ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Resynthesis of Transmasculine Voices to Assess Gender Perception as a Function of Testosterone Therapy

Matti D. Groll,[a,b] Kimberly L. Dahl,[b] Manuel Díaz Cádiz,[b] Brett Welch,[c] Lauren F. Tracy,[d] and Cara E. Stepp[a,b,d]

[a] Department of Biomedical Engineering, Boston University, MA [b] Department of Speech, Language, and Hearing Sciences, Boston University, MA [c] Department of Communication Science and Disorders, University of Pittsburgh, PA [d] Department of Otolaryngology—Head and Neck Surgery, Boston University School of Medicine, MA

ARTICLE INFO

Article History:
Received September 7, 2021
Revision received December 22, 2021
Accepted March 31, 2022

Editor-in-Chief: Bharath Chandrasekaran
Editor: Jack J. Jiang

https://doi.org/10.1044/2022_JSLHR-21-00482

ABSTRACT

Purpose: The goal of this study was to use speech resynthesis to investigate the effects of changes to individual acoustic features on speech-based gender perception of transmasculine voice samples following the onset of hormone replacement therapy (HRT) with exogenous testosterone. We hypothesized that mean fundamental frequency (f_o) would have the largest effect on gender perception of any single acoustic feature.
Method: Mean f_o, f_o contour, and formant frequencies were calculated for three pairs of transmasculine speech samples before and after HRT onset. Sixteen speech samples with unique combinations of these acoustic features from each pair of speech samples were resynthesized. Twenty young adult listeners evaluated each synthesized speech sample for gender perception and synthetic quality. Two analyses of variance were used to investigate the effects of acoustic features on gender perception and synthetic quality.
Results: Of the three acoustic features, mean f_o was the only single feature that had a statistically significant effect on gender perception. Differences between the speech samples before and after HRT onset that were not captured by changes in f_o and formant frequencies also had a statistically significant effect on gender perception.
Conclusion: In these transmasculine voice samples, mean f_o was the most important acoustic feature for voice masculinization as a result of HRT; future investigations in a larger number of transmasculine speakers and on the effects of behavioral therapy-based changes in concert with HRT is warranted.

Gender diverse is an encompassing term representing individuals with a gender identity that does not align with their sex assigned at birth. This includes transmasculine individuals who identify with a different gender than the female sex assigned at birth, such as transgender men, as well as gender-fluid, agender, and nonbinary individuals. Due to widespread transphobia and gender policing, many gender-diverse individuals experience adverse mental and physical health outcomes and are disproportionately the victims of violent crimes (Bauer & Scheim, 2015; Dhejne et al., 2016; James et al., 2016). For these individuals, receiving desired gender attribution from others can impact their health, safety, and well-being (Dacakis et al., 2013; Hancock et al., 2011; Sirin et al., 2020; Watt et al., 2018).

Gender attribution can be directly influenced by an individual's speech (Hancock et al., 2011). The perception of the speaker and the listener both actively and dynamically contribute to gender attribution (Azul & Hancock, 2020).

Correspondence to Matti D. Groll: mgroll@bu.edu. ***Disclosure:*** *Cara E. Stepp has received consulting fees from Altec, Inc./Delsys, Inc., companies focused on developing and commercializing technologies related to human movement. Stepp's interests were reviewed and are managed by Boston University in accordance with their conflict of interest policies. The other authors have declared that no other competing financial or nonfinancial interests existed at the time of publication.*

Speech-based gender perception, or how the gender attributed to the speaker based on perception by the listener, is inherently problematic, given that it is often based on socially constructed gender stereotypes. Despite this, listeners do routinely make assumptions about an individual's gender based on characteristics of their speech, and for many gender-diverse individuals, speech-based gender perception is considered important. In a survey of 16 transmasculine individuals, 14 reported that speech changes were equally as important as surgical alterations for physical appearance (Van Borsel et al., 2000). Another survey of 30 transgender men found that voice masculinity was rated as the most important gender-related trait for transitioning (Hodges-Simeon et al., 2021). In remote settings, such as during phone calls, speech and voice are the most prominent gender cues. As a result, for some people, it is imperative for the speech-based perception of an individual's gender to match their gender identity. However, 96% of gender-diverse individuals have reported experiencing voice–gender incongruence (Kennedy & Thibeault, 2020). The result is that these individuals often seek voice treatment and/or therapy to better align their speech-based gender perception to their gender identity.

Speech treatment for transgender individuals often focuses on altering speech traits that are traditionally different between cisgender male and female speakers. Sex differences in cisgender speakers are well documented. Cisgender male speakers typically have thicker and larger vocal folds, corresponding to a lower average fundamental frequency (f_o) than cisgender female speakers (Titze, 1989). An average cisgender male adult speaker has a f_o of approximately 125 Hz, whereas an average cisgender female adult speaker has an average f_o of approximately 200 Hz (Titze, 1994). The f_o range of cisgender male speakers is often thought to be smaller when compared to cisgender female speakers, but when ranges are reported in semitones (ST), there appears to be little difference (Fitzsimons et al., 2001). However, cisgender male speakers may use fewer upward shifts in intonation during connected speech, resulting in a decrease in f_o variability (Hancock et al., 2014; Skuk & Schweinberger, 2014).

In addition to differences in the anatomy of the vocal folds, cisgender male speakers also have a longer vocal tract, because the male larynx lowers during puberty (Fant, 1966). Although body size also impacts the size of the vocal tract, on average, cisgender male speakers have a vocal tract length of 17–18 cm, whereas cisgender female speakers have a vocal tract length of 14–14.5 cm (Simpson, 2009). A longer vocal tract corresponds to the lower formant frequencies observed in male speakers (Hillenbrand et al., 1995; Skuk & Schweinberger, 2014; Titze, 1989). Thus, a combination of differences in f_o and formants are thought to contribute to speech-based gender perception in cisgender speakers.

Previous studies have used synthetically altered speech samples to observe how changing f_o and formant frequency values of cisgender speech samples to match the typical values of cisgender men and women affects speech-based gender perception. These studies used speech resynthesis algorithms such that speech-based gender perception could be compared across speech samples that had different f_o and formant frequency values but were otherwise identical. Both Whiteside (1998) and Gelfer and Mikos (2005) showed that, in synthesized vowel tokens with conflicting f_o and formant frequency values, gender perception often aligned with f_o. However, Gelfer and Bennett (2013) later observed that cisgender female speakers producing vowels within carrier sentences were still perceived as female even when their f_o was artificially shifted to a typical male range. The authors reasoned that gender perception becomes less of a function of f_o as the length and complexity of the utterance increases. Other studies that have used source-filter synthesizers to investigate the individual effects of changes in f_o and formant frequencies on sustained vowels and sentence length utterances in cisgender speakers concluded that f_o and formant frequencies are equally important for gender perception (Assmann et al., 2006; Smith et al., 2007). Though differences in the relative contribution of f_o and formant frequencies may be unclear, it is well established that combining synthetic changes to both f_o and formant frequency values in cisgender speakers results in a greater change in speech-based gender perception than either measure alone (Hillenbrand & Clark, 2009; Skuk & Schweinberger, 2014). Based on these changes in cisgender speakers, it is possible that individuals seeking to change speech-based gender perception may benefit from changing both f_o and formant frequency values.

For transmasculine speakers, a common intervention is hormone replacement therapy (HRT) with exogenous testosterone. The primary effect of HRT on speech is thought to be a decrease in f_o (Azul, 2015). Although decreases in f_o after the onset of HRT have been documented (Cler et al., 2020; Deuster et al., 2016; Nygren et al., 2016; Papp, 2012), not all individuals experience lowered f_o even after a year of HRT (Ziegler et al., 2018), and those who do show changes experience variable extents and rates (Hancock et al., 2017). Other changes to speech as a result of HRT are not as well documented. Three studies have investigated formant frequencies in transmasculine individuals. A longitudinal case study of one transmasculine individual over the course of a year of HRT observed decreases in fourth formant frequencies (Cler et al., 2020). Similarly, an unpublished thesis observed longitudinal decreases in the first three formants of six transmasculine individuals following HRT onset (Papp, 2012). Additionally, a larger cross-sectional study of 30 transgender men observed that, after onset of HRT, transmasculine speakers had significantly lower formant frequencies than cisgender female speakers, though they were not as low as the formant frequencies of cisgender male speakers (Hodges-Simeon et al., 2021).

Despite these potential speech changes, an estimated 31% of transmasculine speakers are not satisfied with their voice masculinity even after a year or more of HRT (Van Borsel et al., 2000). Furthermore, not all transmasculine individuals are able to or choose to undergo HRT, sometimes due to health care barriers for gender-diverse individuals (Bauer & Scheim, 2015; James et al., 2016). Thus, there is a need for techniques besides HRT that increase odds for desired gender attributions.

Due to their effects on speech-based gender perception in cisgender speakers, f_o and formant frequency values are often the targets of the limited number of speech therapy techniques that have been researched for transmasculine speakers. For example, laryngeal massage is a common voice therapy technique used to relax the laryngeal muscles in individuals with excessive laryngeal tension (Roy et al., 1997). This is thought to lower the larynx, thereby increasing the vocal tract and lowering the formant frequencies of the speaker (Roy & Ferguson, 2001). Laryngeal reposturing via manual therapy has also been shown to lower mean f_o in cisgender speakers with mutational falsetto (Dagli et al., 2008; Roy et al., 2017). Two studies have investigated the use of laryngeal massage and reposturing in a single transmasculine individual and reported increases in vocal tract length (Buckley et al., 2020) and decreases in mean f_o (Myers & Bell, 2020) following therapy. As a result, laryngeal massage and reposturing have potential to be used to target changes in f_o and formant frequencies in transmasculine speakers.

