Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Voices of 'cyborg awesomeness': Posthuman embodiment of nonbinary gender expression in AI speech technologies

Maxwell Hope[1], *Éva Székely*[2]

[1]Department of Linguistics & Cognitive Science, University of Delaware, USA
[2]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
maxhope@udel.edu, szekely@kth.se

Abstract

Speech-generating devices (SGDs) provide users with text-to-speech (TTS) voices that shape identity and self-expression. Current TTS voices enable self-expression but often lack customizable features for authentic voice embodiment, particularly for nonbinary SGD users seeking gender affirmation as existing TTS voices largely reproduce binary, cisgender speech patterns. This study examines how nonbinary SGD users embody, or disembody, synthetic voices and the factors influencing voice affirmation. Through a survey, we analyze experiences of nonbinary SGD users and their impressions of generated speech samples, investigating the role of technological possibilities in gender affirmation and voice embodiment. Findings inform the creation of more user-centered TTS technologies, and challenge dominant paradigms in speech technology, gesturing toward a posthumanist rethinking of voice as co-constructed between human and machine.

Index Terms: nonbinary, transgender, posthuman, human-computer interaction, text-to-speech, voice embodiment

1. Introduction

For many individuals using assistive devices, these technologies become deeply integrated into their sense of self, shaping their physical and cognitive interactions with the world. Neuroscientific studies suggest that users may neurologically incorporate their assistive devices into their body schema, altering their perceptions of movement and autonomy [1]. This phenomenon is less explored in speech-generating devices (SGDs), which are a type of augmentative and alternative communication (AAC) that comprise of computerized text-to-speech (TTS) voices. SGDs serve as both communication tools and intermediaries between the self and the world. Similarly to other assistive devices, SGDs may become an embodied extension of the user's voice, integrating into their identity and influencing their interactions with others.

1.1. Identity and embodiment in text-to-speech systems

Embodiment is a multifaceted process involving self-perception, sensory experiences, and motor actions that can include objects external to the physical body [2]. The importance of embodiment in assistive technologies is exemplified by research on children with severe speech and physical impairments. Despite having access to SGDs, these children often prefer non-technological, embodied communication methods like gestures and body movements because they feel more natural and offer greater speed and control [3]. This underscores a critical gap in AAC design, which has traditionally prioritized functional communication over embodied communication. To better serve all users, particularly those within the transgender and nonbinary community, designers must prioritize embodiment and create technologies that truly integrate with the self. Expanding on this, Preece et al. [4] investigated how disabled individuals who use SGDs personalize their digital voices and navigate identity formation through these tools; for example, Preece, an SGD user, comments: *"It is more than just a voice—it's the voice of my thoughts and my identity."* Furthermore, Preece et al. [4] discuss the challenges of selecting a digital voice that accurately represents one's identity, noting the limitations of current speech synthesis technologies and the need for more expressive, individualized voice options.

1.2. Gender affirmation in voice

1.2.1. Gender dysphoria and euphoria

The existing literature on voice and the gender expansive (i.e. transgender and nonbinary) community suggests that voice is more than a means of communication; it is deeply tied to personal and social identity [5]. For gender expansive individuals, the relationship with voice is often fraught with complexities. Some gender expansive people experience voice dysphoria, a disconnect between how they sound and how they perceive their authentic self [6]. While much of the literature on transgender voice care focuses on mitigating voice dysphoria, a growing body of research highlights the importance of gender euphoria—affirming experiences of joy, alignment, and authenticity in one's gender expression. Austin et al. [7] emphasize that discussions of transgender well-being often center on distress rather than fulfillment, overlooking the impact of gender-affirming experiences; their study, which uses grounded theory and photo-elicitation interviews with gender expansive adults, identifies gender euphoria as a process that unfolds through four key stages: encountering a gender-affirming antecedent, developing an affirming thought, experiencing a positive emotion, and ultimately achieving an enhanced quality of life. These findings suggest that gender euphoria is not merely the absence of dysphoria but a distinct and meaningful phenomenon in its own right.

