gender expressions. Such conditioning would provide structure to the highly variable acoustic signal and lessen the cognitive demands of parsing speech.[65] Our findings demonstrate that, even with a diverse set of speakers who represent a variety of gender identities, cisgender listeners still map these voices to a perceptual schema of masculinity/femininity and gender prototypicality that may serve as a preliminary estimation of structured acoustic variation.[21]

Our findings that listeners use a multidimensional schema in evaluating speaker gender is contrary to many assumptions of a binary categorical gender representation.[2,4,7] MDS analyses revealed two-dimensional solutions, in which listeners demonstrated gradient organization of speakers for each dimension, as the best fit for free classifications when listeners were instructed to group speakers by general similarity. The first dimension was interpreted as masculinity/femininity, where listeners organized speakers from high to low f_o and first formant frequency. The second was interpreted as gender prototypicality, where listeners separated speakers with f_o and first formant frequency at upper and lower extreme values (prototypical) from more intermediate values (non-prototypical). Even when listeners were instructed to group speakers based on attributed gender, most listeners created more than two groups, with some listeners creating as many as seven separate groupings.

For both general similarity-same sentence and general similarity-different sentence conditions, MDS analyses demonstrated that masculinity/femininity and gender prototypicality were similarly weighted in listeners' organizational schema. This multidimensional representation suggests that during speech perception listeners engage in a more fine-grained analysis of speaker gender than has generally been assumed by prior research. This greater perceptual detail cannot be adequately captured by the binary categories that have historically been used in speech communication research. Rather, these results align with prior findings demonstrating finer discrimination of speakers within man and woman categories[3] and a non-binary representation of gender.[27] Research that aims to distinguish speech differences among gender identities and listeners' attribution of gender should, therefore, consider this multidimensional complexity in how speakers are classified and which response options are provided to listeners.

Listeners' classifications were influenced both by investigator-provided instructions and characteristics of the stimulus set. Explicitly directing listeners to attend to speaker gender in the gender identity-same sentence condition resulted in fewer perceptual distinctions among speakers, as evidenced by the collapse of listeners' classifications into a simplified, one-dimensional space in the MDS solution and many fewer perceptual groupings. Hence, the effect of instruction significantly altered listeners' organizational structure and demonstrated that measured representation of speaker gender may considerably vary as a function of experimental conditions. Allowing free classification by general similarity revealed a much more complex organizational structure and provided richer data on representation than was demonstrated in the gender identity condition. Of course, in the general similarity conditions, listeners may have also been creating groups based on other indexical characteristics, such as age, although these did not appear to be consistent enough to appear in the MDS solutions. Similar to our findings, Atagi and Bent[37] observed a gender dimension when listeners grouped second language speakers of English by general similarity, particularly when all talkers produced the same sentence. However, the salience of gender diminished in conditions in which listeners were instructed to group talkers by first language. Thus, gender-related dimensions appear to be highly salient to listeners across a range of stimulus sets, although their prominence can be attenuated when listeners are explicitly instructed to attend to other indexical features.

By using gender-related terms in the provided instructions, listeners may have been biased *a priori* by stereotypes based on the gender binary, which resulted in a more simplified schema. Explicit reference to "gender" in the gender identity condition may have activated these stereotypes and guided listeners' classifications.[66] The simplified schema obtained from the gender identity condition may, thus, represent a distorted characterization of gender representation that is based on these stereotypes rather than listeners' true cognitive organization of these speakers. Overall, these data demonstrate subtle nuance in perceptual representation of speaker gender that may have been obscured in prior research by paradigms that explicitly focused listeners' attention to gender-related terminology.

Although the general similarity-same sentence and general similarity-different sentence conditions were similar in terms of their underlying acoustic and auditory-perceptual correlates and interpretations, some notable differences between the two conditions were seen. Fewer auditory-perceptual features correlated with the MDS solution for the different sentence condition as compared to the same sentence condition. Auditory-perceptual variables of Young/Old, Monotonous/Animated, and I like this voice/I do not like this voice did not demonstrate the significant relations with the MDS solution for the general similarity-different sentence condition that they did with the general similarity-same sentence condition. These findings indicate that listeners' perceptual systems may have been taxed by evaluating multiple indexical dimensions in the presence of greater phonemic and lexical variability. The greater cognitive demands of parsing differing sentences among the speakers would, thus, result in listeners attending to fewer indexical features when organizing speakers along the two dimensions. The fact that the gender related features were the only correlates of the MDS solution for the different sentence condition suggests that they became more pronounced at the expense of other indexical features in listeners' classifications when cognitive demands increased. Although additional research is needed to confirm this hypothesis, these results point to linguistic and indexical dimensions as being processed in a mutually dependent manner and that variability along one dimension may affect processing of the other.[67]

