INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Cues for Perception of Gender in Synthetic Voices and the Role of Identity

Maxwell Hope[1], *Jason Lilley*[2]

[1]University of Delaware, Department of Linguistics and Cognitive Science, Newark, DE, USA
[2]Nemours Biomedical Research, Wilmington, DE, USA
maxhope@udel.edu, Jason.Lilley@nemours.org

Abstract

Perception of gender in voice is not an under-researched area. Previous studies have been conducted in the hopes of pinpointing what aspects of voice (e.g. fundamental frequency, intonation, etc.) carry the largest cues for skewing gender perception. These studies have to date been conducted within the framework of the gender binary, i.e. men's vs. women's voices, which have left out the exploration of perception of something besides simply femininity and masculinity.

The literature thus far has not endeavored to keep pitch in the "androgynous" zone while manipulating other aspects such as the F0 contour or other acoustic parameters. Additionally, past literature on speech perception has neglected to explicitly include members of the gender expansive community. Hence, we recruited participants of all genders and first sought to identify cues for gender perception in synthetically made voices and then examine the relationship between one's own sense of gender identity and the perception of gender in synthetically made voices for native speakers of American English. We found that vocal tract acoustics are most important for swaying perception of gender and one's own gender identity influences gender perception in voice.

Index Terms: speech perception, synthetic voice perception, sociophonetics, gender expansive, gender perception

1. Introduction

The previous studies on gender perception in speech have been limited to the gender binary, for example perceiving masculinity and femininity or men's and women's voices [1, 2] and have not explored how perception of other genders might manifest. These studies also lacked the recruitment of participants outside of the gender binary. Therefore, this study helps to fill a gap by recruiting members of the gender expansive (GE) population, which includes transgender, non-binary and other gender variant people. Additionally, the current literature has only manipulated fundamental frequency (f0) specifically, along with other aspects such as formant frequencies [1, 2], by increasing or decreasing pitch to sound stereotypically masculine or feminine, but has not endeavored to hold the f0 in the "androgynous" range (140-160 Hz [3]) while manipulating other aspects such as the f0 contour or other acoustic parameters. This study will consider the current literature on speech and gender perception while taking a novel approach to speech stimuli and perception ratings.

1.1. Factors regarding speech and gender perception

The typical adult female mean f0 is 196-224 Hz and the typical adult male mean f0 is 107-132 Hz [3]. Thus, f0 can contribute to perception of a voice sounding typically feminine or typically masculine, or somewhere in between. Other factors that contribute to perception include the vocal tract shape. For example, one study involving transwomen examined how changing vocal tract shape such as "spreading the lips wider and bringing the tongue forward" when speaking increases vowel formant frequencies, which increases perception of femininity [4]. Another contributing factor to speech and voice perception is intonation. One study examined the perception of gender through voice among transwomen and transmen, finding that speaking with a falling intonation, as opposed to rising, might help to distinguish speakers as male [5].

Social factors and personal identity are also linked to speech production by non-binary individuals, that is, those who do not identify as men or woman. A recent descriptive study examining the pitch characteristics of those who are gender non-binary found that non-binary people produce pitch in the middle of the pitch range for men and the range for women; in addition to pitch, they appeared to mix intonation patterns of both men and women [6]. Nevertheless, this was a production study and therefore, a perception study investigating the link between gender identity and perception can be helpful in further understanding the production-perception interface for members of the GE community.

1.2. Questions and hypotheses

This study seeks to further examine perception of gender in voice by using synthetically made voices with various acoustic parameters. Synthetic voice construction allows us to isolate certain parameters more easily and precisely than using natural speech. For example, it allows us to create voices within the androgynous pitch range, but which have their F0 contour or vocal tract characteristics manipulated to be either neutral, feminine, or masculine. This lends to an ability to create many combinations of voices to test which cues are most important for perceiving voices with "androgynous" pitch as more masculine, feminine, or other depending on the other manipulated parameters.

