



An investigation of regression-based prediction of the femininity or masculinity in speech of transgender people

Leon Liebig¹, Christoph Wagner¹, Alexander Mainka², Peter Birkholz¹

¹ Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany

² Department of Audiology and Phoniatics, Charité-Universitätsmedizin Berlin, Germany

leon.liebig@tu-dresden.de, peter.birkholz@tu-dresden.de

Abstract

Transgender individuals often seek for voice modification to more closely have their voice matched with their new sex, and avoid potential stigmatization or even discrimination. Whereas treatment options such as voice therapy or surgery exist, a quantitative measure of the treatment outcome is missing. In this paper, we therefore propose a novel regression-based method to predict the perceived femininity or masculinity of a speaker's voice. To this end, 86 speakers (34 male, 35 female, 17 transgender) were recorded reading aloud a German standard passage. Subsequently a group of 28 laypersons and 13 experts rated the femininity/masculinity of these speech samples. Each spoken utterance was automatically analysed with respect to nine different pitch-, resonance- and voice quality-related acoustic features. The ratings were the targets for three prediction models (linear, logistic and decision tree regression) based on the extracted features. The results show that, generally, f_0 and the vocal tract length (VTL) are the main predictors. Furthermore, the continuous outcome logistic regression model with f_0 , smoothed cepstral peak prominence (CPPS), Jitter and VTL as input features performed best and achieved promising results with a cross-validated root mean-squared error of 0.117 on the normalized ratings [0,1].

Index Terms: automatic speech analysis, gender perception, nonlinear regression, acoustic features, vocal tract length, transgender

1. Introduction

The term transgender is used to describe a group of individuals whose gender identity does not correspond to their biological sex [1, 2]. The voice as part of a person's identity provides information of the speaker, including emotion, age and gender [3]. Due to the individual physiological characteristic of the vocal tract and the larynx, each individual's voice is unique. However, gender-specific differences in the voice exist, which makes the voice a secondary sexual characteristics [4].

The average fundamental frequency f_0 as a measure of the pitch of the voice is reported to range between 100-120 Hz for men and 200-220 Hz for women [3]. It is also known that the vocal tract of male adults is longer than that of female individuals, resulting in higher formant frequencies F_1 to F_4 for female voices [3]. In addition, females seem to have a more breathy voice than males [5].

Transgender with a gender-atypical voice often experience negative reactions by society, resulting in discrimination and prejudices [1]. It has been shown that the voice and the quality of life are correlated [6]. Therefore, transgender individuals often seek voice modification to have a gender-appropriate voice which corresponds to their new sex. While hormone therapy for

female-to-male (FtM) transgender people often achieves a satisfactory male voice, it has almost no effect on the vocal folds and vocal tract of male-to-female (MtF) transgender people [7]. As such, MtF-transgenders often need further treatment, including voice therapy or surgical operation. However, quantifying the femininity or masculinity of an individual's voice is difficult and there currently exists no tool to quantitatively assess or measure the treatment outcome w.r.t. the perceived gender of one's voice.

A substantial amount of research has been conducted on determining which acoustic features are most salient to contribute to gender perception. In [8], Leung et al. reviewed studies concerning aspects of speech and the perceived gender of the speaker for the English language. The most studied acoustic feature in this context is f_0 . A meta analysis revealed that f_0 explains 41.6 % of the variance of the listeners perception in gender. However, even if the pitch is attributed a central part of the perceived gender of a voice, it cannot be the only cue in gender perception [3].

The second most investigated feature is the resonance property of the vocal tract, measured with the first four formant frequencies. It was reported that F_1 , F_2 , F_3 and F_4 are all related to gender perception of the voice [8]. However, most studies that investigated formant frequencies used either isolated vowels instead of continuous speech, or manipulated the formant frequencies of small speech samples, which must be considered cautiously, because these stimuli are not well suited to represent natural language.

Whereas a number of studies showed a relation between breathiness as a feature for voice quality and gender perception, its corresponding used acoustic measures (voiced turbulence index and difference of the level of first and second harmonic) showed mixed results [8]. Other aspects of speech, such as articulation and intonation, are believed to contribute to gender perception, but have not been investigated systematically, yet.

Additionally, many studies are based on a limited acoustic sample size. The average sample size in all considered studies according to [8] is 27, which is relatively small. Studies with a larger sample size (greater than 90) [9, 10, 11] only considered speech samples of cis-male and cis-female speakers and did not take other gender identities into account. Also, very few studies exist for the German language.