These findings also provide insight into speech adaptation. A central question in the study of speech processing and representation is how listeners contend with the vast acoustic variability that exists across speakers. Numerous studies suggest listeners retain detailed acoustic information about speakers in long-term memory that allows them to track similarities across social groups, such as gender.[41,44,45,68] Listeners can then use these perceptual models to make inferences about novel speakers.[21] The fact that listeners in this study demonstrated a non-categorical representation of gender implies that they must retain information about speakers that allows them to scale gender in a multivalent manner. This nuanced representation suggests that listeners may develop more expansive or numerous models of speaker gender that underlie their perceptual organization. These models move beyond the gender binary organization often assumed in speech communication research and, importantly, suggest that listeners leverage structured acoustic variability that differentiates speakers of diverse gender identities to inform perceptual representation of speech.[6,10–13,34,55]

Although our data cannot explain *how* listeners develop and refine their perceptual schemas of speaker gender, a likely source of the nuanced representation we found here is listener experience. As described by the ideal adaptor framework,[20,21] listeners' experiences may allow them to continually update their models for speaker gender. Through social networks, media, political involvement, and other types of exposure, listeners' representations of speaker gender and the meaningful ways that speech may vary because of gender diversity evolve. This malleability of speech representation may allow expansion of current or development of new representations of gender that could facilitate adaptation to novel speakers and, thus, listeners' ability to perceptually adapt to changing sociocultural landscapes. Notably, none of our listeners indicated that their immediate social networks were inclusive of individuals who are transgender. This fact may, in part, explain why listeners' representation, although multidimensional, was still binary in some ways (e.g., speakers were categorized generally from feminine to masculine and from less prototypical to more prototypical). Cisgender listeners who have greater gender diversity within their social networks may develop an even more nuanced representation of speaker gender. Future studies should consider how listeners' exposure to individuals with diverse gender identities beyond their immediate social networks may impact their gender classification. For example, survey questions could be developed to probe listeners' interactions with individuals in their wider social circles as well as through media sources (e.g., TV, podcasts, social media). Similarly, listeners who are not cisgender may have greater exposure to individuals whose speech characteristics do not necessarily align with prototypical cisgender men and women. Consequently, their perceptual representation of speaker gender may have greater dimensionality as compared to cisgender listeners. Future studies are recommended to investigate these hypotheses.

While our findings help to clarify the perceptual representation of speaker gender, some limitations of this study should be considered for future work. Our stimulus pool was limited in the number of speakers of diverse gender identities who were represented. The majority of our transgender speaker participants identified as non-binary and only a single speaker identified as agender. Having a greater number of transgender speakers, particularly transgender men, transgender women and agender individuals, may reveal additional nuance in how listeners perceptually organize these diverse voices. Additional speakers would also allow for a more robust characterization of speech acoustic distinctions among varying gender identities.

Speakers' experiences with gender expression may have also influenced our findings. For example, we did not document speaker participants' history of gender affirming speech therapy that may have influenced listeners' perceptual judgements. Future work could seek to compare production and perceptual differences between transgender speakers with and without histories of formal communication training to assess the impact such training may have. For instance, transgender women who have undergone formal speech therapy tend to produce higher speaking f_o,[69–71] more f_o variation,[70,71] and higher formant frequencies[72,73] than those who have not.

Additionally, the fact that one of the identified dimensions in the MDS solutions represented masculinity/femininity is puzzling given that constructs of masculinity and femininity have been argued to be separate dimensions.[26] A potential explanation for this discrepancy may be in our choice of a single rating scale and anchor terms used to collect auditory-perceptual measurements of masculinity/femininity. Houle *et al.*[36] found that varying anchor terms

https://doi.org/10.1121/10.0024521
JASA

along a visual analog scale systematically altered listeners responses. Thus, our decision to utilize a single scale for masculinity/femininity with two end points and the terminology we used as anchors along the scale may have limited listeners' ability to evaluate each of these dimensions independently during the auditory-perceptual rating task. Future work may improve on these methods by eliciting separate ratings of masculinity and femininity from listeners.