However, it is unclear whether changes to f_o and formant frequency values affect speech-based gender perception in transmasculine speakers to the same extent that they do in cisgender speakers. It has been shown that f_o is less correlated with speech-based gender perception in transfeminine speakers than in cisgender women, suggesting that mechanisms of speech-based gender perception may not be identical between transgender and cisgender individuals (McNeill et al., 2008). However, several studies have shown that transfeminine speakers are more likely to be perceived as female when f_o increases (Gelfer & Schofield, 2000; McNeill et al., 2008; Spencer, 1988; Wolfe et al., 1990) and when vowel formant frequencies were higher (Gelfer & Schofield, 2000; Mount & Salmon, 1988). In contrast, there have been less studies investigating speech-based gender perception in transmasculine individuals. Three studies have measured speech-based gender perception, all in relationship to the speech of a single transmasculine individual. Two studies monitored the speech changes of the same transmasculine individual over a year of HRT and observed that listeners reliably rated the individual as male after 28–37 weeks of testosterone (Brown et al., 2021; Cler et al., 2020). A third study investigated the effects of laryngeal massage and reposturing of the larynx on speech-based gender perception in the same transmasculine individual approximately 1.5 years after the start of HRT (Buckley et al., 2020). Though the individual did experience a decrease in f_o and formant frequencies, both as a result of HRT and laryngeal massage/reposturing, it is unclear how changes to these acoustic features individually impacted speech-based gender perception, nor are the results from a single transmasculine individual generalizable. Thus, there is a need to both determine which acoustic features change with HRT in transmasculine individuals and which individual acoustic features are most important for subsequent changes to speech-based gender perception. By monitoring how different acoustic features change in transmasculine individuals following voice masculinization and how those changes individually impact speech-based gender perception, targeted speech modification techniques may be developed to better aid transmasculine speakers seeking voice–gender congruence.

The purpose of this study was to use a speech resynthesis algorithm in order to assess how speech-based gender perception in transmasculine speech samples was affected by changes to individual speech traits. By resynthesizing speech samples using a combination of acoustic features from transmasculine speech samples before the onset of HRT and after a year of HRT, an auditory-perceptual experiment was used to answer the following research question: How do changes in acoustic features (mean f_o, f_o variability, and formant frequencies) affect the speech-based perception of gender in transmasculine speakers? We hypothesized that resynthesized samples with acoustic features from speech samples recorded after HRT onset would result in a more masculine gender perception, with mean f_o individually resulting in a greater change in gender perception than any other individual acoustic feature.

Method

Overview

A resynthesis algorithm was developed to combine the acoustic features of transmasculine speech samples recorded before and after HRT onset. The algorithm was developed using Legacy STRAIGHT speech synthesis, which separates an acoustic signal into its source and spectral content (Kawahara et al., 1999). These signals may be changed to reflect differences in acoustic features as a result of HRT and then resynthesized into novel speech samples. These speech samples were then used for an auditory-perceptual experiment. Listeners evaluated the speech-based gender perception and the synthetic quality of each sample from the stimuli set using a visual analog scale (VAS). These ratings were used to determine (a) which acoustic features had the greatest effect on speech-based gender perception and (b) whether the resynthesis process significantly influenced the synthetic quality of the samples. This research was approved by the Boston University Institutional Review Board, Protocol 2526. All listeners in the auditory-perceptual experiment provided informed consent.

Speech Samples

Speech samples were collected from three transmasculine speakers. Speech samples from two speakers were collected from a larger set of publicly available YouTube videos in which transmasculine individuals record brief speech samples at different time points over the course of HRT to monitor how their voices change. These two speakers were selected because they were native speakers of American English and recorded videos that did not have noticeable background noise in their recordings. Given that these speech samples were collected from prerecorded videos, demographics such as age cannot be reported.

A third speaker (32 years of age) recorded speech samples as part of a separate longitudinal case study (Cler et al., 2020) that monitored changes to speech in a transmasculine individual over a year of HRT. This individual was also a native speaker of American English, and his recordings were collected while seated in a soundproof booth using a standard headset microphone (WH20, Shure) placed approximately 6–10 cm from the mouth at a 45° angle from the midline. Signals were preamplified by an RME QuadMic II and sampled at 44,100 Hz with 16-bit resolution using a MOTU UltraLite-mk3 Hybrid (model UltraLite3Hy). Recordings were made using SONAR software (Cakewalk).

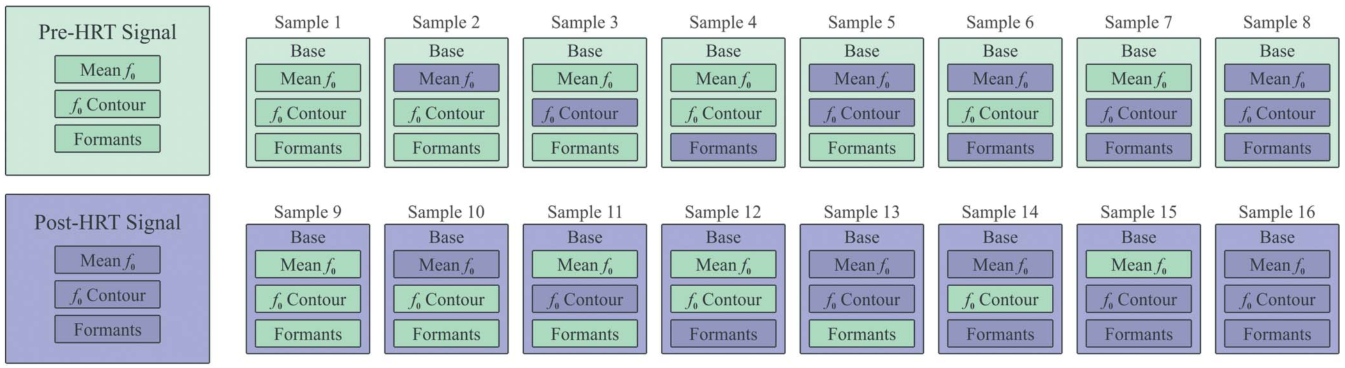
For the purposes of this study, each of the three speakers recorded one speech sample prior to the start of HRT (i.e., pre-HRT sample) and another speech sample after approximately 1 year of HRT (i.e., post-HRT sample). All three speakers demonstrated notable differences in speech-based gender perception between pre-HRT and post-HRT samples based on informal pilot testing. Each sample was a single sentence, 1–3 s in length, extracted from a longer recording. The speech stimulus was consistent within each set of pre-HRT and post-HRT samples but varied across speaker. Speech samples extracted from YouTube were taken from the start of the longer recording where the speaker repeated the same speech stimulus at each time point. This speech was semispontaneous in that it was not a reading passage but was a prepared introduction at each time point (i.e., “What’s up guys, it’s XX.” and “My name is XX and this is my voice.”). Both participants used gender-neutral names in their YouTube recordings. The speech sample recorded in the sound booth was a sentence (“The blue spot is on the key again.”) from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster et al., 2009). All acoustic measures were calculated solely from the extracted speech samples.

These pre-HRT and post-HRT samples were used to create the resynthesized speech samples used for the auditory-perceptual experiment. Resynthesis was performed using Legacy STRAIGHT speech synthesis, because it has been shown to successfully resynthesize samples without compromising the speech quality of the resynthesized samples, thereby reducing potential effects of speech synthesis on ratings of gender perception (Kawahara et al., 1999). For each sample, mean f_o, f_o contour, and formant values were calculated, resulting in three quantitative acoustic features that have been suggested to have an impact on speech-based gender perception. Differences between the pre-HRT and post-HRT values for each acoustic feature per speaker are shown in the Results section. For each pair of pre-HRT and post-HRT samples, these three features could be changed individually or together to reflect the values of the opposite sample. This resulted in a set of 16 new resynthesized samples composed of a combination of pre-HRT and post-HRT features (see Figure 1). Each new resynthesized sample could then be defined by whether its base, mean f_o, f_o contour, and formants were from the pre-HRT or post-HRT sample, with the base representing any other differences between the two samples that were not reflected in changes to f_o and formant frequencies. The final result was a stimuli set of 48 resynthesized speech samples from three transmasculine speakers. More information about the calculation of these three acoustic features and the algorithm used for resynthesis of this stimuli set can be found in the Appendix.

Auditory-Perceptual Study

Following development and implementation of the resynthesis algorithm, all 48 speech samples were used as stimuli for an auditory-perceptual experiment. Twenty young adults (10 men and 10 women, M = 24.2 years, SD = 4.2 years) were recruited for this study. Fifteen participants were cisgender, and five participants were transmasculine. Gender identity was not controlled or balanced during recruitment, because the average speech-based gender perception of transmasculine speakers has been shown to be unaffected by the gender minority status of the listener (Brown et al., 2021). All participants were individuals with no reports of speech, hearing, or language

Figure 1. Schematic showing the resynthesis of 16 unique speech samples via various combinations of the three acoustic features calculated from the pre-HRT (green) and post-HRT (purple) signals: mean fundamental frequency (Mean f_o), fundamental frequency contour (f_o Contour), and the first four formants averaged per vowel (Formants). Base corresponds to remaining acoustic information from the pre-HRT or post-HRT signal. HRT = hormone replacement therapy.



disorders. All provided informed consent in compliance with the Boston University Institutional Review Board.