1.2.2. Gender affirmation in TTS voices

For gender expansive SGD users, therefore, the search for an affirming synthetic voice is not only about avoiding dysphoria but also about seeking euphoria—the feeling of recognition and empowerment that comes from embodying a voice that aligns with one's gender identity. This quest in the realm of SGDs has been constrained by major speech corpora used in synthetic voice development comprising only binary cisgender speakers

10.21437/Interspeech.2025-2229

(e.g. [8] [9] [10]). This is highlighted in recent qualitative studies with gender expansive SGD users who have said that there are not enough nonbinary or "middle-pitch" voice options on their SGD and often they are unable to shift the pitch of the voice on their SGD [11]. However, recent strides have been made to create corpora of gender expansive speech (such as [12] [13]) and to create more inclusive and modifiable TTS systems [14]. Using the Mid-Atlantic Gender Expansive Speech (MAGES) Corpus, [14] applied Constrained Principal Components Analysis to extract gender-independent speaker identity vectors, allowing for nuanced vocal adjustments in TTS models. Community evaluations conducted with nonbinary SGD users revealed that nonbinary SGD users would want voices modeled after gender expansive speakers on their own SGD and that they would want modifiable features, such as the ability to manipulate resonance, breathiness, and vocal tension.

However, gender affirmation in TTS voices may extend beyond manipulation of variables that are considered within the realm of "natural". Nonbinary genders already disrupt binary norms, aligning with posthumanist commitments to fluidity, multiplicity, and process-based identities. Technology, AI, and cybernetic augmentation provide tools for customizing gender expression in ways that parallel transhumanist and posthumanist explorations [15]. Therefore, new possibilities in TTS technology may be able to offer further nonbinary gender embodiment in SGD voices.

1.3. The current study

The process of "making" one's voice—selecting, modifying, and ultimately embodying a TTS voice—thus becomes a critical site of identity formation. This study explores the lived experiences of nonbinary SGD users, examining if and how they embody their synthetic voices and what factors facilitate or hinder this process. By situating voice embodiment within broader discussions of assistive technology, gender identity, and voice euphoria, this research seeks to illuminate the unique challenges and opportunities faced by nonbinary individuals in shaping and claiming their voice. Furthermore, it considers how advancements in TTS technology can better support the complex and diverse needs of all SGD users, ensuring that voice is not only functional but affirming of identity in ways that can be embodied, whether that embodiment is within human "naturalness" or beyond it.

2. Methods and Materials

2.1. Participants

Twelve participants were recruited via email from a listserv of nonbinary individuals who use SGDs. Participants were all over 18 years of age. No personally identifiable information was collected and their responses were anonymous. They described their gender in a variety of ways including "nonbinary", "agender", "gendervague" and "genderqueer".

2.2. TTS sample generation

2.2.1. Modifiable voice quality and acoustic vocal tract length features

We used the TTS system proposed by [14], trained on the MAGES corpus, which has controllable parameters along voice quality dimensions emerging from the speaker embeddings in the corpus, as well as explicit control of acoustic vocal tract length (aVTL) (see Sec. 2.4). We created 2 samples with breathy and tense voice qualities (emergent feature 1 adjusted by plus 1 standard deviation, and emergent feature 2 adjusted by 1 standard deviation), and 2 samples with different vocal tract length settings +5 and -3, which are normalized features translating to approximately 8 standard deviations longer and 5 standard deviations shorter aVTL than the mean for that speaker.

2.2.2. Gradual interpolation between two voice identities

We fine-tuned the publically available base model of XTTSv2 [16] on the 14 speakers of the MAGES corpus (1 GPU for 150k iterations). The XTTS system has an *interpolate* function which allows for the blending of two speakers based on two audio prompts. We modified this function to include weights, which allowed for a gradual interpolation between two speakers. We used this method to create 2 samples for the survey, using two speakers from the MAGES corpus using settings where the first sample was weighted 0.75 first speaker and 0.25 second speaker (MoreFirst) and the second sample was the inverse (MoreSecond).