First, we endeavor to identify what cues (fundamental frequency, f0 contour, vocal tract acoustics) in synthetic voices contribute most to various gender perceptions (feminine, masculine, and other). Second, we seek to examine the relationship between one's own sense of gender identity (rated flexibly on scales of 0 to 100 for femininity, masculinity, and something other than masculinity nor femininity) and the perception of gender in synthetically made voices. In addition to these main objectives, we focus on gathering perception judgements from the GE community, which has a valuable and broader perspective on gender identity and expression, to compare to a control group from the cisgender community. We hypothesize that vocal tract characteristics will be most important for both the cisgender and GE community in perceiving femininity and masculinity and that degree of

http://dx.doi.org/10.21437/Interspeech.2020-2657

feminine, masculine, and other identity will have a relationship with feminine, masculine, and other perception respectively.

2. Methods and Materials

2.1. Speakers and recordings

Forty American English speakers, 20 male and 20 female (age range 20-78, mean 56.9, SD 12.87) who showed no signs of dysarthria, were selected from the ModelTalker database [7]. Voices were selected for a wide range of fundamental frequencies for both males and females. Mean f0 values (over all voiced epochs of all recordings, measured with *reaper* [8]) for males ranged from 94.3 to 148.3 Hz, with an overall mean of 111.4 Hz; the females ranged from 137.6 to 219.8 Hz, mean 174.4 Hz. The mean over all 40 voices was 142.91 Hz. Each speaker had recorded a common set of 1589 utterances, which were used for model training. The 48-kHz recordings were down-sampled to 16 kHz before training.

2.2. Synthetic voice construction

Three "baseline" voices (all-female, all-male, and all-neutral) were built using the standard Merlin [9] DNN synthesis process in which two DNN models were trained: 1) a duration model that takes 229 linguistic features as input, and predicts the number of 5-msec frames per phone as output; and 2) an acoustic model with the same linguistic features as input, and a set of 187 acoustic features per frame as output, which include 180 mel-generalized coefficients (MGC), 3 band aperiodicity (BAP) features, 3 log f0 features, and voicing. The recordings of the 20 female speakers were used to train the all-female models, the 20 male speakers for the all-male models, and all 40 speakers for the all-neutral models. For these models, the natural f0 values extracted from the training material were unaltered for the training.

For the other six synthetic voices, we first modified Merlin scripts to train separate f0-contour and vocal-tract models for each of three gender conditions "feminine," "masculine," and "neutral." As above, the recordings of the 20 female speakers were used to train the "feminine" models, the 20 male speakers for the "masculine" models, and all 40 speakers for the "neutral" models. The same 229 linguistic features were used as the input to both model types. The vocal-tract model was trained to predict the 180 MGC and 3 BAP features, while the f0-contour model was trained to predict the 3 log-f0 features and voicing. Prior to training the f0-contour models, each speaker's mean f0 was adjusted to match the global mean by adding a fixed value per frame equal to the difference between the speaker's mean f0 and the global mean (143 Hz). (If this calculation produced a negative value for a particular frame, the original value for that frame was used instead.)

All training used Theano [10]. All DNN models used six fully-connected layers of 1536 units apiece with tanh activation and no dropout. Training used stochastic gradient descent (initial learning rate 0.002 with exponential decay), with 10 warmup and 30 training epochs (25 for f0 models). Batch sizes were 64 for duration models and 256 for other models. For synthesis we used the WORLD vocoder [11] (D4C edition [12]). For the three baseline voices, we used the trained duration models and matching acoustic models to generate the WORLD vocoder features. For the other six voices, we combined f0-contour models and vocal-tract models as in Table 1. The duration model used was from the baseline voice that matched the gender condition of the f0-contour model. The sentences synthesized were from the Harvard Sentences [13]. The first six of these were used in the speech perception survey.