This auditory-perceptual experiment was completed in a single sitting lasting approximately 1 hr. Each participant completed the study remotely via Gorilla (Cauldron Services) online services while communicating with an experimenter via a web call, such that instructions could be provided to the participant in real time. Participants first completed a brief task to verify that their audio setup satisfied the requirements for completing the study (i.e., audio was able to be automatically played and heard from the headphones of the participant). A Huggins Pitch test (Cramer & Huggins, 1958) developed and made publicly available by Milne et al. (2020) was used to determine that headphones were worn properly and that the study was completed in a quiet setting. Upon completion of these tasks, participants were given an overview of the study.

Participants completed two auditory-perceptual tasks. First, participants listened to each speech sample and were instructed to evaluate speech-based gender perception using a VAS. Participants were not given any information about the sex or gender of the speakers. Gender perception was defined to the participant as "what you think a speaker's gender is based on the characteristics of the speaker's voice." The VAS was scaled from 0 to 100, with anchors at 0 (*definitely a woman*), 25 (*probably a woman*), 50 (*completely uncertain*), 75 (*probably a man*), and 100 (*definitely a man*). The VAS had a slider that began at 50 and could be moved to anywhere along the scale in increments of one. Participants were not permitted to continue to the next sample without moving the slider. Speech samples were played automatically and could be manually replayed by the participant as many times as needed. After submitting a rating, participants were not permitted to return to previous samples. All 48 samples were presented to the participant in a randomized order. In order to ensure the reliability of a participant's evaluation, each sample was presented twice and averaged for a final speech-based gender perception rating.

Following the completion of the speech-based gender perception task, participants were then instructed to evaluate the synthetic quality of each speech sample using a similar VAS. This was used to determine to what extent using the resynthesis algorithm to change different acoustic features impacted the perceived synthetic quality of the speech samples. Synthetic quality was defined to the participant as a measure of "how much the sample sounds like it has been artificially synthesized, as opposed to produced naturally." For this task, the VAS was scaled from 0 to 100, with anchors at 0 (*natural*) and 100 (*synthetic*). Prior to rating samples, participants were also presented with three examples of speech samples that were produced naturally by three different speakers. These three examples were speech samples with various levels of background noise and voice quality that were selected from the larger set of YouTube videos. All three example speech samples were 1–3 s long to match the length of the resynthesized samples. Participants were told that each of these examples would, by definition, be considered to have zero synthetic quality. This was to mitigate the impact of natural differences between voice quality of speakers and recording environments on synthetic quality ratings. Two of the three example speech samples were spoken by speakers used in this study. However, these example speech samples were not part of the stimulus set used to create resynthesized samples. Following these examples, all 48 speech samples were presented to the participant in a randomized order. Participants were not told that the samples had been synthesized. As with speech-based gender perception, each sample was presented a second time such that final synthetic quality ratings were an average of the two ratings.

Descriptive and Statistical Analysis

Descriptive analysis was used to investigate average differences between the acoustic features and speech-based gender perception of pre-HRT and post-HRT speech samples for each speaker. These differences were examined to lend context to our primary analyses, not to make any conclusions about the effects of HRT on voice in general. Mean f_o was calculated via the f_o contour extracted using Legacy STRAIGHT and was reported in hertz. The variability of the f_o contour was quantified by using the standard deviation of the f_o contour in ST. Overall changes in formant frequencies were estimated using the average of the fourth formant in hertz. The fourth formant was selected because higher order formants are considered to be more directly correlated with vocal tract length (Wakita, 1977) and should therefore be more closely connected to speech-based gender perception. In contrast, lower order formant frequencies are more dependent on articulatory changes and are not always lower in male speakers (Hillenbrand et al., 1995). Average speech-based gender perception for pre-HRT and post-HRT speech samples were calculated by averaging the VAS gender perception of all listeners for Sample 1 and Sample 16, respectively, because these samples correspond to the pre-HRT and post-HRT speech samples without any changes in acoustic features (see Figure 1).

Intrarater and interrater reliability for speech-based gender perception and synthetic quality was calculated for all listeners using two-way intraclass correlation (ICC) analysis. Intrarater reliability was calculated by comparing each listener's first set of ratings to their second set of ratings. Interrater reliability was calculated by comparing the average ratings across listeners. The average intrarater reliability was ICC(2, 1) = .70 and .69 for speech-based gender perception and synthetic quality, respectively. The average interrater reliability was ICC(2, 1) = .65 and .61 for speech-based gender perception and synthetic quality, respectively. These results indicate moderate intrarater and interrater reliability for both speech-based gender perception and synthetic quality (Koo & Li, 2016).

Two repeated-measures analyses of variance (ANOVAs) were used to investigate speech-based gender perception and synthetic quality. For each of the 48 speech samples in the stimulus set, the mean f_o, f_o contour, and formants were categorized as either "pre" or "post" based on whether the corresponding acoustic measure originated from the pre-HRT or post-HRT signal. Similarly, the base of the speech sample was also categorized as either "pre" or "post." These categorical variables were main effects in the two ANOVAs. Four-way interaction effects between all four categorical variables were also included. Finally, speaker and listener identity was set as random factors. Partial eta-squared (η_p^2) values were calculated and qualitatively interpreted to determine very small ($\eta_p^2 < .01$), small ($\eta_p^2 \geq .01$), medium ($\eta_p^2 \geq .09$), or large ($\eta_p^2 \geq .25$) effect sizes for statistically significant factors (Witte & Witte, 2010). Statistical analysis was completed using Minitab 18 software with a significance level set a priori as $p < .05$.

Results

Overall Differences Between Pre-HRT and Post-HRT Speech Samples

Figure 2 shows differences in the three acoustic features (mean f_o, f_o contour, and formant frequencies) and average speech-based gender perception between pre-HRT and post-HRT speech samples for all three participants. Mean f_o decreased from an average of 184.4 Hz before HRT onset to 136.4 Hz after 1 year of HRT. Variability of the f_o contour, summarized by the standard deviation of the f_o contour, decreased from an average of 1.97–1.88 ST. The average of the fourth formant decreased from an average of 3841.0–3696.2 Hz. Speech-based gender perception increased from an average of 40.1 (near "completely uncertain") to 66.5 (near "probably a man") via VAS ratings. Although ratings differed for speakers, all ratings became more masculine, with two speakers moving from "completely uncertain" to "probably a man" and one speaker moving from "probably a woman" to approaching "completely uncertain." The average ratings of pre-HRT and post-HRT speech samples from Speaker 1 (samples recorded in a sound booth) were less masculine than pre-HRT and post-HRT speech samples from Speakers 2 and 3 (samples recorded from YouTube) as shown in Figure 2.

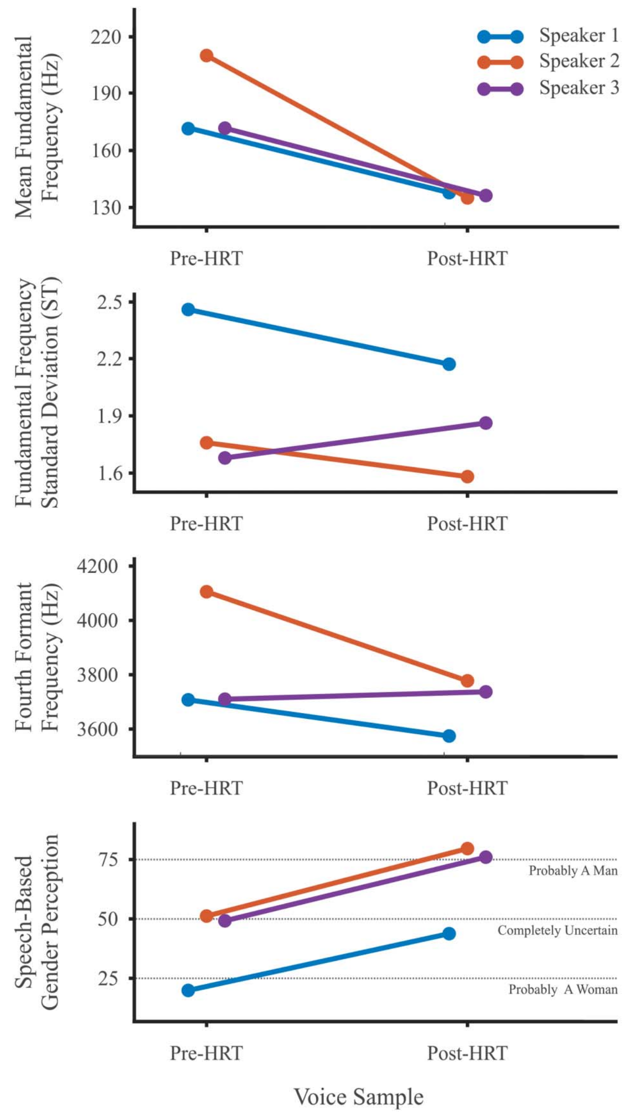
Speech-Based Gender Perception

The results of the ANOVA for speech-based gender perception are shown in Table 1. Speaker and mean f_o both had statistically significant effects on gender perception with large effect sizes. Listener had a medium significant effect, and the base of the speech sample had a small significant effect. Neither f_o contour nor formant frequencies had a statistically significant effect on gender perception. The only significant interaction effect was the two-way interaction between f_o contour and base, which had a very small effect size. The two-way interactions between all four fixed factors are shown in Figure 3.