Last, although the racial and ethnic representation of participants in our study was representative of the U.S. Midwest where the research was conducted,[74] participants were all monolingual American English individuals with largely non-Hispanic white identities. Because acoustic features that distinguish speakers of varying gender identities may vary across languages and cultures,[75–77] the lack of diversity in our participant pool may prevent a true representation of the experiences of many speakers and listeners. Thus, studies that recruit speakers and listeners from diverse linguistic communities and backgrounds could provide insight into cultural linguistic factors that may influence perceptual representation of speaker gender.

V. CONCLUSION

Cisgender listeners' perceptual representation of speaker gender was found to operate within a multidimensional space and in a gradient manner based on organizing factors of masculinity/femininity and gender prototypicality. Listeners appear to engage in fine-grained analyses of speaker gender that cannot be adequately captured by a gender dichotomy. Assumptions of a gender binary in the study of speech communication may, thus, require a critical reexamination to accommodate this multidimensional and gradient representation. Furthermore, the instructions provided from experimenters and characteristics of the stimulus set influenced listeners' categorizations and should be considered in future studies examining gender attribution.

ACKNOWLEDGMENTS

We are grateful to the Department of Speech, Language and Hearing Sciences at Indiana University for providing funding in support of this research. We are also thankful to a member of the transgender community for their review of this manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts of interest to disclose.

Ethics Approval

Ethics approval for this study was obtained from the Indiana University Institutional Review Board. Informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1]We acknowledge that in speech communication literature concepts of sex (e.g., "male" and "female") and gender (e.g., "man/boy," "woman/girl") are used inconsistently and interchangeably. In the context of speech communication, we assume that authors use "male" and "female" to reference the listener perceived construct of gender attribution. Within this paper, we use the terms "male" and "female" to reference the binary categories often used in the cited literature and for an auditory-perceptual scale described in the Methods section. We follow the approach of Azul and Hanock (Ref. 16) to use the term "gender" as the overarching term referring to both the physical and the sociocultural aspects of this construct.

[2]V. G. Skuk and S. R. Schweinberger, "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," J. Speech. Lang. Hear. Res. **57**(1), 285–296 (2014).

[3]J. W. Mullennix, K. A. Johnson, M. Topcu-Durgun, and L. M. Farnsworth, "The perceptual representation of voice gender," J. Acoust. Soc. Am. **98**(6), 3080–3095 (1995).

[4]J. M. Hillenbrand and M. J. Clark, "The role of f0 and formant frequencies in distinguishing the voices of men and women," Atten. Percept. Psychophys. **71**(5), 1150–1166 (2009).

[5]T. Murry and S. Singh, "Multidimensional analysis of male and female voices," J. Acoust. Soc. Am. **68**(5), 1294–1300 (1980).

[6]B. Merritt and T. Bent, "Perceptual evaluation of speech naturalness in speakers of varying gender identities," J. Speech. Lang. Hear. Res. **63**(7), 2054–2069 (2020).

[7]M. Owren, M. Berkowitz, and J. Bachorowski, "Listeners judge talker sex more efficiently from male than from female vowels," Percept. Psychophys. **69**, 930–941 (2007).

[8]E. Coleman, A. E. Radix, W. P. Bouman, G. R. Brown, A. L. C. de Vries, M. B. Deutsch, R. Ettner, L. Fraser, M. Goodman, J. Green, A. B. Hancock, T. W. Johnson, D. H. Karasic, G. A. Knudson, S. F. Leibowitz, H. F. L. Meyer-Bahlburg, S. J. Monstrey, J. Motmans, L. Nahata, T. O. Nieder, S. L. Reisner, C. Richards, L. S. Schechter, V. Tangpricha, A. C. Tishelman, M. A. A. Van Trotsenburg, S. Winter, K. Ducheny, N. J. Adams, T. M. Adrián, L. R. Allen, D. Azul, H. Bagga, K. Başar, D. S. Bathory, J. J. Belinky, D. R. Berg, J. U. Berli, R. O. Bluebond–Langner, M.-B. Bouman, M. L. Bowers, P. J. Brassard, J. Byrne, L. Capitán, C. J. Cargill, J. M. Carswell, S. C. Chang, G. Chelvakumar, T. Corneil, K. B. Dalke, G. De Cuypere, E. de Vries, M. Den Heijer, A. H. Devor, C. Dhejne, A. D'Marco, E. K. Edmiston, L. Edwards-Leeper, R. Ehrbar, D. Ehrensaft, J. Eisfeld, E. Elaut, L. Erickson-Schroth, J. L. Feldman, A. D. Fisher, M. M. Garcia, L. Gijs, S. E. Green, B. P. Hall, T. L. D. Hardy, M. S. Irwig, L. A. Jacobs, A. C. Janssen, K. Johnson, D. T. Klink, B. P. C. Kreukels, L. E. Kuper, E. J. Kvach, M. A. Malouf, R. Massey, T. Mazur, C. McLachlan, S. D. Morrison, S. W. Mosser, P. M. Neira, U. Nygren, J. M. Oates, J. Obedin-Maliver, G. Pagkalos, J. Patton, N. Phanuphak, K. Rachlin, T. Reed, G. N. Rider, J. Ristori, S. Robbins-Cherry, S. A. Roberts, K. A. Rodriguez-Wallberg, S. M. Rosenthal, K. Sabir, J. D. Safer, A. I. Scheim, L. J. Seal, T. J. Sehoole, K. Spencer, C. St. Amand, T. D. Steensma, J. F. Strang, G. B. Taylor, K. Tilleman, G. G. T'Sjoen, L. N. Vala, N. M. Van Mello, J. F. Veale, J. A. Vencill, B. Vincent, L. M. Wesp, M. A. West, and J. Arcelus, "Standards of care for the health of transgender and gender diverse people," Int. J. Transgend. Health **23**(S1), S1–S259 (2022).