Almost all current studies are based on a correlation analysis between gender perception and acoustic features. Whereas automatic voice-based gender classification is routinely performed in human-computer interaction, there currently exists no model to describe the specific relation between different acoustic features and gender perception. In this paper we investigated the regression-based prediction of the femininity/masculinity of the voice. Therefore, we recorded 86 speakers, who were asked to read aloud a German standard passage. We used ratings re-

garding the perceived femininity/masculinity of the speech samples and automatically extracted acoustic features as training data for different regression models.

2. Method

2.1. Speech recordings

A total of 86 participants were recruited, including 17 trans-female, 34 male and 35 female individuals (age 18 to 77, all native German speakers) at the Charité - Universitätsmedizin Berlin. None of the participants reported speech disorders. Written informed consent was obtained from all participants as well as the experiment approval by the Ethical Board Committee of the Charité - Universitätsmedizin Berlin (request number EA4/065/21). All participants were asked to read aloud the phonetically balanced German standard text "Der Nordwind und die Sonne", while their speech signal was recorded in a quiet room with a distance-adjustable head microphone (XION medical) with a fixed distance of 30 cm between the microphone and the mouth, and with a sampling rate of 22050 Hz and a 32 Bit quantisation.

2.2. Listening experiment

The gender perception listening experiment was designed using *Praat* [12]. A 6-point Likert scale from 1 (very masculine) to 6 (very feminine) was chosen to evaluate the perceived gender of each speech sample, because this scale range provided most information as recommended by [13]. From the entire recordings of the standard test, three sentences in the middle were selected to reduce the length of the stimuli to approximately 19 s. The loudness of each stimuli was normalized to -23 LUFS according to the standard *EBU-R128* [14] using *pyloudnorm* [15] which uses the recommended algorithm defined in *ITU-R BS.1770* [2]. Loudness could therefore not be considered as a feature. All stimuli were presented twice and randomly arranged without playing the same stimuli in a row. A group of 28 laypersons (14 male, 14 female) and 13 experts (4 male, 9 female), including speech therapists, took part in the listening experiment. None of them reported any hearing difficulties. As a result, we obtained a total of 41 ratings of femininity/masculinity for each of the 86 speech samples. However, expert ratings were solely used to assess rater reliability and did not enter the model dataset.

2.3. Feature Extraction

Nine different acoustic features were extracted from the speech samples, including f_0 , f_0 variation, harmonic-to-noise ratio (HNR), spectral tilt and slope, smoothed cepstral peak prominence (CPPS), Jitter, Shimmer and vocal tract length (VTL). For feature extraction *parselmouth* [16] was used.

Pitch: Mean f_0 as a measure of the related pitch can be estimated with different methods like *Praat To Pitch (ac)* [17], *STRAIGHT* [18] or others. *Praat's* autocorrelation method is the state-of-the-art and suited for low-noise speech samples. To improve the quality of pitch detection, usually the minimum and maximum frequency of the pitch is determined depending on the sex of the speaker. However, in our case, no a priori knowledge about gender is assumed. f_0 determination is unreliable when no appropriate limits are set because of the large possible pitch range. As such, a two-step approach was taken [19]. In the first step, f_0 values were estimated with the limits of 50 Hz and 600 Hz. In the second step, we used the empirical range

limits [19] of

$$f_{min} = 0.75 \cdot q_1 \text{ and } f_{max} = 1.5 \cdot q_3 \quad (1)$$

based on the 1st and 3rd quartiles according to the f_0 values. The variation of f_0 as a measure of intonation was calculated in terms of the standard deviation of the f_0 samples.

Resonance characteristic: The voice's resonance characteristic was measured in terms of the highly gender-specific VTL, defined along the midline from the glottis to the lips [20]. VTL was chosen over individual formants, because formant frequencies are correlated (which is an undesirable property for regression analysis) and because it reduced the number of features. The VTL \hat{L} can be approximated from the acoustic signal with

$$\hat{L} = \frac{c}{4\hat{\Phi}}, \quad (2)$$

where c is the velocity of sound and the parameter $\hat{\Phi}$ can be considered as a linear combination of formants F_1 to F_4 [20], described by

$$\hat{\Phi} = \beta_0 + \frac{\beta_1 F_1}{1} + \frac{\beta_2 F_1}{3} + \frac{\beta_3 F_3}{5} + \frac{\beta_4 F_4}{7}. \quad (3)$$

The weights $\beta_{0,1,2,3,4}$ were taken from Lammert et al. [20]. Automatic formant estimation was done with the linear predictive coding (LPC) method from *Praat*. Calculating the formant track for the whole speech sample lead to spurious results, since non-vowels were considered as well. To overcome major errors in formant calculation, the automatic segmentation tool *Web-Maus* [21] was used and only formants in the central third of each vowel duration were considered. VTL was then averaged across the relevant time segments. The LPC method was configured to search for 5 formants in the frequency range [0, 5500] Hz with a window length of 25 ms.

Voice quality: The HNR was estimated with the short-term HNR analysis in *Praat* with default settings for voiced parts only. The estimated pitch contour was used to make a voiced-unvoiced-decision across the speech sample. For spectral parameters of voice quality, the slope and tilt were calculated with the long term average spectrum (LTAS) following the calculation in the acoustic voice quality index (AVQI) [22]. For the LTAS a bandwidth of 100 Hz was used. The tilt is defined as the decrease in the regression line of the LTAS. The slope is the ratio of average energy level between the bands of 0-1 kHz and 1-10 kHz [23]. The CPPS is well suited to describe the perceived breathiness in voice [24] and can be used for continuous speech. No voiced-unvoiced decision is needed for this feature. CPPS was obtained from the power cepstrum, which was subsequently smoothed. The CPPS was determined by calculating the difference of the peak and its corresponding value of the regression line through the cepstrum. Short time variation in period length (Jitter) and amplitude (Shimmer) in voiced parts were measured in local surroundings by calculating the perturbation quotient using five periods (PPQ and APQ respectively).

Figure 1 shows the extracted features of all speech samples and their corresponding average laypersons rating. Each histogram displays the distribution of the feature in the sample. No normal distribution can be assumed for the features. Since no major outliers were found, all acoustic features were min-max-normalised to a range from 0 to 1 for further regression analysis. The mean ratings were also normalized to a range from 0 (masculine) to 1 (feminine).

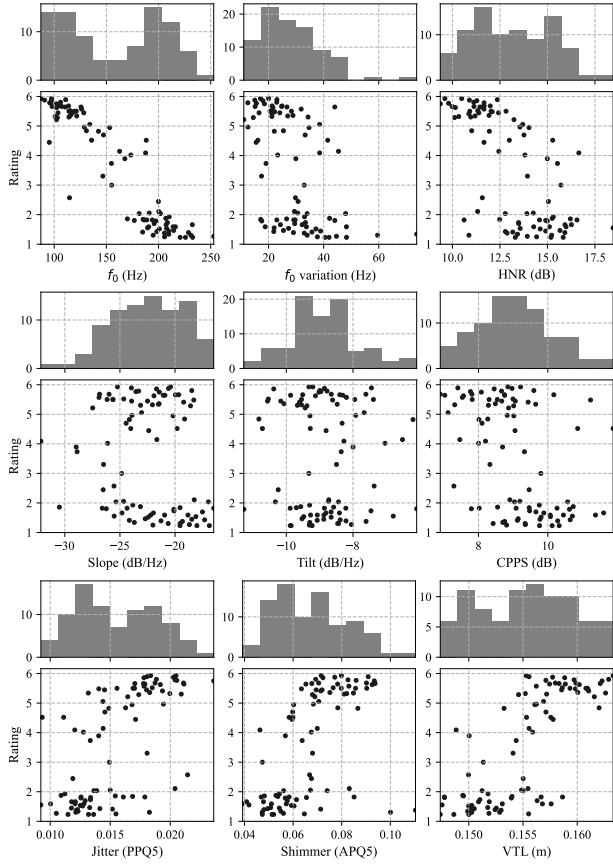


Figure 1: Scatter plots of extracted acoustic feature of the speech samples and the average laymen's rating. The histograms indicate the distribution of the features.