Synthetic Quality

The results of the ANOVA for synthetic quality are shown in Table 2. Listener, speaker, mean f_o, f_o contour,

Figure 2. Differences in measures of mean fundamental frequency, fundamental frequency standard deviation, fourth formant frequency, and speech-based gender perception between speech samples before onset of hormone replacement therapy with exogenous testosterone (pre-HRT) and 1 year after onset (post-HRT) for each speaker are shown. Mean fundamental frequency and fourth formant frequency are plotted in hertz (Hz). Fundamental frequency standard deviation is plotted in semitones (ST). Speech-based gender perception is measured using a visual analog scale from 0 to 100. Dotted lines and labels indicate anchors placed on the visual analog scale.



and formants all had statistically significant effects on synthetic quality. Speaker had a large effect size, listener had a medium effect size, mean f_o and formants had small effect sizes, and f_o contour had a very small effect size. Base did not have a statistically significant effect on synthetic quality. The only statistically significant interaction effect was the two-way interaction between formants and base, which had a medium effect size. The two-way interactions between all four fixed factors are shown in Figure 4.

Discussion

First, it is paramount to acknowledge that judging or assuming one's gender based on their voice, speech, or outward presentation is an inherently problematic practice. The authors strongly discourage anyone from assuming a person's gender based on their physical characteristics, social presentation, and/or voice and speech patterns. This point notwithstanding, listeners do routinely make (potentially incorrect) assumptions and judgments about a person based on their speech. As previously mentioned, these judgments have tangible consequences for gender-diverse individuals, including their safety and well-being. As such, it is important for speech-language pathologists and vocal coaches to have evidence-based goals to target when working with transmasculine individuals who wish to masculinize their voice and speech. This study contributes to the relatively small body of literature that examines how individual acoustic features affect the speech-based gender perception and subsequent gender attribution of transmasculine individuals following a year of HRT.

Differences in Speech-Based Gender Perception Following Onset of HRT

In order to be able to detect possibly subtle differential effects on listener ratings, we intentionally chose three speakers that demonstrated notable differences in speech-based gender perception following the onset of HRT based on informal pilot testing. Thus, our finding that changes in speech-based gender perception for the three transmasculine speakers were clear, albeit moderate in size, is relatively unsurprising. Based on the listener ratings, pre-HRT and post-HRT samples differed by an average of 26.4 on the gender perception VAS. This difference corresponded to an increase from "completely uncertain" to "probably a man" for Speakers 2 and 3 (samples recorded from YouTube) and from "probably a woman" to "completely uncertain" for Speaker 1 (sample recorded in a sound booth).

The listener ratings for Speaker 1 are somewhat unexpected, given that this speaker was the same transmasculine individual who participated in a case study monitoring speech-based gender perception over the course of a year on HRT. In that case study, listeners noted a marked change in speech-based gender perception, with listeners reliably rating the speaker as male after 37 weeks (Cler et al., 2020). The differences in speech-based gender perception between the previous case study and this current study may be due to the fact that listeners

Table 1. Analysis of variance results for voice-based gender perception.

Factor	*df*	*F*	*p*	η_p^2	Effect size
Listener	19	7.23	< .01	.13	Medium
Speaker	2	660.57	< .01	.59	Large
Mean f_o	1	453.73	< .01	.33	Large
f_o Contour	1	0.81	.37	—	—
Formants	1	1.77	.18	—	—
Base	1	80.74	< .01	.08	Small
Mean f_o × f_o Contour	1	0.28	.60	—	—
Mean f_o × Formants	1	0.94	.33	—	—
Mean f_o × Base	1	0.41	.52	—	—
f_o Contour × Formants	1	0.30	.59	—	—
f_o Contour × Base	1	4.39	.04	< .01	Very small
Formants × Base	1	0.78	.38	—	—
Mean f_o × f_o Contour × Formants	1	0.11	.74	—	—
Mean f_o × f_o Contour × Base	1	0.55	.46	—	—
Mean f_o × Formants × Base	1	0.05	.82	—	—
f_o Contour × Formants × Base	1	0.48	.49	—	—
Mean f_o × f_o Contour × Formants × Base	1	0.08	.77	—	—

Note. Listener and speaker were set as random factors. Mean fundamental frequency (Mean f_o), fundamental frequency contour (f_o Contour), formant frequencies (Formants), base, and their four-way interactions were set as fixed factors. Em dashes indicate nonsignificant findings. *df* = degrees of freedom; η_p^2 = partial eta-squared value for evaluation of effect size.

were presented with a longer stimulus in the case study, approximately 13 s of Sentences 2–4 from the *Rainbow Passage* versus a 3-s sentence from the CAPE-V (Kempster et al., 2009) in this study, thereby giving them more information to evaluate speech-based gender perception. Furthermore, listeners in this study were presented with samples from two other speakers, whereas listeners in the case study only listened to samples from the same speaker. A significant effect of speaker and listener on speech-based gender perception indicates that there were inherent differences between each of the speakers, as well as how different listeners evaluated them. For all listeners, both

Figure 3. Interaction plots between the effects of mean fundamental frequency (Mean f_o), fundamental frequency contour (f_o Contour), formant values (Formants), and base on speech-based gender perception are plotted. Each plot corresponds to the interaction between the effects of the factors labeled by the corresponding row and column along the diagonal (e.g., the plot in the first row and second column is the interaction between Mean f_o and f_o Contour). The horizontal axis label represents whether the column factor originates from the speech sample before onset of hormone replacement therapy with exogenous testosterone (Pre) or 1 year after onset (Post). Likewise, line color and marker shape represent the origin of the row factor. **Indicates a statistically significant interaction effect at $p < .05$.

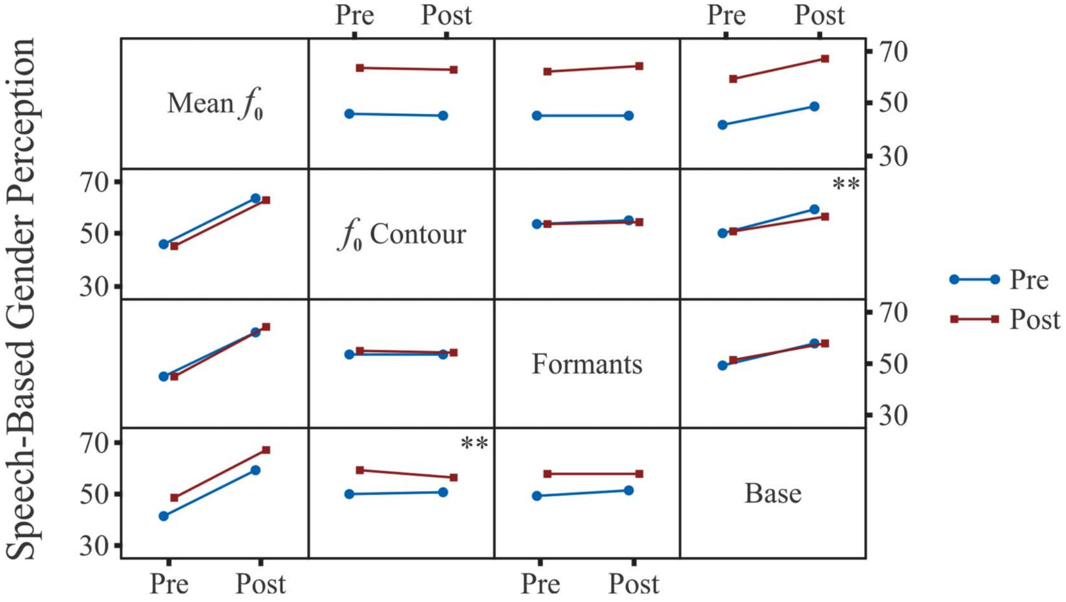


Table 2. Analysis of variance results for synthetic quality.

Factor	*df*	*F*	*p*	η_p^2	Effect size
Listener	19	12.14	< .01	.20	Medium
Speaker	2	633.20	< .01	.58	Large
Mean f_o	1	17.69	< .01	.02	Small
f_o Contour	1	3.89	.05	< .01	Very small
Formants	1	16.79	< .01	.02	Small
Base	1	0.63	.43	—	—
Mean f_o × f_o Contour	1	0.03	.86	—	—
Mean f_o × Formants	1	1.45	.23	—	—
Mean f_o × Base	1	2.38	.12	—	—
f_o Contour × Formants	1	0.60	.44	—	—
f_o Contour × Base	1	0.01	.94	—	—
Formants × Base	1	93.80	< .01	.09	Medium
Mean f_o × f_o Contour × Formants	1	0.38	.54	—	—
Mean f_o × f_o Contour × Base	1	0.01	.93	—	—
Mean f_o × Formants × Base	1	0.55	.46	—	—
f_o Contour × Formants × Base	1	0.12	.73	—	—
Mean f_o × f_o Contour × Formants × Base	1	0.31	.58	—	—