[9]J. Herman, A. Flores, and K. O'Neill, "*How Many Adults and Youth Identify as Transgender in the United States?*," https://williamsinstitute.law.ucla.edu/publications/trans-adults-united-states/ (2022) (Last viewed October 2, 2023).

[10]B. Merritt, "Speech beyond the binary: Some acoustic-phonetic and auditory-perceptual characteristics of non-binary speakers," JASA Express Lett. **3**(3), 035206 (2023).

[11]N. Houle and S. V. Levi, "Acoustic differences between voiced and whispered speech in gender diverse speakers," J. Acoust. Soc. Am. **148**(6), 4002–4013 (2020).

[12]L. Brown and C. Pillot-Loiseau, "Bright voice quality and fundamental frequency variation in non-binary speakers," J. Voice (published online 2022).
[13]L. Zimman, "Variability in /s/ among transgender speakers: Evidence for a socially grounded account of gender and sibilants," Linguistics **55**(5), 993–1019 (2017).
[14]B. Merritt and T. Bent, "Revisiting the acoustics of speaker gender perception: A gender expansive perspective," J. Acoust. Soc. Am. **151**(1), 484–499 (2022).
[15]L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men," J. Lang. Sexuality **2**(1), 1–39 (2013).
[16]A. Agha, *Language and Social Relations* (Cambridge University Press, Cambridge, 2007).
[17]D. Azul and A. B. Hancock, "Who or what has the capacity to influence voice production? Development of a transdisciplinary theoretical approach to clinical practice addressing voice and the communication of speaker socio-cultural positioning," Int. J. Speech Lang. Pathol. **22**(5), 559–570 (2020).
[18]E. Strand and K. Johnson, "Gradient and visual speaker normalization in the perception of fricatives," in *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference* (1996), pp. 14–26.
[19]K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory-visual integration of talker gender in vowel perception," J. Phon. **27**(4), 359–384 (1999).
[20]D. F. Kleinschmidt and F. T. Jaeger, "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," Psychol. Rev. **122**(2), 148–203 (2015).
[21]D. F. Kleinschmidt, K. Weatherholtz, and J. T. Florian, "Sociolinguistic perception as inference under uncertainty," Top. Cogn. Sci. **10**(4), 818–834 (2018).
[22]J. Hay and K. Drager, "Stuffed toys and speech perception," Linguistics **48**(4), 865–892 (2010).
[23]N. Niedzielski, "The effect of social information on the perception of sociolinguistic variables," J. Lang. Soc. Psychol. **18**(1), 62–85 (1999).
[24]K. Drager, "Speaker age and vowel perception," Lang. Speech **54**(1), 99–121 (2011).
[25]B. Munson, S. V. Jefferson, and E. C. McDonald, "The influence of perceived sexual orientation on fricative identification," J. Acoust. Soc. Am. **119**(4), 2427–2437 (2006).
[26]S. L. Bem, "The measurement of psychological androgyny," J. Consult. Clin. Psychol. **42**(2), 155–162 (1974).
[27]M. Hope and J. Lilley, "Gender expansive listeners utilize a non-binary, multidimensional conception of gender to inform voice gender perception," Brain Lang. **224**, 105049 (2022).
[28]A. Hancock, L. Colton, and F. Douglas, "Intonation and gender perception: Applications for transgender speakers," J. Voice **28**(2), 203–209 (2014).
[29]T. L. D. Hardy, C. A. Boliek, D. Aalto, J. Lewicke, K. Wells, and J. M. Rieger, "Contributions of voice and nonverbal communication to perceived masculinity-femininity for cisgender and transgender communicators," J. Speech. Lang. Hear. Res. **63**, 931–947 (2020).
[30]N. Houle and S. V. Levi, "Effect of phonation on perception of femininity/masculinity in transgender and cisgender speakers," J. Voice **35**, 497.e23 (2019).
[31]M. P. Gelfer and K. J. Schofield, "Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female versus those perceived as male," J. Voice **14**(1), 22–33 (2000).
[32]A. B. Hancock and S. F. Pool, "Influence of listener characteristics on perceptions of sex and gender," J. Lang. Soc. Psychol. **36**(5), 599–610 (2017).
[33]T. L. D. Hardy, C. A. Boliek, K. Wells, C. Dearden, C. Zalmanowitz, and J. M. Rieger, "Pretreatment acoustic predictors of gender, femininity, and naturalness ratings in individuals with male-to-female gender identity," Am. J. Speech. Lang. Pathol. **25**, 125–137 (2016).
[34]T. L. D. Hardy, J. M. Rieger, K. Wells, and C. A. Boliek, "Acoustic predictors of gender attribution, masculinity–femininity, and vocal naturalness ratings amongst transgender and cisgender speakers," J. Voice **34**(2), 300.e11–300.e26 (2020).