Table 1: *Nine synthetic voices; F0 is the fundamental frequency, 'contour' indicates the fluctuations in F0 over time, and 'vocal tract' refers to the other acoustic information including BAP and MGC.*

Voice ID	F0	Contour	Vocal Tract
FFF	Fem	Fem	Fem
MMM	Masc	Masc	Masc
NNN	Neutral	Neutral	Neutral
NFN	Neutral	Fem	Neutral
NMN	Neutral	Masc	Neutral
NFF	Neutral	Fem	Fem
NMM	Neutral	Masc	Masc
NNF	Neutral	Neutral	Fem
NNM	Neutral	Neutral	Masc

Fem F0 = ~174 Hz, Masc F0 = ~111 Hz, Neutral F0= ~143 Hz.

2.3. Participants: listeners

Participants over the age of 18 who were native speakers of American English were recruited online via email and social media to partake in a speech perception survey. They were informed that their answers would remain anonymous. A total of 48 participants completed the online survey with 20 of them identifying as being a part of the GE community and 28 identifying as cisgender. The age of the participants ranged from 20 to 69 (M = 31.8, SD = 13.3).

2.4. Speech perception survey

Participants first answered demographic questions including age, languages spoken other than American English, and questions about their gender identity, including three questions that asked them to rate their own degrees of feminine, masculine, and other gender identity on scales from 0 to 100. Next, in the first listening trial, they were presented with a screen in which they could repeatedly listen to a Harvard sentence generated by each of the nine voices (presented in a pre-randomized order that was the same for all participants), allowing the listener to compare the voices directly. They were asked to rate the "femininity" of each stimulus on a scale from 0 to 100. This was similar to a MUSHRA design [14] except it did not contain a reference voice, as gender perception is more subjective than sound quality and we did not want to prime the listeners in any direction. In the second trial, this was repeated with a different sentence, which was rated for all nine voices for "masculinity." Next, a different sentence was rated for "something other than masculinity or femininity." Finally, the whole procedure was repeated with three new sentences, for a total of six trials. In total, each voice was listened to and rated twice for each perception condition, and the two ratings were averaged to get the listener's femininity rating, masculinity rating, and other rating for each voice.

2.5. Statistical analyses

Statistical analyses were computed in Excel. One-way analyses of variance (ANOVA) were computed to discover main effects of voice condition on perception of femininity, masculinity, and otherness. Post-hoc Tukey HSD tests were then used to find significantly different perceptions of voices while correcting for multiple comparisons. A correlation matrix of Pearson's r values between the different gender identity aspects (feminine,

masculine, and other) and different gender perceptions (feminine, masculine, and other) was created to identify potential trends in the data. This led to computations of significance of the strongest correlations identified.

3. Results

Three one-way ANOVAs were conducted for the three different ratings of the nine voices for all participants together ("whole group"). Post-Hoc Tukey HSDs were then computed for all possible pairs (n = 36) for each ANOVA. This process was repeated for two participant subgroups: cisgender and GE. Ratings for FFF, MMM, and NNN were significantly different from each other for femininity and masculinity in both the whole group and between the two subgroups ($p < 0.001$).

3.1 Ratings for all voices for whole group

As determined by separate one-way ANOVAs, there were statistically significant main effects of voice condition on femininity rating ($F(8,423) = 126.93$, $p < 0.0001$), masculinity rating, ($F(8,423) = 111.06$, $p < 0.001$), and 'other' rating ($F(8,423) = 8.59$, $p < 0.0001$). The distributions of ratings for femininity, masculinity and other perception of all nine voices for the whole group are shown in the boxplots in Figures 1, 2, and 3.