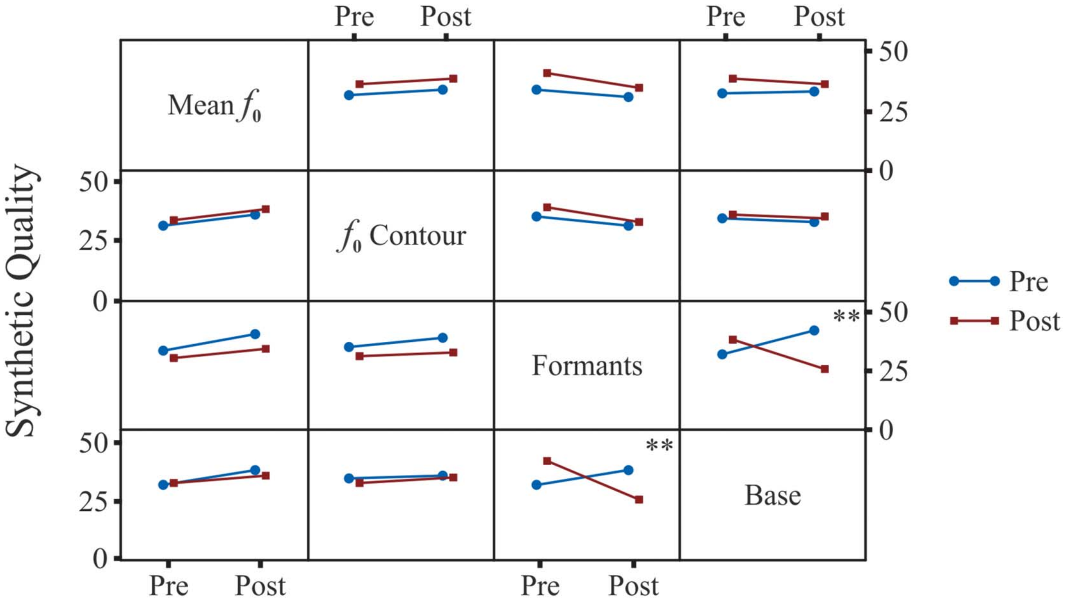
Note. Listener and speaker were set as random factors. Mean fundamental frequency (Mean f_o), fundamental frequency contour (f_o Contour), formant frequencies (Formants), base, and their four-way interactions were set as fixed factors. Em dashes indicate nonsignificant findings. *df* = degrees of freedom; η_p^2 = partial eta-squared value for evaluation of effect size.

Speakers 2 and 3 were consistently rated as sounding more masculine. This may have caused listeners to consistently rate Speaker 1 as less masculine by comparison. Likewise, gender perception for Speakers 2 and 3 may have also had a larger range if listeners exclusively heard samples from a single speaker.

Mean f_o Had a Large Effect on Changes in Speech-Based Gender Perception

Of the three individual acoustic factors investigated, only mean f_o had a statistically significant effect on speech-based gender perception in transmasculine speech

Figure 4. Interaction plots between the effects of mean fundamental frequency (Mean f_o), fundamental frequency contour (f_o Contour), formant values (Formants), and base on synthetic quality are plotted. Each plot corresponds to the interaction between the effects of the factors labeled by the corresponding row and column along the diagonal (e.g., the plot in the first row and second column is the interaction between Mean f_o and f_o Contour). The horizontal axis label represents whether the column factor originates from the speech sample before onset of hormone replacement therapy with exogenous testosterone (Pre) or 1 year after onset (Post). Likewise, line color and marker shape represent the origin of the row factor. **Indicates a statistically significant interaction effect at p < .05.



samples. Mean f_o had a large effect size, suggesting that, although other acoustics features also had an effect, in these three transmasculine individuals, mean f_o was a primary driver of the changes in speech-based gender perception. The importance of mean f_o on speech-based gender perception in these transmasculine individuals aligns with previous studies that have found similar results in cisgender individuals (Gelfer & Mikos, 2005; Whiteside, 1998) and transfeminine speakers (Gelfer & Schofield, 2000; McNeill et al., 2008; Spencer, 1988; Wolfe et al., 1990). This result supported our hypothesis that mean f_o would have the largest effect on speech-based gender perception of any single measure and suggests that the changes as a result of HRT in these samples were primarily driven by decreases in mean f_o.

Mean f_o also had a small but statistically significant effect on the synthetic quality of the speech sample. Specifically, when the mean f_o of the speech sample was from the post-HRT sample, listeners consistently rated the speech sample as more synthetic. Since there was no interaction effect between mean f_o and base, this was true regardless of whether the mean f_o was synthetically altered or not, indicating that the resynthesis process itself was not the reason for this change in synthetic quality. Given that mean f_o demonstrated a larger change following HRT onset than f_o contour or formant frequencies, it is possible that post-HRT speech samples presented listeners with conflicting gender cues. In the post-HRT condition, the mean f_o of all three participants was outside the typical range of mean f_o values for cisgender female speakers. In contrast, fourth formant frequency values decreased by an average of 144.8 Hz. This is a notably smaller difference than the average difference of 623.5 Hz observed by Hillenbrand et al. (1995) across cisgender male and female speakers. These conflicting gender cues may explain the inherent increase in synthetic quality in post-HRT speech samples. Specifically, listeners may perceive speech with contradictory gender cues as inherently more synthetic than speech without conflicting gender cues. Additionally, individuals have reported changes in voice quality as a result of HRT that may contribute to changes in the perception of synthetic quality, including a coarsening of the voice and strained voice quality (Azul et al., 2017; Gooren & Giltay, 2008; Nygren et al., 2016). In order to further explore this potential interpretation, future studies should investigate how ratings of synthetic quality change following HRT onset for a larger set of speech samples that have not been altered by a resynthesis algorithm.

Changing f_o Contour Did Not Contribute to the Perception of Voice Masculinization

Unlike mean f_o, f_o contour did not have a statistically significant effect on speech-based gender perception. Previous studies have shown that transfeminine speakers who were perceived as female used a higher number of pitch fluctuations and a larger proportion of upward intonational contours (Hancock et al., 2014; Wolfe et al., 1990). Thus, gender differences in f_o variability are likely the result of volitional changes made by the speaker. In this study, there appeared to be little-to-no differences in f_o variability between the pre-HRT and post-HRT speech samples of these three transmasculine speakers, at least when estimated by the standard deviation of f_o. Across all three speakers, the standard deviation of f_o decreased on average by a factor of 1.05 from pre-HRT to post-HRT samples, suggesting that these speakers did not use changes in f_o variability to increase voice masculinization. If speakers had actively decreased f_o variability, it is possible that there would have been larger differences in speech-based gender perception than those predominantly driven by a decrease in mean f_o. Future research should investigate whether volitional control of f_o variability drives additional changes in gender perception after mean f_o has decreased; if so, this could become an evidence-based therapeutic target, as it is in transfeminine speakers.

Although f_o contour did not have an overall effect on speech-based gender perception, there was an interaction effect between f_o contour and base. As shown in Figure 3 (Row 4, Column 2), the base of the speech sample affected whether changes in f_o contour resulted in increases or decreases in speech-based gender perception following HRT onset. Initially, this interaction effect appears to present contradicting effects of f_o contour on gender perception, but the effect size of this interaction effect is very small ($\eta_p^2 < .01$) and thus is unlikely to be clinically meaningful.

Changes in Formant Frequencies Did Not Affect Speech-Based Gender Perception

Formant frequency values also did not have a statistically significant effect on speech-based gender perception. Given that previous studies have shown that f_o and formant frequencies are both important for speech-based gender perception, especially in connected speech (Hillenbrand & Clark, 2009), these results are surprising. In this study, which used connected speech, mean f_o had a statistically significant effect on speech-based gender perception, but formant frequencies did not. However, there are differences in the design of this study when compared to previous studies that must be considered when interpreting these results.

Previous studies that have investigated the effects of f_o and formant frequencies on speech-based gender perception have used resynthesis algorithms to artificially shift f_o and formant frequency values to the typical range of cisgender speakers of a different sex (e.g., Assmann et al., 2006; Smith et al., 2007). As a result, these previous studies were shifting acoustic features by the theoretical difference between a typical cisgender male and female speaker. In contrast, this study shifted acoustic features of transmasculine speech based on differences between pre-HRT and post-HRT speech samples from the same individual. Our goal was to capture actual changes observed in this set of transmasculine speakers, which may differ from the theoretical differences between cisgender male and female speakers. For example, the fourth formant frequencies of vowels produced by cisgender male and female speakers may differ by over 600 Hz in certain vowels (Hillenbrand et al., 1995). Research on how formant frequencies change in transmasculine voices is limited, but in a longitudinal case study of a single transmasculine individual, Cler et al. (2020) observed a decrease of 140 Hz in fourth formant frequencies a year after HRT onset. In a study by Hodges-Simeon et al. (2021), transmasculine speakers were observed to have formant frequencies that fell between those of cisgender male and female speakers. Thus, transmasculine individuals may be unlikely to experience changes as large as the differences observed between cisgender male and female speakers. By using changes that had actually occurred within a transmasculine individual instead of artificially altering speech to match a cisgender standard, we were able to investigate how speech-based gender perception changed as a result of realistic changes that may occur during voice masculinization via HRT.

Using actual changes between the pre-HRT and post-HRT samples may have resulted in differences that were not large enough to meaningfully affect speech-based gender perception. Indeed, across all three speakers, the fourth formant value decreased, on average, by a factor of 1.04 (factors for first, second, and third formants were 1.01, 1.04, and 1.04, respectively). This decrease is notably smaller than the average decrease in f_o, corresponding to a factor of 1.35, which may explain why f_o had an effect on gender perception in these speakers, but formant frequencies did not. In this study, the moderate increase in voice masculinization was not significantly affected by small changes in formant frequencies, but it is possible that larger changes in formant frequencies may have resulted in greater changes in gender perception.