[35]R. Shrivastav, "Multidimensional scaling of breathy voice quality: Individual differences in perception," J. Voice **20**(2), 211–222 (2006).
[36]N. Houle, D. Goudelias, M. P. Lerario, and S. V. Levi, "Effect of anchor term on auditory-perceptual ratings of feminine and masculine speakers," J. Speech Lang. Hear. Res. **65**, 2064–2080 (2022).
[37]E. Atagi and T. Bent, "Auditory free classification of nonnative speech," J. Phon. **41**(6), 509–519 (2013).
[38]C. Brennan, A. C. Hannah, J. Romick, J. W. Lewon, and C. Meyers, "Differences and similarities in the perception of voice gender for individuals who are or are not members of the LGBT+ community," J. Voice. (published online 2022).
[39]Y. Leung, J. Oates, and S. P. Chan, "Voice, articulation and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis," J. Speech. Lang. Hear. Res. **61**, 266–297 (2018).
[40]C. Clopper, "Auditory free classification: Methods and analysis," Behav. Res. Methods **40**(2), 575–581 (2008).
[41]C. G. Clopper and D. B. Pisoni, "Free classification of regional dialects of American English," J. Phon. **35**, 421–438 (2007).
[42]M. C. Hout, M. H. Papesh, and S. D. Goldinger, "Multidimensional scaling," WIRES Cogn. Sci. **4**, 93–103 (2013).
[43]K. L. Lansford, J. M. Liss, and R. E. Norton, "Free-classification of perceptually similar speakers with dysarthria," J. Speech. Lang. Hear. Res. **57**(6), 2051–2064 (2014).
[44]S. Alcorn, K. Meemann, C. G. Clopper, and R. Smiljanic, "Acoustic cues and linguistic experience as factors in regional dialect classification," J. Acoust. Soc. Am. **147**(1), 657–670 (2020).
[45]S. Singh and T. Murry, "Multidimensional classification of normal voice qualities," J. Acoust. Soc. Am. **64**(1), 81–87 (1978).
[46]F. Ho and A. J. Mussap, "The gender identity scale: Adapting the gender unicorn to measure gender identity," Psychol. Sex Orientat. Gend. Divers **6**(2), 217–231 (2019).
[47]M. L. Haupert, "Considerations for the development and implementation of transgender-inclusive gender demographic questions," Ph.D. thesis, Indiana University, Bloomington, IN (2019).
[48]A. R. Bradlow, "ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings," https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar (Last viewed 3/1/2022).
[49]Sexual orientation information was not collected from speakers in the ALLSSTAR corpus.
[50]B. Merritt, Perceptual representation of speaker gender, Ph.D. thesis, Indiana University, Bloomington, IN (2022).
[51]M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099 (1994).
[52]M. P. Gelfer, "Perceptual attributes of voice: Development and use of rating scales," J. Voice **2**(4), 320–326 (1988).
[53]J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," Behav. Res. **51**, 195–203 (2019).
[54]A. Dhand, C. C. White, C. Johnson, Z. Xia, and P. L. De Jager, "A scalable online tool for quantitative social network assessment reveals potentially modifiable social environmental risks," Nat. Commun. **9**(1), 3930 (2018).
[55]M. P. Gelfer and V. A. Mikos, "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels," J. Voice **19**(4), 544–554 (2005).
[56]M. P. Gelfer and Q. E. Bennett, "Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender," J. Voice **27**(5), 556–566 (2013).
[57]M. Guzman, D. Muñoz, M. Vivero, N. Marín, M. Ramírez, M. Trinidad Rivera, C. Vidal, J. Gerhard, and C. González, "Acoustic markers to differentiate gender in prepubescent children's speaking and singing voice," Int. J. Pediatr. Otorhinolaryngol. **78**(10), 1592–1598 (2014).
[58]A. Hasanvand, A. Salehi, and M. A. Ebrahimipour, "A cepstral analysis of normal and pathologic voice qualities in Iranian adults: A comparative study," J. Voice **31**(4), 508.e17–508.e23 (2017).
[59]H. K. Vorperian and R. D. Kent, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," J. Speech. Lang. Hear. Res. **50**(6), 1510–1545 (2007).
[60]P. Boersma and D. Weenink. "Praat: Doing phonetics by computer [computer program]," http://www.praat.org/ (Last viewed January 22, (2021)).
[61]C. Clopper. "Free classification workshop," https://www.asc.ohio-state.edu/clopper.1/nu07/ (Last viewed February 11, 2023).
[62]IBM Corporation, "IBM SPSS Statistics for Windows" (2020).