2.4. Regression analysis

The purpose of the regression analysis was to find a mapping function $f(\mathbf{x}, \mathbf{y})$ to predict the degree of femininity, y , i.e.,

$$\hat{y} = f(\mathbf{x}, \mathbf{w}), \quad (4)$$

with the (independent) input features $\mathbf{x} = (x_1, x_2, x_3 \dots)^T$ and weights $\mathbf{w} = (w_0, w_1, w_2, \dots)$. Three different regression models were compared: linear, nonlinear and decision tree regression. R^2 or adjusted R^2 (R_{adj}^2) are widely used as performance measure for regression [25]. It is known that R^2 is not well suited for nonlinear fits. Instead, other criterions like Akaike information criterion (AIC) are preferred [25]. However, the AIC is not defined for decision tree regression. Model evaluation was therefore performed using mainly the mean-squared-error (MSE), and compared to R_{adj}^2 when applicable. Evaluations were performed with k-fold cross validation with varying fold sizes of $k \in \{10, 20, 43, 86\}$. Cross-validated MSE was averaged across the used fold sizes.

Multiple linear regression was used as the baseline model. For feature selection, the lasso regularisation with L1-norm was computed, where the weight penalty λ was increased iteratively from 10^{-15} to 1. Using lasso regularisation, some weights w_j can become zero and the corresponding feature can be removed. The results of Lasso regularisation were compared with the approach of forward stepwise selection [26].

Visually inspecting the features in figure 1 already sug-

gested that a nonlinear regression might be more suitable for this particular application. Specifically, the logistic sigmoid function in the form of

$$\hat{y} = f(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-(w_0 + \sum_{i=1}^K w_i \cdot x_i))} \quad (5)$$

was assumed for the nonlinear mapping. Typically, logistic regression is used in binary response modelling for classification problems [27]. Here, we used the characteristics of the logistic regression function for a regression-based prediction of a continuous outcome. Feature selection in nonlinear regression is not as well explored compared to linear regression [28]. To find a subset of suitable features, we also performed a k-fold cross validation for every possible combination of $N = 1, 2, \dots, 9$ features and measured the cross-validated MSE. The set of features with the best cross-validated MSE was chosen for the final model. When no significantly better MSE was achieved by an additional feature, fewer features were chosen to avoid the curse of dimensionality and overfitting.

For the decision-based regression (using the CART-algorithm [29]), hyperparameters were determined with an initial random search, followed by a more detailed grid search using *scikit-learn* [30]. Hyperparameters and their respective search interval were: `max_depth`: {5, 6, 7, 8, 9}, `min_samples_leaf`: {1, 2, 3, 4}, `min_samples_split`: {3, 4, 5, 6, 7}, `max_features`: {4, 5, 6} and `max_leaf_nodes`: {30, 40, 50, 60, 70}.

3. Results

3.1. Rater reliability

Evaluating the reliability of the ratings obtained from the listening experiment is highly important. The question arises whether the data are a correct representation and allow reproducibility. The consensus of the ratings between different raters (inter-rater) and between one and the same rater on multiple repetitions of the stimuli (intra-rater) was evaluated. Krippendorff's α is a flexible method which can be used for an ordinal scale to measure rater reliability [31]. For the ratings of laymen, an inter-rater reliability of $\alpha_{L,1} = 0.86$ and an intra-rater reliability of $\alpha_{L,2} = 0.91$ was computed. The professional ratings show an inter-rater reliability of $\alpha_{P,1} = 0.85$ and an intra-rater reliability of $\alpha_{P,2} = 0.93$. Hence, all reliability results indicate strong to almost perfect reliability according to the interpretation of Cohen's Kappa [32].

3.2. Correlation

Correlation coefficients give a first impression about the importance of the features. Correlational analysis was carried out using Spearman's rank-order correlation coefficient r_s , which can be used for monotonic nonlinear relations. The correlation between the mean laymen's rating and the different extracted acoustic variables are shown in Table 1. f_0 and VTL correlate most with laymen's rating. f_0 has a significant positive correlation ($r_s = 0.91$) and VTL has a significant negative correlation ($r_s = -0.75$). An additional correlation analysis between the acoustic features showed multicollinearity among them. The highest inter correlation was determined between HNR and Shimmer ($r_s = -0.81$), HNR and Jitter ($r_s = -0.76$), f_0 and Jitter ($r_s = -0.75$), f_0 and VTL ($r_s = -0.68$) and between f_0 and f_0 variation ($r_s = -0.67$).

Table 1: Spearman’s correlation coefficient r_s for acoustic features and the gender-specific voice perception rating of lay participants ($p < 0.05$ indicates significance).

	f_0	Var(f_0)	HNR	Slope	Tilt	CPPS	PPQ	APQ	VTL
Spearman’s r_s	0.91	0.56	0.67	0.23	0.06	0.48	-0.69	-0.65	-0.75
p -value	0.000	0.000	0.000	0.036	0.610	0.000	0.000	0.000	0.000

3.3. Model evaluation

The results of the regression analysis are presented in Table 2, displaying the performance by averaged cross-validated MSE and R_{adj}^2 , and the selected features.

Linear regression with lasso regularisation and forward stepwise selection yielded the same results for feature selection, with only f_0 and VTL remaining. The specific regression coefficients were calculated using all laymen’s ratings for training, resulted in the equation:

$$\hat{y} = 0.175 + 1.037 \cdot f_0 - 0.277 \cdot VTL . \quad (6)$$

In logistic regression, the feature selection resulted in a feature subset of $\{f_0, CPPS, PPQ \text{ and } VTL\}$, and training using all data yielded in the equation:

$$\hat{y} = \frac{1}{1 + \exp(-\phi)} , \quad (7)$$

with $\phi = -3.502 + 6.900 \cdot f_0 + 1.078 \cdot CPPS + 2.509 \cdot PPQ - 2.274 \cdot VTL$. Feature importance was also calculated for decision-based regression, leading to the selection of f_0 , CPPS and VTL as final features.

With respect to R_{adj}^2 , the best model fit was found for decision tree regression. In terms of the MSE, however, logistic regression performed best. We performed a model selection based on MSE, since k-fold cross-validation tests on unseen data. Figure 2 shows a scatter plot of predicted ratings \hat{y} with logistic regression and the averaged laymen’s ratings y .

Table 2: Regression evaluation: averaged cross-validated MSE, R_{adj}^2 and selected features.

Model	Criteria		Features
	MSE	R_{adj}^2	
Linear	0.0173	0.88	f_0 , VTL
Logistic	0.0137	0.92	f_0 , CPPS, PPQ, VTL
Decision tree	0.0230	0.93	f_0 , CPPS, VTL

3.4. Follow-up evaluation for the clinical use case

The model evaluation was done taking both cis- and transgender data into account. The low cross-validated MSE is expected to be caused by the cisgender data influencing the interpretation for clinical application and possibly overestimating model performance. Specifically, the question arises what error is produced by testing on transgender data only. This is why we performed an additional assessment, using the best-performing model, the logistic regression. In a first experiment, we took the cisgender data for feature selection and training, and subsequently tested on all (unseen) transgender samples, investigating if cisgender data can be used to effectively “anchor” the sigmoid model function on cisgender data only. We calculated an

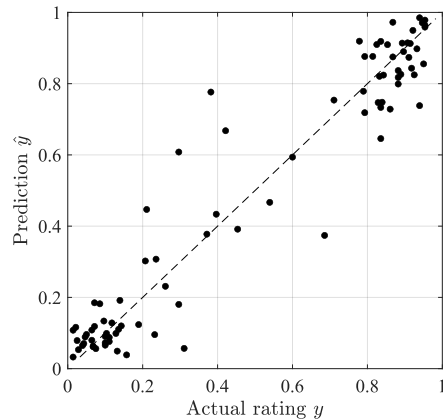


Figure 2: Scatter plot of prediction \hat{y} by logistic regression and averaged laymen’s ratings y used for training.

$MSE = 0.0529$. In a second experiment, we used the nested leave-one-out cross-validation, holding back a single transgender speech sample in each iteration and calculated an MSE of 0.0497.

4. Discussion and outlook

The results of the correlation analysis and the regression analysis revealed that the pitch- and the resonance-related features are most important for gender-specific perception, which is supported by current literature [8]. Additionally, features of voice quality also contribute significantly.

The root-mean-squared-error (RMSE, calculated from the MSE) indicates the average prediction error, with logistic regression having an RMSE of 0.117. Considering the range from 0 to 1, the model predicts quite well. For the clinical application, where only transgender voices are taken into account as test data (nested LOOCV), the RMSE of 0.2229 showed a less precise prediction. Also, for logistic regression, the gradient and steepness is highest in the transition region, where most transgender voices are located. As such, deviations in feature extraction will impact the estimation error most. The visualisation of features in Figure 1 and the scatter plot in Figure 2 show that ratings in the mid-range are less represented, compared to ratings around 1-2 and 5-6. Although the number of transgender speech samples is relatively large, compared to previous studies, it still remains under-sampled. As such, our experimental results present a proof-of-concept, an even larger database of speech samples would be necessary for actual clinical use. As for model development, we deliberately chose to average the ratings into a single continuous score for each speech sample. An arguably more refined approach would be to omit averaging, use a mixed model and treat the listeners as a random factor.