This relatively small change in formant frequencies is difficult to contextualize, because very few studies have investigated how formant frequencies change in transmasculine individuals. Although longitudinal changes were not calculated, Hodges-Simeon et al. (2021) observed that the formant frequencies of transmasculine speakers were lower than cisgender female speakers by an average factor of 1.05, similar to the factor observed in this study. In a longitudinal study, Papp (2012) observed significant changes in formant frequencies of six transmasculine individuals following a year of HRT, though a factor could not be calculated, because average frequency values were not reported. However, Papp continued to monitor formant frequencies of two of the transmasculine individuals for 2–3 years and observed that formant frequencies continued to change even after the 1 year mark. This indicates that larger changes in formant frequencies may occur after a longer amount of time on HRT. Likewise, a case study with a single transmasculine individual measured a decrease in the fourth formant frequency by a factor of 1.14 following a session of laryngeal massage and reposturing (Buckley et al., 2020), suggesting that speech modification may be used to further lower formant frequencies as well. These larger changes in formant frequencies may drive additional changes in speech-based gender perception than those observed in this study.

The results of the ANOVA for synthetic quality should also be considered when evaluating the contributions of formant frequencies on speech-based gender perception in this study. As shown in Figure 4 (Column 4, Row 3), the two-way interaction between formant frequencies and base had a statistically significant effect on synthetic quality with a medium effect size (η_p^2 = .09). Specifically, the base of the speech sample affected whether listener perceptions of synthetic quality increased or decreased when formant frequencies were changed. This indicates that when the resynthesis algorithm was used to shift formants, listeners perceived the speech samples as more synthetic. This weakens the validity of listeners' ratings of gender perception in samples with resynthesized formants. Combined with the fact that there were only small changes in formants between the pre-HRT and post-HRT samples, conclusions about formant frequencies from the results of this study should be interpreted with caution. It is clear that more work is needed to understand the impact of formant frequencies on gender perception in transmasculine speakers.

Additional Acoustic Features Contribute to Changes in Speech-Based Gender Perception

Speech-based gender perception was statistically significantly affected by the base of the speech sample with a small effect size (η_p^2 = .08). In other words, gender perception was affected by whether the synthesized speech sample originated from the pre-HRT sample or the post-HRT sample, independent of whether the three selected acoustic features were changed. In the resynthesis process, the base of the speech sample may be considered as a representation of all other differences between the pre-HRT and post-HRT samples that are not captured by f_o and formant frequencies. Based on these results, other acoustic differences between the pre-HRT and post-HRT samples may have an effect on gender perception, though the combined effect size of these acoustic differences is notably smaller than the large effect size of mean f_o (η_p^2 = .08 vs. η_p^2 = .33). This result is similar to a previous study that also used Legacy STRAIGHT to investigate the effects of changes to acoustic features on gender perception. Assmann et al. (2006) observed that utterances originally spoken by cisgender female speakers were more likely to be heard as feminine even after changing f_o and formant frequencies to match typical values of cisgender male speakers, and vice versa, for utterances spoken by cisgender male speakers. The authors suggested that there were residual indicators of speech-based gender not captured by f_o and formant frequencies.

Previous studies have suggested that voice quality may differ between male and female speakers, with female speakers using breathier and creakier voices (Becker et al., 2014; Klatt & Klatt, 1990). Additional studies have also suggested that female speakers produce the sibilant /s/ at a higher frequency than male speakers (Flipsen et al., 1999; Zimman, 2018). Potential differences in these acoustic features are likely represented by the base of the speech sample, though they were not evaluated in this study. Future studies should explore to what extent individual changes to these acoustic features impact transmasculine speech-based gender perception.

Limitation to Study Design

In this study, sets of pre-HRT and post-HRT samples were used to investigate the effects of changes to acoustic features on speech-based gender perception such that actual changes in transmasculine speech could be evaluated, as opposed to artificial changes based on typical cisgender ranges. Therefore, although the results of this study indicate that the changes in speech-based gender perception in these samples occurring due to real-life changes are largely driven by f_o, it is important to note that this result may not hold true for potentially larger changes in gender perception.

Though this study investigated acoustic features that had changed in transmasculine speakers as a result of HRT, this study did not investigate which additional acoustic features could change in these speakers. Specifically, future research is needed to show the effects of intentional behavioral modifications (e.g., posturing the larynx lower to elongate the vocal tract and altering intonation patterns), which could be part of therapeutic approaches regardless of the use of HRT. The same resynthesis approach as in this study would enable further investigation into these factors independently.

Speech-based gender perception of the listener is only part of what contributes to a speaker's gender attribution, which has been suggested to be determined by interacting factors from the speaker, the listener, and biocultural forces (Azul & Hancock, 2020). These interactions may lead to a dynamic gender attribution in gender-diverse speakers. Future studies should also investigate a similar resynthesis approach to the one used in this study to explore potential dynamic changes in gender attribution as a result of these additional factors.

This study was limited by the number of speakers, a common trend with research relating to the transgender community. In this study, samples from one transgender speaker were collected from a previous longitudinal study (Cler et al., 2020). The remaining two transgender speakers were collected through publicly available YouTube videos. Though there were additional recordings available, they were excluded due to the presence of excessive background noise, the lack of a repeated utterance across pre-HRT and post-HRT samples, or the language of the speaker.

It is also impossible to fully characterize each participant's HRT as a result of using data for two of the three speakers from publicly available YouTube videos. It is unknown what the exact treatment that each participant entailed and whether they participated in any other forms of voice treatment between time points. As a result, conclusions about differences between pre-HRT and post-HRT samples in this study with a limited number of speakers should be interpreted with caution.

It is clear that there is a need for a study that can record a large number of transmasculine speakers before and after the onset of HRT in a quiet recording environment, which could apply a similar resynthesis approach to the one used in this study. Participants can then also provide specific information about their individual course of HRT and other additional differences between time points. Recruiting a larger number of participants will likely result in transmasculine speakers that demonstrate variable magnitudes of change across a range of acoustic features. This larger study will then be able to investigate how gender perception differs between these individuals in order to further understand the effects of individual acoustic features on speech-based gender perception in transmasculine speakers.

Conclusions

The results of this study demonstrate the use of a resynthesis algorithm using Legacy STRAIGHT to evaluate the effects of individual acoustic features on speech-based gender perception in transmasculine speakers. Mean f_o had a statistically significant effect on gender perception with a large effect size. This indicates that, based on these three transmasculine speakers, mean f_o is the primary driving force in voice masculinization as a result of HRT. However, given that the transmasculine speakers in this study only experienced a moderate change in speech-based gender perception and that there was little-to-no change in f_o contour and formant frequencies, it is possible that larger changes in speech-based gender perception may be achieved by making changes to these other acoustic features. To better understand the effects of all of these acoustic features, there is a need for larger studies on the speech-based gender perception of transmasculine speakers.

Acknowledgments

This work was supported by a Boston University Hariri Institute Research Incubation Award and by the National Institutes of Health Grants R01DC015570 (C.E.S.), R01DC020061 (C.E.S.), and T32DC013017 (C.A.M. and C.E.S.). This article was reviewed by a transmasculine individual with an expertise in speech science. The authors would like to thank him for his anonymous contributions. The authors would also like to thank Megan Cushman for aiding with recruitment.

References

Assmann, P. F., Nearey, T. M., & Dembling, S. (2006). *Effects of frequency shifts on perceived naturalness and gender information in speech*. Paper presented at the Proceedings of the 9th International Conference on Spoken Language Processing.

Azul, D. (2015). Transmasculine people's vocal situations: A critical review of gender-related discourses and empirical data. *International Journal of Language & Communication Disorders, 50*(1), 31–47. https://doi.org/10.1111/1460-6984.12121

Azul, D., & Hancock, A. B. (2020). 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. *International Journal of Speech-Language Pathology, 22*(5), 559–570. https://doi.org/10.1080/17549507.2019.1709544

Azul, D., Nygren, U., Södersten, M., & Neuschaefer-Rube, C. (2017). Transmasculine people's voice function: A review of the currently available evidence. *Journal of Voice, 31*(2), 261.e9–261.e23. https://doi.org/10.1016/j.jvoice.2016.05.005

Bauer, G. R., & Scheim, A. I. (2015). *Transgender people in Ontario, Canada: Statistics from the Trans PULSE Project to inform human rights policy*. Trans PULSE. https://transpulseproject.ca/wp-content/uploads/2015/06/Trans-PULSE-Statistics-Relevant-for-Human-Rights-Policy-June-2015.pdf

Becker, K., Khan, S. U. D., & Zimman, L. (2014). Voice quality variation and gender. *The Journal of the Acoustical Society of America, 136*(4), 2295–2295. https://doi.org/10.1121/1.4900303

Brown, K. M., Dahl, K. L., Cler, G. J., & Stepp, C. E. (2021). Listener age and gender diversity: Effects on voice-based perception of gender. *Journal of Voice, 35*(5), 739–745. https://doi.org/10.1016/j.jvoice.2020.02.004

Buckley, D. P., Dahl, K. L., Cler, G. J., & Stepp, C. E. (2020). Transmasculine voice modification: A case study. *Journal of Voice, 34*(6), 903–910. https://doi.org/10.1016/j.jvoice.2019.05.003

Cler, G. J., McKenna, V. S., Dahl, K. L., & Stepp, C. E. (2020). Longitudinal case study of transgender voice changes under testosterone hormone therapy. *Journal of Voice, 34*(5), 748–762. https://doi.org/10.1016/j.jvoice.2019.03.006

Cramer, E. M., & Huggins, W. H. (1958). Creation of pitch through binaural interaction. *The Journal of the Acoustical Society of America, 30*(5), 413–417. https://doi.org/10.1121/1.1909628

Dacakis, G., Davies, S., Oates, J. M., Douglas, J. M., & Johnston, J. R. (2013). Development and preliminary evaluation of the transsexual voice questionnaire for male-to-female transsexuals. *Journal of Voice, 27*(3), 312–320. https://doi.org/https://doi.org/10.1016/j.jvoice.2012.11.005

Dagli, M., Sati, I., Acar, A., Stone, R. E., Dursun, G., & Eryilmaz, A. (2008). Mutational falsetto: Intervention outcomes in 45 patients. *The Journal of Laryngology & Otology, 122*(3), 277–281. https://doi.org/10.1017/S0022215107008791

Deuster, D., Matulat, P., Knief, A., Zitzmann, M., Rosslau, K., Szukaj, M., am Zehnhoff-Dinnesen, A., & Schmidt, C. M. (2016). Voice deepening under testosterone treatment in female-to-male gender dysphoric individuals. *European Archives of Oto-Rhino-Laryngology, 273*(4), 959–965. https://doi.org/10.1007/s00405-015-3846-8

Dhejne, C., Van Vlerken, R., Heylens, G., & Arcelus, J. (2016). Mental health and gender dysphoria: A review of the literature. *International Review of Psychiatry, 28*(1), 44–57. https://doi.org/10.3109/09540261.2015.1115753

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *STL-QPSR, 7*(4), 022–030.

