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

An Objective Voice Gender Scoring System and Identification of the Salient Acoustic Measures

Fuling Chen[1], *Roberto Togneri*[1], *Murray Maybery*[2], *Diana Tan*[2,3]

[1]Dept. of Electrical, Electronic and Computer Engineering, University of Western Australia
[2]School of Psychological Science, University of Western Australia
[3]Telethon Kids Institute, Perth, Australia

fuling.chen@uwa.edu.au, roberto.togneri@uwa.edu.au, murray.maybery@uwa.edu.au, diana.tan@uwa.edu.au

Abstract

Human voices vary in their perceived masculinity or femininity, and subjective gender scores provided by human raters have long been used in psychological studies to understand the complex psychosocial relationships between people. However, there has been limited research on developing objective gender scoring of voices and examining the correlation between objective gender scores (including the weighting of each acoustic factor) and subjective gender scores (i.e., perceived masculinity/ femininity). In this work we propose a gender scoring model based on Linear Discriminant Analysis (LDA) and using weakly labelled data to objectively rate speakers' masculinity and femininity. For 434 speakers, we investigated 29 acoustic measures of voice characteristics and their relationships to both the objective scores and subjective masculinity/femininity ratings. The results revealed close correspondence between objective scores and subjective ratings of masculinity for males and femininity for females (correlations of 0.667 and 0.505 respectively). Among the 29 measures, F0 was found to be the most important vocal characteristic influencing both objective and subjective ratings for both sexes. For female voices, local absolute jitter and Harmonic-to-Noise Ratio (HNR) were moderately associated with objective scores. For male voices, F0 variance influenced objective gender scores more than the subjective ratings provided by human listeners.

Index Terms: gender scoring, masculinity, femininity, LDA

1. Introduction

The perceived degree of masculinity in males and femininity in females is associated with the development of secondary sex characteristics [1]. Influenced by genetics, hormones and the environment, these secondary sex characteristics have been found to correlate with health status, physical strength and mating success [2, 3, 4, 5]. The human voice is one form of sexually selected morphological trait, and is amenable to judgements of masculinity and femininity. Several studies have shown that vocal masculinity/femininity plays an important role in social behaviours [6, 7, 8, 9]. A common research method used in examining vocal masculinity/femininity is to collect subjective gender scores, in which listeners assign degrees of masculinity/femininity to human voices. However, the acquisition of perceptual gender scores is both time and resource consuming.

Several studies have investigated relationships between various acoustic measures and the perceived masculinity/femininity of voices, as well as the utility of the acoustic measures in objectively discriminating male and female speakers. Vocal-tract length (VTL) has been shown to influence male and female speakers' acoustic quality where longer VTL has been associated with lower F0 [10, 11], lower formant dispersion [7], and higher perceived masculinity and attractiveness, as rated by female listeners [7]. It was also demonstrated that male speakers who were taller and had higher testosterone levels had lower F0 and resonance (ΔF), and their voices were rated as more masculine [6, 12]. Similarly, speakers of both sexes whose voices had naturally either low F0 or low formant frequencies (Fn), or both were rated as being more masculine [13]. Apart from the conventional cues, jitter, shimmer and Harmonic-to-Noise Ratio (HNR) have also been investigated [14, 15]. However, only jitter parameters were found to be statistically significantly higher in males than in females [14]. Whereas the study of voice quality [16] found that HNR, jitter and shimmer were correlated significantly with biological sex. It was reported that the classification of sex using a combination of F0, jitter, shimmer and HNR achieved accuracy of 99% for both males and females [16]. Therefore, the existing literature suggests that acoustic measures such as F0, F0 variance, Fn, ΔF, VTL, HNR, jitter and shimmer could be valid cues to vocal masculinity and femininity as assessed by human listeners. Several studies have focused on developing computation models for the binary classification of sex based on voice samples and using particular acoustic measures. Examples include applying support vector machines (SVMs) on F0 and Fn [17], a linear prediction (LP) model on F0 related measures [18] and LDA on voice quality measures [19].