[63]J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition," Psychometrika **35**(3), 283–319 (1970).
[64]B. Winter, *Statistics for Linguists* (Taylor & Francis, New York, 2020).
[65]J. Pierrehumbert, "Phonological representation: Beyond abstract versus episodic," Annu. Rev. Linguist. **2**, 33–52 (2016).
[66]D. L. Hamilton, S. J. Sherman, and C. M. Ruvolo, "Stereotype-based expectancies: Effects on information processing and social behavior," J. Soc. Issues **46**(2), 35–60 (1990).
[67]J. W. Mullennix and D. B. Pisoni, "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**(4), 379–390 (1990).
[68]C. G. Clopper and D. B. Pisoni, "Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization," Lang. Var. Change **16**(1), 31–48 (2004).
[69]G. Merrick, A. Figol, J. Anderson, and R. J. Lin, "Outcomes of gender affirming voice training: A comparison of hybrid and individual training modules," J. Speech. Lang. Hear. Res. **65**(2), 501–507 (2022).
[70]K. A. Chadwick, R. Coleman, K. Andreadis, M. Pitti, and A. Rameau, "Outcomes of gender-affirming voice and communication modification for transgender individuals," Laryngoscope **132**, 1615–1621 (2021).
[71]A. B. Hancock and L. M. Garabedian, "Transgender voice and communication treatment: A retrospective chart review of 25 cases," Intl. J. Lang. Comm. Disor. **48**(1), 54–65 (2013).
[72]D. Kawitzky and T. McAllister, "The effect of formant biofeedback on the feminization of voice in transgender women," J. Voice **34**(1), 53–67 (2018).
[73]A. Hancock and L. Helenius, "Adolescent male-to-female transgender voice and communication therapy," J. Commun. Disord. **45**(5), 313–324 (2012).
[74]United States Census Bureau, "Population and Housing Unit Estimates. United States Census Bureau," https://data.census.gov/table/ACSDP1Y2022.DP05?q=population+of+monroe+county+indiana+2022 (Last viewed September 29, 2023).
[75]M. Gordon, P. Barthmaier, and K. Sands, "A cross-linguistic acoustic study of voiceless fricatives," J. Int. Phon. Assoc. **32**(2), 141–174 (2002).
[76]K. Johnson, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," J. Phon. **34**(4), 485–499 (2006).
[77]R. Van Bezooijen, "Sociocultural aspects of pitch differences between Japanese and Dutch women," Lang Speech. **38**(3), 253–265 (1995).
[78]See the supplementary material for correlations between multidimensional scaling dimension coordinates and all acoustic-phonetic, auditory perceptual, and prototypicality indices measures for each free classification condition.