2.2.3. Fluid transition of voice identities across a single sentence

To create a similar gradually transitioning effect over a single synthesized utterance, we modified the prosody-controllable gender-ambiguous TTS model open-sourced by [17]. The frame-level features originally used for prosody control were replaced by a vector interpolating between the two speaker embeddings. We refer to this method as the *Ombre* setting. For the survey, we synthesized two sentences beginning with one speaker embedding and transitioning to the other, over the 7.5s long sample.

2.3. Survey

A survey containing both close-ended and open-ended questions was administered through Qualtrics. The survey asked participants about their experiences using SGDs and their biological voice, if applicable. To demonstrate several of the features we were interested in, they were presented with the speech samples generated as described above and asked to imagine they were going to use this voice on their SGD and elaborate on their impressions of the voice, what they experienced in their body as they listened to the voice if anything, and whether or not the voice was compatible with their gender.

2.4. Acoustic analyses

Acoustic analyses were conducted in PRAAT [18]. The third formant (F3) at the midpoint of each vowel, minimum, maximum and average fundamental frequency (F0), and harmonics-to-noise (HNR) ratio were extracted from each of the speech samples presented in the survey. To calculate aVTL, we used [19]'s method as presented in [20]:

$$\Delta F = \text{mean}\left(\frac{F_3}{2.5}\right) \tag{1}$$

$$aVTL = \frac{34000}{2 \times \Delta F} \tag{2}$$

For the Ombre voice, because it shifted in speaker identity along 7.53 seconds, we divided the sample into three parts of 2.51 seconds for the aVTL calculation and took the aVTL at the

beginning portion and aVTL at the end portion in order to observe how this acoustic feature was presented at the beginning and end of the speech sample.

3. Results

3.1. Acoustic profiles of speech samples

Table 1 and Table 2 show the acoustic profiles of the first six voices participants heard. Table 3 shows the acoustic profile of the last voice, the Ombre voice.

Table 1: *Minimum, maximum and average F0 for the first six voices*

Voice	Min Pitch (Hz)	Max Pitch (Hz)	Avg Pitch (Hz)
Breathy	162.4	249.9	202.7
Tense	73.5	213.7	126.8
ShortVTL	173.8	233.8	193.1
LongVTL	75.8	125.3	98.5
MoreFirst	129.8	273.5	194.0
MoreSecond	98.1	160.3	122.6

Table 2: *Harmonics-to-Noise ratio (HNR) and acoustic Vocal Tract Length (aVTL) for the first six voices*

Voice	HNR (dB)	aVTL (cm)
Breathy	19.5	14.6
Tense	11.6	15.3
ShortVTL	21.3	12.6
LongVTL	11.0	16.3
MoreFirst	16.5	14.2
MoreSecond	14.0	14.6

Table 3: *Minimum, maximum and average F0, Harmonics-to-Noice ratio (HNR) and acoustic Vocal Tract Length (aVTL) for Ombre Voice*

Voice	Min Pitch (Hz)	Max Pitch (Hz)	Avg Pitch (Hz)
Ombre	79.3	494.2	168.5
Voice	**HNR (dB)**	**aVTL start (cm)**	**aVTL end (cm)**
Ombre	14.5	14.4	16.7

3.2. Participant reflections

3.2.1. Breathy voice

After listening to the speech sample that had increased breathiness, nine of the 12 participants indicated that they would want to control breathiness on their SGD. Eleven of the participants shared that it did *not* affirm their gender. Participants shared the following reflections:

Overall reflection:

> *"It felt clearer, more open. Breathy isn't the word I'd use for that; I associate breathy with, like, less voiced sound and more whisper, [this felt like] just as much voiced sound with more breath alongside it."*

Gender complication:

> *"I had a very strong 'wow, that isn't my voice' response to the point it was visceral."*
>
> *"I prefer a low-pitched voice for my gender expression, but the breathiness adds a quality I feel more connected to than the sample without it. I don't particularly identify with being human, so the breathiness feels right because it doesn't feel like Perfectly Standard Human Speech."*

3.2.2. Tense voice

After listening to the speech sample that had increased vocal tension, eleven of the 12 participants indicated that they would want to control vocal tension on their SGD. Seven participants indicated that this voice affirmed their gender. Participants shared the following reflections:

Overall reflections:

> *"It sounds rougher/gravellyer a little bit like the way I conceptualize the, like, stereotypical 'person experiencing testosterone-induced puberty (at any age)' quality of voice."*
>
> *"My instinct was that [it] feels so much more like my mouth speech in a way that I felt in my body."*
>
> *"I felt a sensation in my chest similar to when I use noises to communicate."*

Gender affirmation:

> *"This feels relatable in terms of my gender because it sounds a little more 'in between' typical cis norms (like [...] seeming relevant to T-induced puberties that are recent/new)."*
>
> *"Hell yes [this voice affirms my gender]. This knob embodies trans and tired, which is me."*

3.2.3. Exaggerated vocal tract length

After listening to the speech samples with greatly exaggerated and greatly shortened vocal tract lengths, all 12 participants indicated that they would want the ability to control vocal tract length to extreme lengths on their SGD. Regarding gender affirmation, five participants felt that the greatly lengthened vocal tract affirmed their gender, while only three participants felt the greatly shortened vocal tract affirmed their gender. Several participants indicated that the shortened vocal tract felt too feminine for their gender. Participants shared the following reflections:

Overall reflections:

> *"Wow that was so interesting, it felt like it's not trying to be human and I love that. I hadn't expected to latch onto this so much but I want something like this so much."*
>
> *"The shorter one felt weirdly monotone and tense; the longer one felt a little unnatural but pretty relaxed, and I might not've guessed it was artificially lengthened."*
>
> *"The shortened vocal tract was slightly unsettling but the lengthened one felt very nice to listen to. It's a similar sensation to playing notes on an instrument that are low enough you feel them as well as hear them."*

Gender affirmation (LongVTL):

> *"I loved the longer sample. It felt low and masculine enough but the robotic qualities made it feel less like a man's voice and more androgynous. The short sample was too high for me personally."*
>
> *"I want that voice. I love it. It feels deep, open, cavernous... It's not overly deep in pitch but it's deep in the ways that actually resonate with me. I've actually been very disappointed with the lack of options that sound like this."*
>
> *"This was very compatible with my gender. It had a lower pitch which was way more comfortable for me, but the editing of the length made it feel more nonbinary/androgynous than a cis man's voice. I might use this as a voice on my SGD depending on circumstances."*

Gender complication (ShortVTL):

> *"It's completely not in line with my gender expression but I definitely know people who might like to have a voice like that. It feels feminine, sharp, neither of which particularly describe my preferred voices."*

3.2.4. Voices with two speakers blended together

After listening to the speech samples with two speakers mixed together, all twelve participants indicated that they would want the ability to do this on their SGD. Regarding gender affirmation, none of the participants felt that the MoreFirst voice resonated with them. However, nine participants felt that the MoreSecond voice affirmed their gender. Participants shared the following reflections:

Overall reflections:

> *"Remarkably well-blended, I'd love to have this knob."*
>
> *"Euphoria at the range and possibilities. Deep desire for the ability to do this."*

Gender affirmation (MoreSecond voice):

> *"Loved it. Femme on balance but with the lower notes that make a good trans femme voice."*
>
> *"Extremely affirming, perhaps the most affirming I have heard of all voice options ever. Can I get it somewhere??? Perfect mix of my queerness and how I present myself and like to be read."*
>
> *"It was my voice. I've literally never found a voice that sounded so much like my verbal speech without voice banking. And the nasally-ness sounds how I speak and hear my voice on the inside. I think this would be my primary voice on an SGD."*

3.2.5. Ombre voice

After listening to the speech sample that changed gradually in speaker identity, nine of the twelve participants indicated that they would want this feature on their SGD. Regarding gender affirmation, six participants felt that this voice affirmed their gender. Some participants who did not find it affirming mentioned that the two endpoints were too binary-coded and suggested they might want this feature with different voices at the endpoints. Participants shared the following reflections:

Overall reflections:

> *"It is very cyborg awesomeness!"*
>
> *"It was weird. I would like this option, especially if there was the option to rapidly shift back and forth between registers/identities so that no one identity is clearly heard, but it felt very sing-song to me. Like the spoken equivalent of a vocalist singing scales."*

Gender affirmation:

> *"I loved how the voice was able to change over time like how someone going through puberty or on [Hormone Replacement Therapy]. It also seems like it would be affirming for gender-fluid people."*
>
> *"I really like it. I want it so bad. My gender is very multifaceted and having the ability to have my voice change like this as I talk sounds like a recipe for gender euphoria. Amazing. I love it."*
>
> *"The instability, the unpredictability, the no-I-don't-match-what-you-expect—that felt so good. This absolutely affirmed my gender, and while I'd not always want something that did this for the pure being easier on my communication partners, I would love it to be an option."*

Gender complication:

> *"*This* voice started at a Nope Not Me and ended at a different Nope Not Me, but the concept of transitioning like this is interesting to me, with different voices that are both/all at least sometimes me-ish."*
>
> *"It doesn't really suit my gender with the current voices."*

4. Discussion

Nonbinary SGD users provided insightful reflections that pointed towards voice embodiment and gender affirmation. Additionally, these reflections were also grounded in the acoustic profiles of the speech samples. Participants associated the Breathy voice with "openness" and "clarity", which corresponds to its higher HNR, while the Tense voice was described as "gravelly" or "rough", reflecting its lower HNR, consistent with existing literature on voice qualities [21]. The 3.7 cm difference between the exaggerated long and short vocal tract lengths likely exceeds normal human manipulation, contributing to perception that the voices *"felt like it's not trying to be human."* This aspect may support gender affirmation for individuals who do not associate with human gender. Similarly, in the mixed voices, samples from the same speakers yielded significantly different acoustics, highlighting the potential for creating diverse voices by combining acoustic features to varying degrees. Finally, the Ombre voice, with its wide F0 range (414.9 Hz) and notable aVTL difference (2.3 cm), demonstrated how varying acoustics over time can foster gender affirmation by challenging both gender norms and humanness.

The participants' responses to the generated voices were not only cognitive or evaluative; they frequently described visceral sensations in their bodies and strong emotional reactions to the voices, whether affirming or unsettling. One quote that particularly illuminates this is *"I had a very strong 'wow, that isn't my voice' response to the point it was visceral."* This highlights that voice embodiment is not just about having a voice but about how that voice interacts with self-perception, agency, and affective experience. The discomfort some participants expressed when a voice felt "wrong" was as pronounced as the gender euphoria experienced when a voice aligned with their identity.

These results illustrate that SGD users are not simply seeking to "correct" or "normalize" their voices to fit able-bodied standards. Instead, they desire a process of posthuman embodiment – one in which voice is not a fixed, biological given but an adaptable, modifiable extension of the self. So, what would it take for this to become everyday reality? Integrating novel technologies into real-world SGDs remains a complex challenge. Factors such as text input constraints, system latency, compatibility with existing architectures, user interface limitations and privacy concerns pose significant hurdles. Despite these practical barriers, it is important to stress that the stimuli in this study were not manually engineered or handcrafted; they were generated through fully reproducible processes. This means that the results are not only replicable and can be produced on-the-fly to facilitate real interaction, but also adaptable to new speech corpora and evolving technological frameworks. Thus, while implementation hurdles exist, the fundamental capability of current technology to support more expressive and identity-affirming TTS voices is already within reach – it is not a hypothetical or utopian endeavor, but a matter of design prioritization and deployment.

5. Conclusion

This study found that nonbinary SGD users desire voice features that add humanness such as the ability to manipulate voice texture (e.g. breathiness, tenseness) as well as features that go beyond traditional human constraints. Unlike those using biological voices, SGD users are not bound by the fixed physical properties of a vocal tract, allowing for a different kind of voice embodiment – one that is flexible, adaptable, and unconstrained by normative expectations of speech and gender expression. Some users preferred voices that actively subverted human norms, reinforcing a posthumanist vision of voice as adaptable, fluid, and co-constructed with technology.

6. Acknowledgments

This research is supported by the Swedish Reearch Council project Perception of speaker stance (VR-2020-02396), and the Riksbankens Jubileumsfond project CAPTivating (P20-0298). The authors would also like to extend deep gratitude to our participants from the gender expansive community who, despite much adversity, continue to bring light and hope into our world.

7. References

[1] M. Pazzaglia and M. Molinari, "The embodiment of assistive devices—from wheelchair to exoskeleton," *Physics of Life Reviews*, vol. 16, 2016.

[2] M. J. Giummarra, S. J. Gibson, N. Georgiou-Karistianis, and J. L. Bradshaw, "Mechanisms underlying embodiment, disembodiment and loss of embodiment," *Neuroscience & Biobehavioral Reviews*, vol. 32, no. 1, pp. 143–160, 2008.

[3] S. B. Ibrahim, A. Vasalou, and M. Clarke, "Design opportunities for aac and children with severe speech and physical impairments," in *Proc. CHI Conference on Human Factors in Computing Systems (CHI '18)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. Paper 227, 1–13.

[4] J. Preece, E. Sullivan, F. Tams-Gray, and G. Pullin, "Making my voice and owning its future," *Medical Humanities*, vol. 50, no. 4, pp. 624–634, 2024.

[5] L. Zimman, "Transgender voices: Insights on identity, embodiment, and the gender of the voice," *Language and Linguistic Compass*, 2018.

[6] J. Holmberg, "On voice dysphoria: Placing the transgender and gender diverse client at the centre of gender-affirming voice training," Ph.D. dissertation, Umeå University, Faculty of Medicine, Department of Clinical Sciences, Speech and Language Therapy, 2025.

[7] A. Austin, R. Papciak, and L. Lovins, "Gender euphoria: a grounded theory exploration of experiencing gender affirmation," *Psychology & Sexuality*, vol. 13, no. 5, pp. 1406–1426, 2022.

[8] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset, 2017.

[9] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," *ArXiv (Cornell University)*, 2021.

[10] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *ArXiv*, 2019.

[11] L. J. Martin and M. Nagalakshmi, "Bridging the social & technical divide in augmentative and alternative communication (aac) applications for autistic adults," *arXiv*, 2024, available at https://arxiv.org/pdf/2404.17730.

[12] D. V. Dolquist and B. Munson, "A palette of transmasculine voices," Retrieved from the Data Repository for the University of Minnesota, n.d., [Online]. Available: https://doi.org/10.13020/0fas-n510.

[13] M. Hope, "The mid-atlantic gender expansive speech (mages) corpus," [Online]. Available: https://maxwell-hope.com/mages-corpus/, n.d.

[14] É. Székely and M. Hope, "An inclusive approach to creating a palette of synthetic voices for gender diversity," in *Proc. Interspeech*, 2024, pp. 3070–3074.

[15] N. B. Mellamphy, "Challenging the humanist genre of gender: Posthumanisms and feminisms," in *Different Voices: Gender and Posthumanism*. V&R Unipress, 2022, ch. Chapter, pp. 15–28.

[16] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "XTTS: a massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech*, 2024, pp. 4978–4982.

[17] É. Székely, J. Gustafson, and I. Torre, "Prosody-controllable gender-ambiguous speech synthesis: a tool for investigating implicit bias in speech perception," in *Proc. Interspeech*, 2023, pp. 1234–1238.

[18] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]," http://www.praat.org/, 2021, version 6.1.48, retrieved 18 February 2021.

[19] P. E. Nordström and B. Lindblom, "A normalization procedure for vowel formant data," in *Proc. 8th International Congress of Phonetic Sciences*, Leeds, England, 1975, as presented in Johnson, K. 2020, *The F method of vocal tract length normalization for vowels*, Laboratory Phonology, 11(1): 10, pp. 1–16, DOI: 10.5334/labphon.196.

[20] K. Johnson, "The δf method of vocal tract length normalization for vowels," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, pp. 10, 1–16, 2020.

[21] G. de Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 4, pp. 794–811, 1995.