Several studies have identified the acoustic measures that differentiate male and female speakers. Other studies have identified acoustic measures that correlate with listeners' ratings (subjective gender scores) of masculinity/femininity. To the best of our knowledge, there has been no studies that have attempted to build an objective gender scoring model by using a comprehensive set of acoustic measures to differentiate between males and females. Additionally, limited effort has been invested in generating a continuous gender score based on input data with binary labels of males and females. As a result, it is currently unclear whether an objective gender score based on biological sex classification would correlate with subjectively rated masculinity/femininity. A previous study utilised 3D facial measurements to generate objective gender scores based on measures that could accurately classify individuals on biological sex using a computational model based on LDA [20]. The objective gender score was verified by investigating its correlations with subjective gender scores within each sex. We adapted this approach to the study of voice masculinity/femininity.

Using a new dataset of speech segments from 434 speakers which is more than 3 times larger than the datasets used in pre-

http://dx.doi.org/10.21437/Interspeech.2020-1627

vious research [6, 7, 8, 10, 11, 12, 13, 14, 16], the current study aims to:

1. Establish a novel model to derive an objective gender score based solely on using a comprehensive set of acoustic measures that differentiate between the two biological sexes.
2. Propose a machine learning method for determining the relative weighting of the acoustic measures in accounting for gender classification.
3. Assess the extent to which an objective gender score derived from the gender classification correlates with the subjective masculinity/femininity ratings for males/females.

The rest of this paper is organized as follows. Section 2 gives a detailed description of the proposed model. Section 3 describes the databases. Section 4 presents the obtained results, and is followed by analysis and discussion in Section 5. Section 6 concludes the paper with a brief summary and direction for future work.

2. System Description

An overview of the gender scoring system is given in Figure 1.

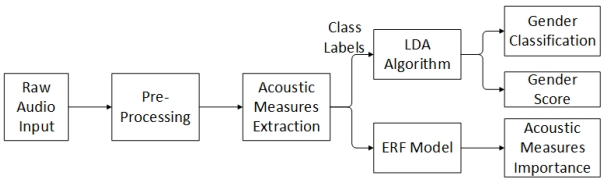

Figure 1: *Block Diagram of the Proposed System*

2.1. Acoustic Measures

All the audio files (see Section 3) were used to obtain the targeted speakers' utterances and were segmented into 1 second time frames. A set of 29 widely known acoustic measures were extracted for each time frame. Among these measures, meanF0Hz, stdevF0Hz, Harmonic-to-Noice Ratio (HNR), all jitter measures (local jitter, local absolute jitter, rap jitter, ppq5 jitter and ddp jitter)[1] and all shimmer measures (local shimmer, local db shimmer, apq3/5/11 shimmer and dda shimmer)[2] were obtained from Parselmouth 0.3.3 which is a Python library for the Praat software. The mean and median of F1, F2, F3 and F4 measure formants at each glottal pulse using the formant position formula [21]. The VTL is estimated in seven measures: formant position (pF) [21], formant dispersion (fdisp) [22], average formant frequency (avgFormant) [23], geometric mean formant frequency (mff) [24], estimation of VTL (VTL_Fn) [25], formant spacings (ΔF) [26] and estimation of the maximum VTL (VTL_ΔF).

On inspection, each of the acoustic measures was approximately normally distributed for each sex. Some of the 29 parameters were highly correlated with each other, such as the mean and median values of each formant frequency, as well as measures of jitter and of shimmer. In this study because correlations are not expected to affect the computational modelling negatively, we decided to keep as much information as possible.

[1] http://www.fon.hum.uva.nl/praat/manual/Voice_2__Jitter.html

[2] http://www.fon.hum.uva.nl/praat/manual/Voice_3__Shimmer.html

Given these acoustic measures, a data-driven model is proposed to be used in generating objective gender scores for comparison with subjective ratings of masculinity and femininity.

2.2. Computational Gender Scoring Model

A Linear Discriminant Analysis (LDA) classifier was used on the acoustic measures and the corresponding binary labels (0 for females and 1 for males). LDA benefits from its ability to reduce dimensions of the feature set while retaining the information that discriminates output classes. In this study, LDA tries to find a decision boundary around the clusters of classes of males and females. It then projects the 29-dimensional data points to new dimensions, in a way that the two clusters are as separate as possible and the individual data points within a cluster are as close to the centroid of the cluster as possible. The transformed new dimensions are ranked on the basis of their ability to maximize the distance between the clusters and minimize the distances between the data points within each cluster. While some of the acoustic measures are highly correlated, LDA takes advantage of information on the multiple dimensions and transforms them into the LDA space, without adverse effect.