Fitzsimons, M., Sheahan, N., & Staunton, H. (2001). Gender and the integration of acoustic dimensions of prosody: Implications for clinical studies. *Brain and Language, 78*(1), 94–108. https://doi.org/10.1006/brln.2000.2448

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeny, J. (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research, 42*(3), 663–677. https://doi.org/10.1044/jslhr.4203.663

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender. *Journal of Voice, 27*(5), 556–566. https://doi.org/10.1016/j.jvoice.2012.11.008

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice, 19*(4), 544–554. https://doi.org/10.1016/j.jvoice.2004.10.006

Gelfer, M. P., & Schofield, K. J. (2000). Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female versus those perceived as male. *Journal of Voice, 14*(1), 22–33. https://doi.org/10.1016/S0892-1997(00)80092-2

Gooren, L. J. G., & Giltay, E. J. (2008). Review of studies of androgen treatment of female-to-male transsexuals: Effects and risks of administration of androgens to females. *Journal of Sexual Medicine, 5*(4), 765–776. https://doi.org/10.1111/j.1743-6109.2007.00646.x

Hancock, A., Colton, L., & Douglas, F. (2014). Intonation and gender perception: Applications for transgender speakers. *Journal of Voice, 28*(2), 203–209. https://doi.org/10.1016/j.jvoice.2013.08.009

Hancock, A. B., Childs, K. D., & Irwig, M. S. (2017). Trans male voice in the first year of testosterone therapy: Make no assumptions. *Journal of Speech, Language, and Hearing Research, 60*(9), 2472–2482. https://doi.org/10.1044/2017_JSLHR-S-16-0320

Hancock, A. B., Krissinger, J., & Owen, K. (2011). Voice perceptions and quality of life of transgender people. *Journal of Voice, 25*(5), 553–558. https://doi.org/10.1016/j.jvoice.2010.07.013

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*(5), 1150–1166. https://doi.org/10.3758/APP.71.5.1150

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hodges-Simeon, C. R., Grail, G. P. O., Albert, G., Groll, M. D., Stepp, C. E., Carré, J. M., & Arnocky, S. A. (2021).

Testosterone therapy masculinizes speech and gender presentation in transgender men. *Scientific Reports, 11*(1), Article 3494. https://doi.org/10.1038/s41598-021-82134-2

James, S. E., Herman, J. L., Rankin, S., Keisling, M., Mottet, L., & Anafi, M. (2016). *The report of the 2015 U.S. transgender survey*. National Center for Transgender Equality. https://transequality.org/sites/default/files/docs/usts/USTS-Full-Report-Dec17.pdf

Kawahara, H. (2018). *Legacy STRAIGHT (GitHub repository)*. https://github.com/HidekiKawahara/legacy_STRAIGHT

Kawahara, H., Masuda-Katsuse, I., & de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication, 27*(3–4), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124–132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Kennedy, E., & Thibeault, S. L. (2020). Voice–gender incongruence and voice health information-seeking behaviors in the transgender community. *American Journal of Speech-Language Pathology, 29*(3), 1563–1573. https://doi.org/10.1044/2020_AJSLP-19-00188

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820–857. https://doi.org/10.1121/1.398894

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

McNeill, E. J. M., Wilson, J. A., Clark, S., & Deakin, J. (2008). Perception of voice in the transgender client. *Journal of Voice, 22*(6), 727–733. https://doi.org/10.1016/j.jvoice.2006.12.010

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Mount, K. H., & Salmon, S. J. (1988). Changing the vocal characteristics of a postoperative transsexual patient: A longitudinal study. *Journal of Communication Disorders, 21*(3), 229–238. https://doi.org/10.1016/0021-9924(88)90031-7

Mustafa, K., & Bruce, I. C. (2006). Robust formant tracking for continuous speech with speaker variability. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(2), 435–444. https://doi.org/10.1109/TSA.2005.855840

Myers, B., & Bell, T. (2020). Adapting vocal function exercises for voice masculinization. *Perspectives of the ASHA Special Interest Groups, 5*(4), 861–866. https://doi.org/10.1044/2020_PERSP-20-00076

Nygren, U., Nordenskjold, A., Arver, S., & Sodersten, M. (2016). Effects on voice fundamental frequency and satisfaction with voice in trans men during testosterone treatment-a longitudinal study. *Journal of Voice, 30*(6), 766.e23–766.e34. https://doi.org/10.1016/j.jvoice.2015.10.016

Papp, V. G. (2012). *The female-to-male transsexual voice: Physiology vs. performance in production* [Doctoral dissertation, Rice University].

Roy, N., Bless, D. M., Heisey, D., & Ford, C. N. (1997). Manual-circumlaryngeal therapy for functional dysphonia: An evaluation of short- and long-term treatment outcomes. *Journal of Voice, 11*(3), 321–331. https://doi.org/10.1016/s0892-1997(97)80011-2

Roy, N., & Ferguson, N. A. (2001). Formant frequency changes following manual circumlaryngeal therapy for functional dysphonia: Evidence of laryngeal lowering? *Journal of Medical Speech-Language Pathology, 9,* 169–175.

Roy, N., Peterson, E. A., Pierce, J. L., Smith, M. E., & Houtz, D. R. (2017). Manual laryngeal reposturing as a primary approach for mutational falsetto. *The Laryngoscope, 127*(3), 645–650. https://doi.org/10.1002/lary.26053

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621–640. https://doi.org/10.1111/j.1749-818X.2009.00125.x

Sirin, S., Polat, A., & Alioglu, F. (2020). Voice-related gender dysphoria: Quality of life in hormone naïve trans male individuals. *Alpha Psychiatry, 21*(1), 53–60. https://doi.org/10.5455/apd.41947

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research, 57*(1), 285–296. https://doi.org/10.1044/1092-4388(2013/12-0314)

Smith, D. R. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *The Journal of the Acoustical Society of America, 122*(6), 3628–3639. https://doi.org/10.1121/1.2799507

Spencer, L. E. (1988). Speech characteristics of male-to-female transsexuals: A perceptual and acoustic study. *Folia Phoniatrica, 40*(1), 31–42. https://doi.org/10.1159/000265881

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Titze, I. R. (1994). *Principles of voice production*. Prentice Hall.

Van Borsel, J., De Cuypere, G., Rubens, R., & Destaerke, B. (2000). Voice problems in female-to-male transsexuals. *International Journal of Language & Communication Disorders, 35*(3), 427–442. https://doi.org/10.1080/136828200410672

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 25*(2), 183–192. https://doi.org/10.1109/TASSP.1977.1162929

Watt, S. O., Tskhay, K. O., & Rule, N. O. (2018). Masculine voices predict well-being in female-to-male transgender individuals. *Archives of Sexual Behavior, 47*(4), 963–972. https://doi.org/10.1007/s10508-017-1095-1

Whiteside, S. P. (1998). The identification of a speaker's sex from synthesized vowels. *Perceptual and Motor Skills, 87*(2), 595–600. https://doi.org/10.2466/pms.1998.87.2.595

Witte, R. S., & Witte, J. S. (2010). *Statistics*. Wiley.

Wolfe, V. I., Ratusnik, D. L., Smith, F. H., & Northrop, G. (1990). Intonation and fundamental frequency in male-to-female transsexuals. *Journal of Speech and Hearing Disorders, 55*(1), 43–50. https://doi.org/10.1044/jshd.5501.43

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America, 123*(5), 3878–3878. https://doi.org/10.1121/1.2935783

Ziegler, A., Henke, T., Wiedrick, J., & Helou, L. B. (2018). Effectiveness of testosterone therapy for masculinizing voice in transgender patients: A meta-analytic review. *International Journal of Transgenderism, 19*(1), 25–45. https://doi.org/10.1080/15532739.2017.1411857

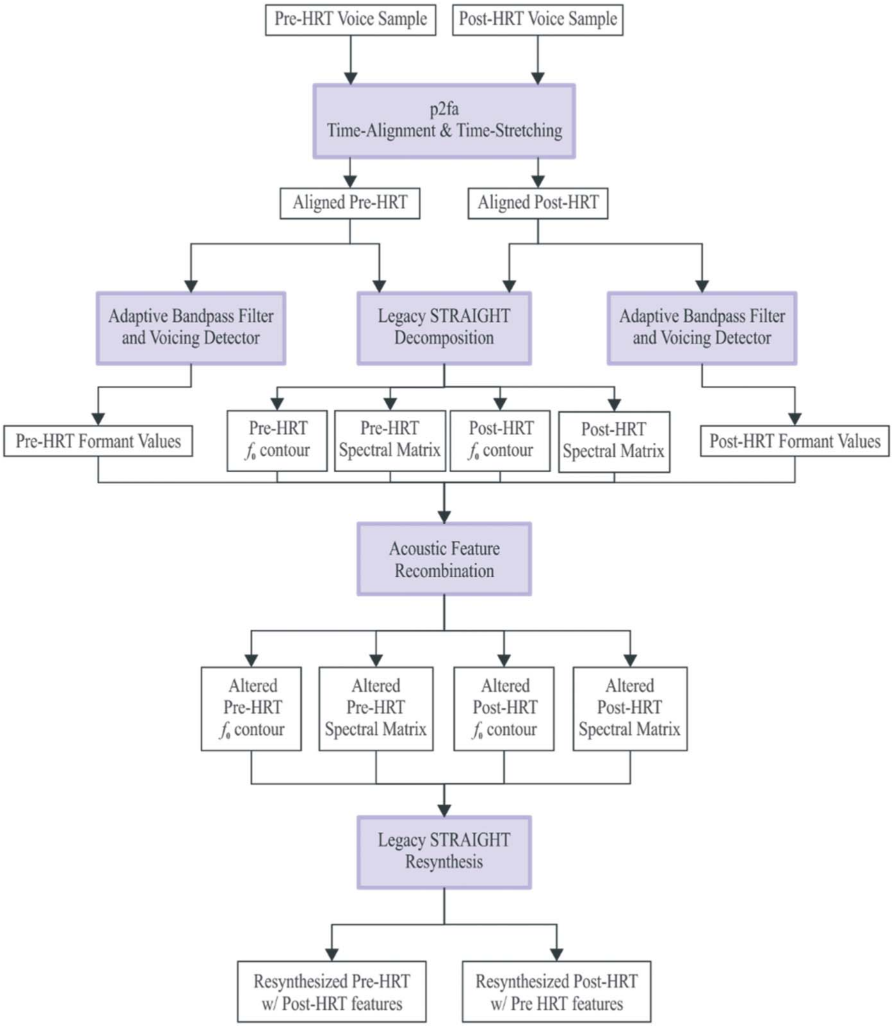
Zimman, L. (2018). Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass, 12*(8). https://doi.org/10.1111/lnc3.12284

Appendix (p. 1 of 2)

Supplemental Information on Resynthesis Method

The following sections detail the algorithm used to calculate acoustic features and resynthesize the stimuli set of 48 speech samples used in the listening experiment. For simplicity, these sections are written with the assumption that the pre-HRT speech sample is being changed to reflect acoustic feature values of the post-HRT speech samples (i.e., Samples 1–8 in Figure 1). For speech samples in which the post-HRT speech sample is being changed to reflect acoustic feature values of the pre-HRT speech sample (i.e., Samples 9–16 in Figure 1), the same steps are followed, but with the roles of the pre-HRT and post-HRT speech sample switched. A summary flowchart of the resynthesis algorithm is shown in Figure A1.

Figure A1. A summary schematic of the resynthesis algorithm used to create the 48 speech samples in the stimulus set. The initial input to the algorithm is a set of speech samples recorded prior to and following a year of hormone replacement therapy with exogenous testosterone (Pre-HRT and Post-HRT, respectively). p2fa = Penn Phonetics Lab Forced Aligner; STRAIGHT = Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum; f_o = fundamental frequency.



Pre-HRT Voice Sample
Post-HRT Voice Sample
p2fa
Time-Alignment & Time-Stretching
Aligned Pre-HRT
Aligned Post-HRT
Adaptive Bandpass Filter and Voicing Detector
Legacy STRAIGHT Decomposition
Adaptive Bandpass Filter and Voicing Detector
Pre-HRT Formant Values
Pre-HRT f_0 contour
Pre-HRT Spectral Matrix
Post-HRT f_0 contour
Post-HRT Spectral Matrix
Post-HRT Formant Values
Acoustic Feature Recombination
Altered Pre-HRT f_0 contour
Altered Pre-HRT Spectral Matrix
Altered Post-HRT f_0 contour
Altered Post-HRT Spectral Matrix
Legacy STRAIGHT Resynthesis
Resynthesized Pre-HRT w/ Post-HRT features
Resynthesized Post-HRT w/ Pre HRT features

Appendix (p. 2 of 2)

Supplemental Information on Resynthesis Method

Preliminary Signal Processing

The pre-HRT and post-HRT speech samples were single sentences of running speech manually cropped from longer recordings. Prior to acoustic feature extraction, pre-HRT and post-HRT speech samples were time-aligned and time-stretched at the phonemic level. A MATLAB script was used to implement the Penn Phonetics Lab Forced Aligner (p2fa), which detected the locations of each phoneme in the running speech sample (Yuan & Liberman, 2008). A speech VOCODER (Legacy STRAIGHT) was then used to separate each speech sample into a source vector (corresponding to the f_o contour) and a spectral matrix (Kawahara et al., 1999). Legacy STRAIGHT is described in the following section. Linear interpolation was then used to stretch the source and spectral components of the post-HRT sample to match the locations of the phonemes in the pre-HRT sample. After interpolation, the post-HRT sample was resynthesized with the new time-aligned and time-stretched source and spectral components. Since the process of separating and resynthesizing a speech sample results in a small loss of quality in the speech signal, the original pre-HRT sample was also resynthesized without changes to its source vector or spectral matrix in order to generate baseline samples (i.e., Sample 1 in Figure 1) that were comparable to the other resynthesized samples. Finally, both pre-HRT and post-HRT speech samples were normalized by the root-mean-square (RMS) of the pre-HRT sample. The result of this preprocessing is two speech samples per speaker with identical durations at a phonemic level and similar RMS values. All additional processing was performed on these preprocessed signals.

Legacy STRAIGHT Spectral Separation and Resynthesis

Following time alignment and time stretching, the first four formants for each speech sample were automatically calculated using a custom MATLAB script adapted from Mustafa and Bruce (2006). Each formant was calculated using an adaptive voicing detector and an adaptive bandpass filter that were updated based on the values of the preceding formant frequencies. This approach allowed for more robust estimates of formant frequencies over a range of signal-to-noise ratios (Mustafa & Bruce, 2006). Formant values were averaged per vowel, resulting in an $m \times 4$ matrix of average formant values per speech sample, where m is the total number of vowels in the running speech sample. The standard deviation of each formant value per vowel was also calculated.

Next, Legacy STRAIGHT was again used to separate each speech sample into the source vector and the spectral matrix. Legacy STRAIGHT applies pitch-adaptive spectral smoothing to fully remove interferences from source periodicity in the spectral matrix (Kawahara et al., 1999). This completely decouples the spectral matrix from the voice source such that the spectral matrix does not include fine structures from higher order harmonics. As a result, changes can be made to the spectral matrix without impacting the harmonics of the source. Likewise, changes to f_o in the source vector will also change the harmonic structures of the speech sample upon resynthesis. The Legacy STRAIGHT algorithm was implemented in MATLAB using publicly available code provided by the original authors (Kawahara, 2018).

When using Legacy STRAIGHT, the source vector corresponded directly to the f_o contour, wherein f_o was calculated at a rate of 1000 Hz. For unvoiced segments of running speech, the f_o contour was undefined. The mean f_o per speech sample was calculated using the average of the f_o contour. The f_o contour per speech sample was then converted to STs using the mean f_o, such that each f_o contour represented ST deviations from the mean f_o of the signal. Thus, the f_o contour could be separated from the mean f_o such that two speech samples may have different means but identical f_o contours in ST. Prior to resynthesis, the source vector of the pre-HRT sample could be adjusted to reflect the mean f_o and the f_o contour of the post-HRT sample, as shown in Figure 1.

In order to change the formants of a speech sample, linear interpolation was used to stretch the spectral matrix along the frequency axis at each time frame (also sampled at 1000 Hz) that corresponded to a vowel in the running speech. Interpolation points were defined as the average formants of the corresponding vowel ± 1 *SD*. This made it so that the frequency content around each formant band would be shifted to a new formant value without changing the overall spectral content of the signal. This newly aligned spectral matrix was then used to resynthesize a speech sample with different formant values.

During algorithm design, it was observed that the formants of the resulting resynthesized speech samples were not consistently matching the target formants. For example, when the spectral matrix of the pre-HRT sample was altered to reflect the formants of the post-HRT sample, the new speech sample did not match the formants of the post-HRT sample. This is because changes to the spectral matrix are not identical to changes in the spectrum of the final speech sample. It was generally observed that the difference in formant values between the original values and the target values was larger than the difference between the original values and the final values. As a result, the difference between the original formant values and the target formant values was multiplied by a scale value to compensate for this discrepancy. This scale value could range from 0 to 2 and was different for each formant. Each scale value was automatically optimized for each set of speech samples using a feedback loop in the algorithm. The scale value that corresponded to the smallest difference between target formant value and final formant value was used in final resynthesis.

Using the above process, 16 resynthesized speech samples were created per speaker, representing every combination of the three acoustic features from the pre-HRT and post-HRT samples. In order to mitigate potential changes in voice quality between samples as a result of the speech resynthesis process, the unaltered pre-HRT and post-HRT samples were decomposed and resynthesized as well. Thus, all 16 of the samples used for the listening experiment were resynthesized samples. Similarly, in order to mitigate differences in signal amplitude, every sample was normalized by the RMS of the resynthesized pre-HRT sample.