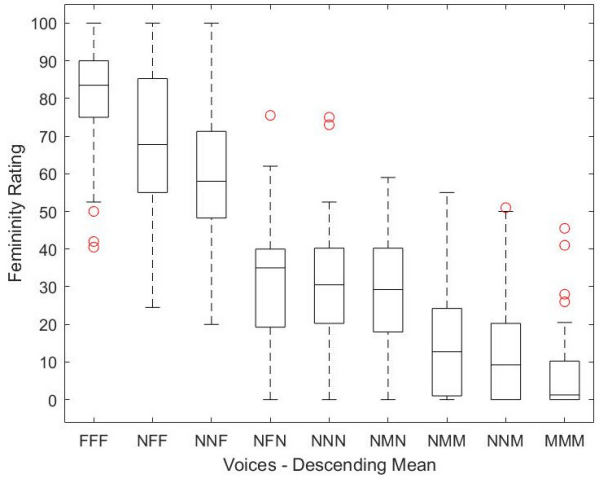

Figure 1: *Femininity ratings, all voices and listeners.*

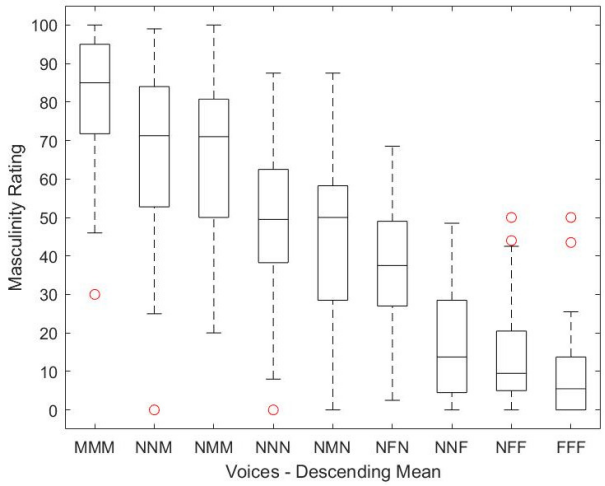

Figure 2: *Masculinity ratings, all voices and listeners.*

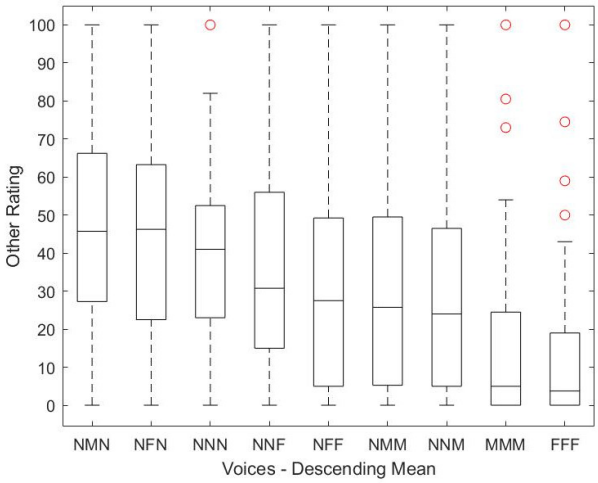

Figure 3: *Other ratings, all voices and listeners.*

Regarding the femininity ratings in Figure 1, the Tukey HSD results indicate that NFF vs NNF, NNN vs NFN, NNN vs NMN, NFN vs NMN, NMM vs MMM, and NNM vs MMM were not significantly different ($p > 0.05$). All other pairs were significant ($p < 0.01$). Therefore, while FFF had an additional effect on perception of femininity compared to NFN, MMM did not further decrease perception of femininity when compared to NMM. These results also indicate a separation of the nine voices into three "categories". This occurs at the switch of the vocal tract parameters from feminine to neutral and again from neutral to masculine.

For the masculinity ratings shown in Figure 2, the Tukey HSD results indicate that NNM vs NMM, NNN vs NMN, NMN vs NFN, NNF vs NFF, NNF vs FFF, and NFF vs FFF were not significantly different ($p > 0.05$). All other pairs were significant at $p < 0.01$, except NNN vs NFN which was significant at $p < 0.05$. Similarly to the results in Figure 1, this shows three "categories" of the nine voice. This division of categories again occurs at the shift of the vocal tract, this time from masculine to neutral and then from neutral to feminine.

Finally, for the other perception ratings shown in Figure 3, FFF vs NFF, MMM vs NMM, MMM vs NNM, NNN vs NFN, NNN vs NMN, NFN vs NMN, NMN vs NMM, NFF vs NNF, and NMM vs NNM were not significantly different ($p > 0.05$) while all others were significant ($p < 0.05$). For these ratings, there is much more variability and less distinct categories; however, the two binary voices MMM and FFF are rated lowest for other perception.