5. References

- [1] G. Dacakis, S. Davies, J. M. Oates, J. M. Douglas, and J. R. Johnston, "Development and preliminary evaluation of the transsexual voice questionnaire for male-to-female transsexuals," *Journal of Voice*, vol. 27, no. 3, pp. 312–320, 2013.
- [2] WHO, "Eleventh revision of the international statistical classification of diseases and related health problems (icd-11)," World Health Organization, Tech. Rep., 2019.
- [3] 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.
- [4] S. Salm, K. Hower, S. Neumann, and L. Ansmann, "Validation of the german version of the transsexual voice questionnaire for male-to-female transsexuals," *Journal of Voice*, vol. 34, no. 1, pp. 68–77, 2020.
- [5] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.
- [6] A. B. Hancock, J. Krissing, and K. Owen, "Voice perceptions and quality of life of transgender people," *Journal of Voice*, vol. 25, no. 5, pp. 553–558, 2011.
- [7] M. L. Gray and M. S. Courey, "Transgender voice and communication," *Otolaryngologic Clinics of North America*, vol. 52, no. 4, pp. 713–722, 2019.
- [8] Y. Leung, J. Oates, and S. P. Chan, "Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis," *Journal of Speech, Language, and Hearing Research (JSLHR)*, vol. 61, no. 2, pp. 266–297, 2018.
- [9] 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.
- [10] 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.
- [11] Y. Leung, J. Oates, S.-P. Chan, and V. Papp, "Associations between speaking fundamental frequency, vowel formant frequencies, and listener perceptions of speaker gender and vocal femininity-masculinity," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2600–2622, 2021.
- [12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.1.38, retrieved 2 January 2021 <http://www.praat.org/>, 2021.
- [13] L. J. Simms, K. Zelazny, T. F. Williams, and L. Bernstein, "Does the number of response options matter? psychometric perspectives using personality questionnaire data," *Psychological Assessment*, vol. 31, no. 4, pp. 557–566, 2019.
- [14] E. B. U. (EBU), "R 128 - loudness normalisation and permitted maximum level of audio," Operating Eurovision and Euroradio, Geneva, Switzerland, Tech. Rep., 08 2020.
- [15] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *Proc. of 150th Audio Engineering Society Convention*, Online, 2021.
- [16] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [17] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. of 17th IFA*, Amsterdam, Netherlands, 1993, pp. 97–110.
- [18] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. of the Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Florence, Italy, 2001, pp. 59–64.
- [19] D. Hirst, "A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation," in *Proc. of 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 2007, pp. 1233–1236.
- [20] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *Public Library of Science (PLOS) ONE*, vol. 10, no. 7, p. e0132193, 2015.
- [21] F. Schiel, C. Draxler, and J. Harrington, "Phonemic segmentation and labelling using the maus technique," in *Workshop New Tools and Methods for Very-Large-Scale Phonetics Research*, Philadelphia, USA, 2011.
- [22] Y. Maryn, P. Corthals, P. V. Cauwenberge, N. Roy, and M. D. Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," *Journal of Voice*, vol. 24, no. 5, pp. 540–555, 2010.
- [23] Y. Maryn, M. D. Bodt, B. Barsties, and N. Roy, "The value of the acoustic voice quality index as a measure of dysphonia severity in subjects speaking different languages," *European Archives of Otorhinolaryngology*, vol. 271, no. 6, pp. 1609–1619, 2014.
- [24] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.
- [25] A.-N. Spiess and N. Neumeyer, "An evaluation of r^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a monte carlo approach," *BMC Pharmacology*, vol. 10, no. 6, 2010.
- [26] T. Hastie, R. Tibshirani, and R. Tibshirani, "Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons," *Statistical Science*, vol. 35, no. 4, pp. 579–592, 2020.
- [27] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011.
- [28] Y. Sun, J. Yao, and S. Goodison, "Feature selection for nonlinear regression and its application to cancer research," in *Proc. of the 2015 SIAM International Conference on Data Mining*, Vancouver, Canada, 2015, pp. 73–81.
- [29] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, 1st ed. UK: Chapman and Hall/CRC, 1984.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: <http://jmlr.org/papers/v12/pedregosa11a.html>
- [31] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication Methods and Measures*, vol. 1, no. 1, pp. 77–89, 2007.
- [32] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemistry Medica*, vol. 22, no. 3, pp. 276–282, 2012.