LDA has been commonly used in classification models. In this study, LDA was used for classification based on the work of [20]. In [20], a gender score was obtained by measuring the ratio of the distance between each projected sample data point and the midpoint between the means for the male and female classes to the distance between the means of the two classes. However, in the present study, given two clusters with irregular hyperplanes, this method may not be applicable. Thus for this study, an objective gender score was derived from the confidence score of the LDA algorithm attributes. The confidence score of each sample belonging to two classes is the signed distance of this sample to the boundary point c, which is defined as:

$$\begin{cases} O_{gsm} = \| \omega^* x - c \|_1, if\ sample\ x\ is\ labelled\ as\ male \\ O_{gsf} = - \| \omega^* x - c \|_1, if\ sample\ x\ is\ labelled\ as\ female \end{cases} \quad (1)$$

Where ω^* is the projection vector of LDA, which is learned from the development data labelled for male and female; and c is the boundary point of the two classes, which is calculated as $w^* \frac{1}{2}(\mu_0^* + \mu_1^*)$, and where μ_0^* and μ_1^* are the means of the two classes. To normalise the gender scores for comparison with the subjective scores, the gender scores were converted to z-scores for each sex. As the objective scores have a mean of 0 and a distribution spread of up to 4 standard deviations, for ease of presentation the final objective scores were shifted by +4 for males and -4 for females denoting as O_{gsm}^z and O_{gsf}^z. Figure 2 depicts the distribution of the objective gender scores and the calculation method, where O_{gsm}^z is in blue and O_{gsf}^z is in red.

2.3. Acoustic Measure Importance

The Extreme Random Forest (ERF), which was deployed to extract the importance of the acoustic measures, generates a set of weights summing up to 100% across all the acoustic measures. The ERF is more suitable than other methods in the case of numerical input and categorical output, given multi-dimensional data. Most psychoacoustic studies ranked the importance of acoustic dimensions by means of the effects of manipulations on human ratings or by using classification accuracy, which constrains the statistical analysis on the acoustic measures. The ERF, popularly used in clinical research [27] to estimate feature importance in multi-dimensional data, is selected as the tool in

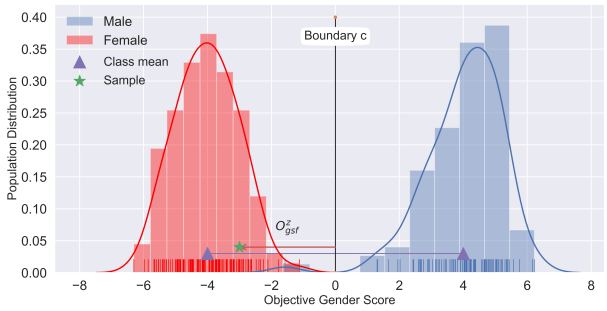

Figure 2: *Distribution of Objective Gender Score*

this study. The ERF is preferred as it is much faster than the random forest method and is less prone to overfitting.

2.4. Evaluation Criteria

The main focus of this study is to establish a computational model that can mimic subjective perceptual ratings of vocal masculinity/ femininity. The outcomes of this study are the objective gender scores and the importance of acoustic measures. The evaluation criterion for the objective gender scores is the correlation between the subjective masculinity/femininity ratings and the objective scores derived from the model. The importance of the most discriminating acoustic measures, in the objective rating, will be analysed by comparing them with values reported in the literature.

3. Data

The database used was obtained from the School of Psychological Science at the University of Western Australia. Voice recordings were collected from another study which investigated the association between perceived gender ratings and autistic traits [28]. This database was chosen because it contains more speakers with available perceived gender ratings than any public database.

The database (see description in Table 1) is composed of three cohorts of 434 adult participants (268 females and 166 males) who were undergraduates and fluent in English. Tested individually in a soundproof room, each participant provided voice samples by reading the Rainbow passage [29] using a conversational tone. Only the second sentence from the passage was used for the masculinity and femininity ratings.

Table 1: *Database Description*

Cohort No.	**1**	**2**	**3**
Collected year	2015	2017-2018	2019
Speakers mean age	18.9 yrs	20.87 yrs	19.09 yrs
Number of speakers	22 m* 22f*	70 m 139 f	74 m 107f
Subjective rating	Yes	No	Yes
Number of raters	30	–	25
Rating scale	1-10	–	1-100

* m - males; f - females

Human gender ratings were provided by raters who did not know the speakers. For each rater, the voices for each class were presented in random order. Following the presentation of each voice through enclosed headphones, a rating scale appeared on the screen. The scale ranged from 1 to 10 for Cohort 1 and 1 to 100 for Cohort 3, with the extreme points labelled 'not at all masculine' and 'extremely masculine' for male voices, and 'not at all feminine' and 'extremely feminine' for female voices.

In summary, among the 434 speakers (166 males and 268 females), 96 male speakers were rated for their masculinity and 129 female speakers were rated for their femininity, with the remaining speakers used only in gender classification. To correct for how listeners may have used the masculinity/femininity rating scales differently, the ratings for each listener were first converted to z-scores. This also enabled the merging of ratings across Cohorts 1 and 3. In order to visualize the results, the mean value was shifted to -2 for the femininity ratings for females and +2 for the masculinity ratings for males, as all subjective z-scores were initially had a mean of 0 and a distribution spread of up to 2 standard deviations. As a result, the mean values of human ratings for each sex may differ from the corresponding mean values of the objective scores. However, this shift in the standardized ratings did not affect correlations or the analysis of acoustic measure importance.

The recruitment and testing of all participants were conducted in accordance with the ethics approval obtained for this study from the Human Research Ethics Committee at the University of Western Australia.

4. Results

4.1. Classification and Objective Scores Analysis

The LDA model classified 99.77% of the speakers correctly as male or female from the entire 3 cohorts. This confirms the ability of the LDA to accurately separate males and females. The gender scores derived from the LDA model were moderately to highly correlated with the corresponding subjective femininity and masculinity ratings, with Pearson correlations (r) of 0.505 and 0.667 obtained for females and males, respectively. Figure 3 shows the scatter plot of the relationship between subjective and objective gender scores in males and females. The objective gender score is more highly correlated with the corresponding subjective gender score in males than in females. Possible reasons will be considered in the next section.

Figure 3: *Correlations between Subjective Gender Scores and Objective Gender Scores*

4.2. Acoustic Measures Importance

The weights of the 29 acoustic measures, provided by the ERF, demonstrate the important cues in gender classification. Among all the acoustic measures, whose weights sum up to 100%, the F0_mean (weight of 36.46%), consistent with literature, was the most discriminating measure in differentiating males and females. The contributions of the next most important acoustic measures were F0_stdev (7.9%), local absolute jitter (7.21%), mean value of F1 (3.55%), HNR (3%), VTL_Fn (2.44%), formant position (pF, 2.4%) and geometric mean formant frequency (mff, 2.35%).

Table 2 shows the correlations of the acoustic measures with the corresponding subjective and objective scores in males and females.

Table 2: *Correlations of Acoustic Measures with Subjective and Objective Scores in Males/Females*

Measures	Male Subjective Score	Male Objective Score	Female Subjective Score	Female Objective Score
F0_mean	0.60	0.92	0.49	0.90
F0_stdev	0.14	0.52	0.21	0.17
local abs jitter	0.33	0.45	0.37	0.69
F1_mean	0.20	0.12	0.30	0.20
HNR	0.17	0.17	0.13	0.56
VTL_Fn	0.46	0.36	0.14	0.14
pF	0.46	0.46	0.15	0.17
mff	0.46	0.37	0.11	0.18

Considering the top three acoustic measures from Table 2, in Figure 4 we show the scatter plots for each of these measures with the objective and subjective gender scores with female data in red and male data in blue.

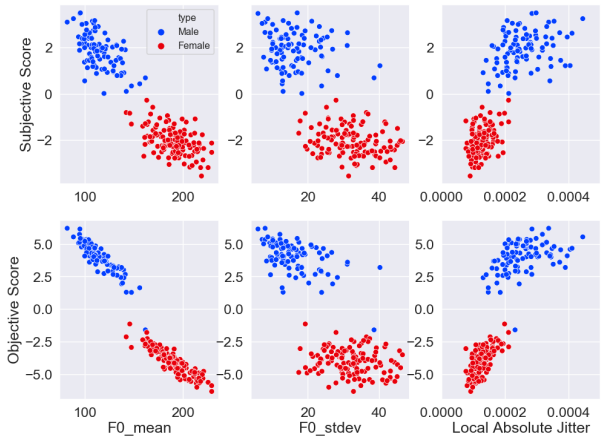

Figure 4: *Scatter Plots of Subjective Gender Score and Objective Gender Score with Three Key Acoustic Measures*

5. Discussion

The objective gender scores, based on all 29 acoustic measures, correlated more highly with the subjective gender scores (r = .667 in males; r = .505 in females), than any other single measure (F0_mean: r = .6 in males; r = .49 in females). Compared to the facial gender scoring [20] (r = .895 in females and r = .794 in males), the voice gender scoring shows weaker correlations between objective and subjective gender scores. In the study of [20], raters were asked to nominate the facial regions that they used to make their judgements on the masculinity of males and the femininity of females. Regarding voice, raters may not be able to easily identify the acoustic measures that influence their gender judgements. Thus the relationship between each measure and its effect on perceptual gender scoring may not be evident. Raters from different cultural backgrounds may vary in their experience of masculinity and femininity, which could influence their ratings and therefore the correlations with objective scores.

Consistent with the literatures [6, 7, 10, 11, 12, 13, 16], the mean value of F0 (F0_mean) had the strongest association of any of the acoustic measures with both the objective and subjective gender scores for males and females, aligning with its primary role in gender classification. From Table 2 it was shown that F0_mean correlates only moderately with the subjective gender scores, with the correlation larger for males than for females. For the objective scores, it is noticed that the computational model is more dependent on F0_mean, thus it strongly dominates the scoring. As demonstrated in Table 2 and Figure 4, the standard deviation of F0 (F0_stdev) was the second most important measure in categorizing the two sexes and correlated highly with the objective gender scores for males, but less so for females. Raters did not rely on F0_stdev to assess masculinity and femininity, with weak correlations observed. Consistent with [14, 15, 16], local absolute jitter was associated at moderate levels with both the subjective and objective gender scores for both sexes. Furthermore, in the present study, local absolute jitter was found to be more associated with the femininity ratings. The estimation of vocal-tract length (VTL_Fn), formant position (pF), and geometric mean formant frequency (mff) had moderate associations with both objective and subjective gender scores for males, but not for females. The HNR, another moderate-level contributor to gender scores for females, was associated with objective scores, aligning with study [16].

6. Conclusions

This study proposed a computational model based on LDA to rate human speakers' vocal masculinity and femininity, verified by the correlations between the objective scores and subjective scores given by human listeners (r = 0.667 in males and r = 0.505 in females). The study used ERF to analyse the importance of the 29 acoustic measures. Moreover, it investigated the relationships between these measures and the subjective and objective scores. The results show that mean value of F0 is the most important factor in subjective scoring and gender classification, especially in assessing males' masculinity. The standard deviation of F0, vocal-tract length, formant position and geometric mean formant frequency contribute more in assessing males' masculinity than in assessing females' femininity, while local absolute jitter and HNR stand out in determination of females' femininity. Generally, the model's performance is highly coherent with human perceptual ratings. A key limitation is that the computational model was trained for gender classification and a regression approach should be considered when there is sufficiently labelled data. In this way a more reliable analysis of the acoustic factors that affect both subjective scores and objective scores can be made.

7. References

[1] M. Andersson, *Sexual selection*. Princeton University Press, 1994, vol. 72.

[2] B. Fink, N. Neave, and H. Seydel, "Male facial appearance signals physical strength to women," *American Journal of Human Biology*, vol. 19, no. 1, pp. 82–87, 2007.

[3] J. Hönekopp, U. Rudolph, L. Beier, A. Liebert, and C. Müller, "Physical attractiveness of face and body as indicators of physical fitness in men," *Evolution and Human Behavior*, vol. 28, no. 2, pp. 106–111, 2007.

[4] M. M. Samson, I. Meeuwsen, A. Crowe, J. Dessens, S. A. Duursma, and H. Verhaar, "Relationships between physical performance measures, age, height and body weight in healthy adults." *Age and ageing*, vol. 29, no. 3, pp. 235–242, 2000.

[5] R. Thornhill and S. W. Gangestad, "Facial sexual dimorphism, developmental stability, and susceptibility to disease in men and women," *Evolution and Human Behavior*, vol. 27, no. 2, pp. 131–144, 2006.

[6] V. Cartei, R. Bond, and D. Reby, "What makes a voice masculine: Physiological and acoustical correlates of women's ratings of men's vocal masculinity," *Hormones and Behavior*, vol. 66, no. 4, pp. 569–576, 2014.

[7] D. R. Feinberg, B. C. Jones, M. L. Smith, F. R. Moore, L. M. DeBruine, R. E. Cornwell, S. Hillier, and D. I. Perrett, "Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice," *Hormones and behavior*, vol. 49, no. 2, pp. 215–222, 2006.

[8] A. C. Little, J. Connely, D. R. Feinberg, B. C. Jones, and S. C. Roberts, "Human preference for masculinity differs according to context in faces, bodies, voices, and smell," *Behavioral Ecology*, vol. 22, no. 4, pp. 862–868, 2011.

[9] D. R. Feinberg, L. M. DeBruine, B. C. Jones, and A. C. Little, "Correlated preferences for men's facial and vocal masculinity," *Evolution and Human Behavior*, vol. 29, no. 4, pp. 233–241, 2008.

[10] S. A. Collins, "Men's voices and women's choices," *Animal behaviour*, vol. 60, no. 6, pp. 773–780, 2000.

[11] D. R. Feinberg, B. C. Jones, A. C. Little, D. M. Burt, and D. I. Perrett, "Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices," *Animal behaviour*, vol. 69, no. 3, pp. 561–568, 2005.

[12] S. J. Ko, C. M. Judd, and I. V. Blair, "What the voice reveals: Within-and between-category stereotyping on the basis of voice," *Personality and Social Psychology Bulletin*, vol. 32, no. 6, pp. 806–819, 2006.

[13] K. Pisanski and D. Rendall, "The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2201–2212, 2011.

[14] J. P. Teixeira and P. O. Fernandes, "Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices," *Procedia technology*, vol. 16, pp. 1228–1237, 2014.

[15] A. Lovato, W. De Colle, L. Giacomelli, A. Piacente, L. Righetto, G. Marioni, and C. de Filippis, "Multi-dimensional voice program (mdvp) vs praat for assessing euphonic subjects: a preliminary study on the gender-discriminating power of acoustic analysis software," *Journal of Voice*, vol. 30, no. 6, pp. 765–e1, 2016.

[16] M. Biemans, *Gender variation in voice quality*. Netherlands Graduate School of Linguistics, 2000.

[17] Y.-L. Shue and M. Iseli, "The role of voice source measures on automatic gender classification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4493–4496.

[18] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research*, vol. 43, no. 2, pp. 469–485, 2000.

[19] M. Lugger and B. Yang, "Classification of different speaking groups by means of voice quality parameters," *Proceedings of ITG-Sprach-Kommunikation*, 2006.

[20] S. Z. Gilani, K. Rooney, F. Shafait, M. Walters, and A. Mian, "Geometric facial gender scoring: objectivity of perception," *PloS one*, vol. 9, no. 6, 2014.

[21] D. A. Puts, C. L. Apicella, and R. A. Cárdenas, "Masculine voices signal men's threat potential in forager and industrial societies," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1728, pp. 601–609, 2012.

[22] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *The Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, 1997.

[23] K. Pisanski and D. Rendall, "The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2201–2212, 2011.

[24] D. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3177–3186, 2005.

[25] A. Paige and V. Zue, "Calculation of vocal tract length," *IEEE Transactions on audio and electroacoustics*, vol. 18, no. 3, pp. 268–270, 1970.

[26] D. Reby and K. McComb, "Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags," *Animal behaviour*, vol. 65, no. 3, pp. 519–530, 2003.

[27] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," 2013.

[28] D. W. Tan, S. N. Russell-Smith, J. M. Simons, M. T. Maybery, D. Leung, H. L. Ng, and A. J. Whitehouse, "Perceived gender ratings for high and low scorers on the autism-spectrum quotient consistent with the extreme male brain account of autism," *PloS one*, vol. 10, no. 7, 2015.

[29] G. Fairbanks, *Voice and articulation drillbook*. Harper & Brothers, 1940.