3.2 Participant group differences

The differences in the post-hoc tests for the GE group versus the cisgender group are reported in Table 2. If the comparison does not appear, it means that this comparison had the same result from the post-hoc test for both groups (either both significant or both insignificant). A total of 11 out of 108 possible pairs were different.

The first three rows in Table 2 show that cisgender participants had significantly different perceptions in femininity and masculinity when only the fundamental frequency was manipulated (for FFF vs NFF and MMM vs NMM) as well as different perception in masculinity when both the fundamental frequency and contour were changed, while the GE participants did not have significantly different perceptions

for femininity and masculinity based off of these changes alone. The remainder of Table 2 shows that overall, changes in F0, contour, and vocal tract do change the perception of otherness for GE participants but not for cisgender participants. Figure 4 highlights some of these key differences in the groups for the baseline voices.

Table 2: *Tukey HSD Post-hoc result differences with p-values.*

Perception	Comparison	GE	Cis
Femininity	FFF vs NFF	0.66	0.008**
Masculinity	MMM vs NMM	0.63	0.009**
Masculinity	MMM vs NNM	0.66	0.01*
Other	FFF vs NNN	0.005**	0.11
Other	FFF vs NNF	0.17	0.03*
Other	MMM vs NNN	0.001**	0.60
Other	MMM vs NFN	0.001**	0.15
Other	MMM vs NMN	0.001**	0.21
Other	MMM vs NNF	0.04*	0.31
Other	NMN vs NFF	0.03*	0.90
Other	NMN vs NNM	0.03*	0.63

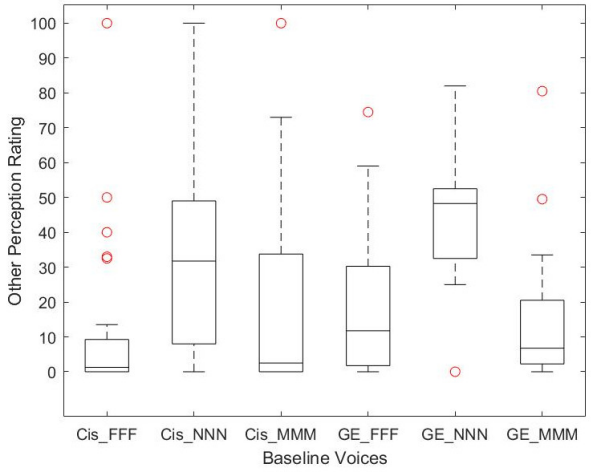

Figure 4: *Cisgender and Gender Expansive other perception for the three baseline voices.*

3.3 Relationship between gender identity and perception

The correlation matrix resulted in identification of which relationships between gender identity and perception were strongest. Participant's feminine gender was not significantly correlated with any of the three gender perceptions. Masculine identity was negatively correlated with perception of masculinity in the three voices perceived most neutrally. Other gender was positively correlated with other perception in 6 of the voices. Results for these two relationships are in Table 3.

4. Discussion and conclusions

The results from our whole group analysis showed that the nine voices were essentially divided into three groups: a "feminine" group, a "neutral" group, and a "masculine" group. In general, people used vocal tract acoustics to inform their gender perception more heavily than pitch contour. For example, for feminine and masculine perception as shown in Figures 1 and 2, the vocal tract parameter was the key parameter for shifting or skewing perception of gender (e.g. from NNF to NFN and NMN to NNM in Figure 1). However, when looking at some comparisons between groups in Table 2, it was found that cisgender participants were more influenced by change in f0 alone for perception of femininity and masculinity compared to GE listeners.

Table 3: *Pearson's r correlation values between degree of masculine identity and perception and degree of other identity and perception. Correlation significance values: * for p < 0.05, ** for p < 0.01*

Voice	Masc & Masc	Other & Other
FFF	-.22	.35 *
MMM	-.04	.11
NNN	-.39 *	.46 **
NFN	-.35 *	.36 *
NMN	-.31 *	.46 **
NFF	-.22	.23
NMM	-.15	.35 *
NNF	-.15	.22
NNM	-.10	.36 *

For the correlational analysis, our hypothesis that there was a relationship between gender identity and gender perception was only partially confirmed. There was no relationship between feminine identity and any of the gender perceptions. There was, on the other hand, a negative relationship between masculine identity and masculine perception; and those with stronger other gender identity would rate voices as being perceived as more other. Based off the current results, it appears that those with greater other identity are more likely to perceive things as other – similar to a native speaker picking up on native cues of voicing and phonemic contrasts. This is shown in the results in Figure 4 as well: as those who were part of the GE community were more likely to have a stronger other identity, we see that they group the totally neutral voice separately from the totally masculine and totally feminine ones in terms of otherness, whereas cisgender participants did not do this. We conclude that those with other identity, to a degree, create a third separate category of other perception in addition to "feminine" and "masculine".

One limitation of this study is that we were not able to conduct an experiment with all possible voice combinations such as repeating our procedures for voices with pitches in the feminine range and masculine range. It will be important to repeat this study looking at other pitch/vocal tract/contour combinations.

This study has shown that gender perception is more complex than a binary choice between masculine and feminine or men and women. Additionally, the greatest cue for skewing this gender perception was the vocal tract parameter. While this is not a completely new finding, it expands our knowledge of how those from a wide variety of genders might categorize different speakers. Finally, this study shows the important role of identity, especially identity outside of the binary, to distinguish "otherness" as a unique and distinct category.

5. Acknowledgements

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

6. References

[1] J. M. Hillenbrand and M. J. Clark, "The role of f0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, & Psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009. DOI: 10.3758/app.71.5.1150

[2] V. G. Skuk and S. R. Schweinberger, "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," *Journal of Speech, Language, and Hearing Research,* vol. 57, no. 1, pp. 285–296, 2014. DOI: 10.1044/1092-4388(2013/12-0314)

[3] S. Davies and J. M. Goldberg, "Clinical aspects of transgender speech feminization and masculinization," *International Journal of Transgenderism,* vol. 9, no. 3-4, pp. 167–196, 2006. DOI: 10.1300/j485v09n03_08

[4] L. Carew, G. Dacakis, and J. Oates, "The effectiveness of oral resonance therapy on the perception of femininity of voice in male-to-female transsexuals," *Journal of Voice,* vol. 21, no. 5, pp. 591-603, 2007.

[5] A. Hancock, L. Colton, and F. Douglas, "Intonation and gender perception: Applications for transgender speakers," *Journal of Voice,* vol. 28, no. 2, pp. 203-209, 2014.

[6] M. Schmid and E. Bradley, "Vocal pitch and intonation characteristics of those who are gender non-binary." Presented at the 2019 International Congress of Phonetic Sciences. [Online]. Available: https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_178.pdf

[7] H. T. Bunnell, J. Lilley, and K. McGrath, "The ModelTalker project: A web-based voice banking pipeline for ALS/MND patients," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings,* 2017, pp. 4032-4033.

[8] D. Talkin (2015). *REAPER: Robust Epoch And Pitch EstimatoR.* [Online]. Available: https://github.com/google/REAPER

[9] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW9)*, *Sunnyvale, CA, USA*, Sep. 13-15, 2009, pp. 218-233. Available: http://ssw9.talp.cat/download/ssw9_proceedings.pdf

[10] Theano Development Team (May 2016). *Theano: A Python framework for fast computation of mathematical expressions.* [Online]. Available: http://arxiv.org/abs/1605.02688

[11] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99-D, No. 7, pp. 1877-1884, 2016.

[12] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57-65, Nov. 2016. Available: http://www.sciencedirect.com/science/article/pii/S0167639316300413

[13] "IEEE Recommended Practice for Speech Quality Measurements," in *IEEE Transactions on Audio and Electroacoustics,* vol. 17, no. 3, pp. 225-246, September 1969. DOI: 10.1109/TAU.1969.1162058.

[14] International Telecommunications Union (Oct. 2015). "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems." Available: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf